2-2020

On the Distribution of Genetic Variation in Ecological Communities

Isaac Overcast
The Graduate Center, City University of New York

On the distribution of genetic variation in ecological communities

by

Isaac Overcast

A dissertation submitted to
the Graduate Faculty in Biology
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
The City University of New York

2020

On the distribution of genetic variation in ecological communities

by

Isaac Overcast

This manuscript has been read and accepted for the Graduate Faculty in Biology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

_____	_____
Date	Chair of Examining Committee Dr. Michael J. Hickerson, City College of New York
_____	_____
Date	Executive Officer (Acting) Dr. Christine Li

Supervisory Committee:

_____
Dr. Ana C. Carnaval, City College of New York

_____
Dr. Andrew J. Rominger, Santa Fe Institute

_____
Dr. Brian Tilston Smith, American Museum of Natural History

_____
Dr. Frank Burbrink, American Museum of Natural History

THE CITY UNIVERSITY OF NEW YORK

Abstract

On the distribution of genetic variation in ecological communities

by

Isaac Overcast

Advisor: Michael J. Hickerson, Ph.D.

Biodiversity in ecological communities is structured hierarchically across spatial and temporal scales. Many open questions remain as to how this structure accumulates. For example, what are the relative contributions of dispersal versus in situ speciation? Or, how important are stochastic drift versus deterministic processes? Up to this point, these questions have been investigated by isolated disciplines (e.g. macroecology, comparative phylogeography, macroevolution) using tools and data that tend to focus on only one axis of community scale data (e.g. phylogenies, relative abundances, and/or trait information). Yet we know that there are feedbacks among processes that respond on short, medium, and long time scales (local changes of abundance, accumulation of population genetic variation, and speciation processes, respectively). Therefore, the focus of my work is: first, to develop a model of the distribution of genetic variation in ecological communities; second, to construct a multi-scale model of the accumulation of biodiversity in ecological communities that jointly models three axes of data that respond on ecological, population genetic, and phylogenetic timescales; and third, to incorporate abiotic variables with community-scale genetic data in a machine learning framework to make predictions about the distribution of genetic variation across the landscape. First, I will present a modelling approach that involves merging Hubbell's neutral theory with neutral population genetic theory to construct a joint model of species abundance and genetic diversity. This model simulates joint distributions of abundance and genetic variation assuming both ecological and

population genetic neutrality, and captures both equilibrium and non-equilibrium dynamics. These simulations can be used for a variety of applications, including estimating the shape of the abundance distribution using only a sample of community-scale genetic data. Next, I will present a model that extends the double neutral model to incorporate non-neutral processes (such as ecological interactions) and to introduce a speciation process. The goal of this work is to fully integrate abundance and trait data with phylogenies and population genetic data into a unified framework with the aim of testing community assembly models and estimating ecological parameters using observed community data. One result of this work is the finding that genetic diversity is distributed more uniformly in ecological communities than abundance. Another critical insight is that community-scale genetic data provide a record of community history on a population-genetic timescale, which can complement ecological information obtained from sampled abundance data, and deep time community history recorded in phylogenies. Finally, I will describe a machine learning framework that integrates community-scale genetic data and abiotic variables (climatic/environmental) to make predictions about genetic diversity across the landscape. I demonstrate this method using densely sampled abundances and community-scale sequence data collected from 10 decapod crustacean communities distributed throughout the Coral Triangle. The observed distributions of abundance and genetic diversity in these communities largely agree with model predictions, in that abundance distributions demonstrated higher dominance. The machine learning inference procedure identified mean annual sea surface temperature and proximity of the sampling site to deep water as key factors contributing to the shape and magnitude of community-scale genetic diversity. As community-scale genetic data becomes easier to cost-effectively obtain, this only increases the importance of hierarchical models of biodiversity accumulation that account for feedbacks across timescales to make the most accurate inference about community history from this data.

Acknowledgements

I would like to thank my dissertation committee: Ana C. Carnaval, Andrew J. Rominger, Brian Tilston Smith, Frank Burbrink, and in particular my PhD advisor, Michael J. Hickerson, AKA "Mr. Low Tide". I would also like to individually thank the departmental administrators at the various institutions I've been affiliated with, without whom none of this would have been possible: Joan Ried (GC), Christine Klusko & Yolanda Pitt (CCNY), and Marnelli Candelario & Anna Manuel (AMNH). Thank you to all the current and past members of the Hickelab, the Smith lab, and the Burbrink lab, my inner circle of collaborators in the city. I will also individually thank Heather Heying, Stephane Boissinot and Deren Eaton. I thank the staff of the AMNH Herpetology department for graciously adopting me as an honorary herpetologist, and for helping me understand the most important part of the scientific process: Snacking! At the end of the day, I have made so many good friends and collaborators over these five years that I would run out of room trying to acknowledge everyone individually, so in the interest of space I will just say: Thank you so much and I am grateful to be a part of this community. Funding for my work was provided by grants from FAPESP (BIOTA, 2013/50297-0 to MJH and ACC), NASA through the Dimensions of Biodiversity Program (DOB 1343578), the National Science Foundation (DEB-1253710 to MJH; DEB 1745562 to ACC), and the Mina Rees Dissertation Fellowship in the Sciences provided by the Graduate Center of the City University of New York. Additionally, I am grateful for financial support provided by sDiv, the Synthesis Centre of the German Centre for Integrative Biodiversity Research (iDiv) Halle Jena Leipzig, and the Santa Fe Institute. I am, of course, infinitely grateful to my family and friends for all their support and encouragement. Finally, I thank Fubee, Sasquatch, and Lady Godiva.

Table of Contents

List of Figures and Tables

Chapter 1: An integrated model of population genetics and community ecology

Introduction

The species abundance distribution (SAD) is a classic summary of the structure of ecological communities (McGill *et al*. 2007), which is gaining increasing interest in areas of applied ecology and biodiversity management (Matthews & Whittaker 2015), community assembly (Fattorini *et al*. 2016), and biogeography in general (Matthews *et al*. 2017). However, unbiased comparative species abundance data is often challenging to obtain, a problem that is recognised to be particularly acute for invertebrates (Cardoso *et al*. 2011). Standardised sampling protocols can be implemented to improve comparability within studies (e.g. Emerson et al. 2017), but these do not account for idiosyncratic phenological or microhabitat differences among species that may affect sampling probability, potentially skewing estimates of relative abundance. Genetic sequence data retains a record of population size changes through time (Griffiths & Tavaré 1994; Drummond *et al.* 2005), yet this axis of information has rarely been exploited by community ecologists (Vellend 2005; Laroche *et al*. 2015), and never at the scale of the full community. Therefore, a model linking abundance and effective population size at the community scale could enable a new way to characterize abundance distributions indirectly from genetic data alone. Such rapid and cost effective estimation of SADs could greatly enhance understanding of the structure of ecological communities, with potential to aid in the design of conservation strategies, and to improve forecasts of changes in aggregate population dynamics in the context of global climate change.

The accumulation of sequence data for non-model organisms from over two decades of comparative phylogeographic studies (Avise, Bowen, & Ayala, 2016) large-scale DNA barcoding initiatives (Bucklin, Steinke, & Blanco-Bercial, 2011; Schoch et al., 2012), and forthcoming community-scale genome-wide data (Coissac, Hollingsworth, Lavergne, & Taberlet, 2016;

Garrick et al., 2015), presents us with an exciting opportunity for linking abundances and aggregate population genetic data . However, what is lacking is a flexible joint model that links existing models in comparative phylogeography (Carstens, Gruenstaeudl, & Reid, 2016; Huang, Takebayashi, Qi, & Hickerson, 2011; Jordan D. Satler & Carstens, 2016, 2017; Xue & Hickerson, 2017) with existing biogeographic models of community assembly (Etienne & Haegeman, 2011; Rosindell & Harmon, 2013; Rosindell, Harmon, & Etienne, 2015; Rosindell, Hubbell, He, Harmon, & Etienne, 2012).

Despite the potential of comparative phylogeography to leverage the power of aggregated demographic histories to answer fundamental questions about community assembly and macroecology (Avise et al. 1987; Hickerson et al. 2010; Avise et al. 2016), such approaches have generally neglected the growing body of theory from community ecology that seeks to accommodate the relative importance of deterministic (Tilman 2004; Maire et al. 2012) and stochastic processes (MacArthur & Wilson 1963; Hubbell 2001; Rosindell et al. 2012) governing the assembly of communities. For instance, comparative phylogeographic approaches that do incorporate community assembly have tended to focus on general models of shared demographic histories (Burbrink et al., 2016; Satler & Carstens, 2017; Stone et al., 2012), rather than models that are explicitly parameterized from ecological community assembly theory (but see Bunnefeld *et al*. 2018).

Ecological theory has been fundamental for understanding processes underlying spatial patterns of biodiversity as typically quantified by regional SADs and species area relationships (McGill *et al*. 2007; Matthews & Whittaker 2014). However, ecological models of community assembly tend to view communities as static pools with an ahistorical focus on equilibrium expectations (Weiher *et al*. 2011). Although there have been efforts to incorporate non-

equilibrium history in models of community assembly (Clark & McLachlan 2003), as well as a long tradition of incorporating phylogenetic information (Webb *et al*. 2002; Pearse *et al*. 2014) that also accommodates non-equilibrium historical dynamics (Pigot & Etienne 2015; Manceau *et al*. 2015), there has only been limited, yet promising, effort in considering intraspecific genetic polymorphism within a dynamic non-equilibrium assembly framework (Vellend *et al*. 2014; Laroche *et al*. 2015; McGaughran 2015) or within statistical models of macro-ecology (Miraldo *et al*. 2016; Smith *et al*. 2017; Pelletier & Carstens 2018). These efforts in bridging the gap between ecological models and population genetics have focused on characterizing the correlation between species diversity and genetic diversity in ecological communities (the species-genetic diversity correlation'; Vellend 2005; Papadopoulou *et al*. 2011; Vellend et al. 2014, Laroche *et al*. 2015) while other efforts have looked at the relationships between adaptive genetic diversity and community dynamics (Hughes *et al*. 2008; Becks *et al*. 2010; Schoener 2011).

Despite these important efforts to unify our understanding of ecological and evolutionary dynamics, a community-scale model linking species abundances and genetic diversities under a dynamic model of assembly has yet to be proposed. Here we describe, test, and demonstrate a joint inferential framework that bridges ecological neutral theory with population genetics in order to make joint predictions of community-wide distributions of species abundances, genetic diversities, and genetic divergences under a (Papadopoulou et al., 2011; Vellend, 2005; Vellend et al., 2014). The unified framework we present combines a forward-time model of island assembly with a backward-time coalescent model, linking abundance and colonization history with aggregated population genetic samples from multiple taxa.

First, forward-time community assembly simulations are performed using an

island/mainland metacommunity model following Rosindell & Harmon (2013). The individual-based neutral model of Rosindell & Harmon (2013) unifies MacArthur and Wilson's equilibrium theory of island biogeography (ETIB) with Hubbell's unified neutral theory of biodiversity (UNTB) to generate time-dependent non-equilibrium and equilibrium predictions of local richness and abundances (MacArthur & Wilson 1963; Hubbell 2001; Rosindell & Harmon 2013). We use these predicted temporal changes in abundance distributions and colonization times to parameterize a hierarchical multi-species model to simulate a sample of aggregate population genetic data backwards in time under the coalescent (Rosenberg & Nordborg, 2002). The former allows for inference about the time series progression of community change while the later links predicted changes in community population genetic data to this community assembly process.

We use simulation experiments to validate the power and accuracy of our method using an approximate Bayesian computation framework (ABC; Beaumont, Zhang, & Balding, 2002; Csilléry, François, & Blum, 2012; Lintusaari, Gutmann, Dutta, Kaski, & Corander, 2017). Subsequently, we demonstrate an application of our method to a sample of community wide mitochondrial DNA sequence data and corresponding densely sampled abundance estimates obtained from an assemblage of 57 spider species from the island of Réunion (Emerson et al., 2017). Using only the sequence data, we accurately estimate the Shannon's Index summary of the the observed SAD, and additionally obtain an estimate of the fraction of equilibrium obtained by the community. The joint model, implemented in Python, and all ipython notebooks for reproducing simulations and analysis are freely available on GitHub: https://github.com/isaacovercast/gimmeSAD.

Methods

Forward-time model - Forward time simulations of community assembly follow the

spatially implicit neutral model of (Rosindell & Harmon, 2013) that unifies the ETIB with the UNTB whereby abundance distributions, and immigration and extinction rates proceed under a birth/death/colonization process in the biogeographical context of a focal local community and a regional source pool (metacommunity). In this model the carrying capacity (*K*) of the local community consists of the sum of population sizes of all species on the island. This value is fixed, of finite size, and constantly saturated. The colonization rate is modeled as a single parameter (*c*) that specifies the probability of a colonization event. Colonizing species are sampled from a metacommunity composed of species with abundances that are logseries distributed, and which is static with respect to the timescale of local assembly. At each time-step one individual is randomly sampled for removal from the local community. With probability 1 - *c*, this individual is replaced by the offspring of a randomly sampled individual from the local community. With probability *c*, the individual is replaced by a randomly sampled member of the mainland metacommunity, where the probability of sampling from any given species is weighted by the relative metacommunity abundance (A_{meta}; Table 1.1).

Each time interval in the forward time simulation model can be described by a vector of $T_i^j = \{\tau_1^j, \ldots, \tau_{S_{local}}^j\}$ times since colonization (in generations) for each of the S_{local} species in the local community as well as a jointly associated vector of $A_i^j = \{A_1^j, \ldots, A_{S_{local}}^j\}$ local island abundances across the same S_{local} species in the local community. With respect to any particular time interval, the *j*th element for the *i*th local species in T_i^j and A_i^jdenotes time since the original colonization of the *i*th species from the meta-community. Therefore, $A_i^j = \{A_i^{\tau_i - 0}, \ldots, A_i^{\tau_i - \tau_i}\}$ and $T_i^j = \{\tau_i^{\tau_i - 0}, \ldots, \tau_i^{\tau_i - \tau_i}\}$ such that $j = \tau_i - 0$ at the final time interval declines going back in time at

previous time intervals until $\tau_i - \tau_i$. As the simulations progress forward in time, the species that go locally extinct become omitted sequentially through time, and the count of post-colonization migration events are accumulated per species in the vector $M = \{m_1, \ldots m_{S_{local}}\}$ (Table 1.1). Two emergent parameters (model response variables) are then c' (effective colonization rate) and † (effective extinction rate) which are defined as the realized number of colonization and extinction events per generation, respectively (Table 1.2).

Scaling forward time model to backward time coalescent model - For the *i*-th island species that is extant at a particular time interval with an abundance of A_j^i, there exists the history of changes in abundance over time since colonization τ_j^i, from a source species in the metacommunity. To relate raw sample-based abundances with the effective population sizes that parameterize the backwards time coalescent process of the gene tree lineages, we make the assumption of a random spatial distribution of individuals that is predicted to lead to a simple scaling relationship whereby the sample-based and regional-based abundance distributions have the same functional form (Green & Plotkin 2007). To approximate this expectation, we incorporate a rescaling that is based on the assumption that the observed abundances from direct sampling are proportional to actual abundances and current effective population sizes.

To this end we rescale the time-dependent abundance of each species (A_i^j) into a time-dependent effective population size (N_i^j) using the scaling factor σ such that $A_j^i \sigma = N_i^j$ whereby the numbers of individuals per species over time (A_i^j) is scaled to the number of demes of size σ over time per species. Across all species sampled genetically at a time interval, this yields time dependent vectors (N_i^j) of the effective population sizes for the i = {1, …, S_{local}} species, the

associated times since colonization in units of generations $T_i^j = \{\tau_1^j, \ldots, \tau_{S_{local}}^j\}$, and temporally static effective population size vectors for the corresponding source metacommunity species ($N_{meta}^{\square}$). Under this assumption, each island species consists of a metapopulation of σ demes of size A_i^j with strong migration conditions that reduce to the temporally dynamic predictions of a panmictic effective population of size $A_j^i\sigma$. Under this assumption of a metapopulation with strong migration conditions, the "collecting phase" is predicted to dominate the entire history of ancestry thereby approaching the standard panmictic coalescent expectations of a time dependent effective population size ($A_j^i\sigma = N_i^j$) as the number of demes become large (Wakeley & Aliacar 2001; Wakeley 2001; Wakeley 2004). Importantly, This rescaling is based on the assumption that the observed abundances from direct sampling are proportional to actual abundances and current effective population sizes.

While this rescaling assumes that the birth/death demographic changes in the number of individuals over time are proportional to the changes in the number of demes over time with strong migration, and that abundances are likewise proportional to effective population sizes even though these relationships are known to be complex (Luikart *et al*. 2010; Palstra & Frasier 2012). However, how σ changes the timescale of both forward and backward processes is not not determined in our model and therefore it is critical to determine if a chosen σvalue results in a model that can generate the observed data. As a check, one should assess the ability of the model to generate the data by statistical goodness of fit tests or model evaluation (Gelman, 2003; Lemaire, Jay, Lee, Csilléry, & Blum, 2016). Alternatively, σcould be treated as an unknown model and estimated given the data.

Given the parameters of the backwards time model (Tables 1.1 & 1.2), we use the

msPrime coalescent simulator program (Kelleher, Etheridge, & McVean, 2016) to generate genetic polymorphism data matching an arbitrary sampling regime of the island and/or mainland species pair sample sizes (with respect to numbers of individuals sampled at a mtDNA locus of length *L*). Instead of parameterizing the coalescent simulations of the *i*th species following the $\tau_i^{\tau_i}$ stochastic changes in effective population sizes since colonization according to N_i^j = { $N_i^{\tau_i-0}, \ldots, N_i^{\tau_i-\tau_i}$}, we use $\left(N_e\right)_i$, the harmonic mean of each species' effective population size across all time steps indicated by the $\tau_i^{\tau_i}$elements within N_i^j (Karlin 1968; Pollak 1983). One gene genealogy is simulated for each sampled species pair corresponding to a 570bp segment of the mitochondrial COI gene given an assumed invertebrate mitochondrial divergence rate (1.1% per species per million years; e.g. (Brower, 1994).

Initial conditions - We implement two different starting conditions to simulate volcanic versus continental island formation. Our initial conditions under the volcanic model deviate from those of Rosindell & Harmon (2013), in that at time zero they assume that one initial colonizing lineage consumes all available space in the community, thereby saturating *K*. In our model we select the most abundant species in the metacommunity and introduce one individual onto the unpopulated island. This initial condition is both biologically more realistic, and also avoids the assumption that volcanic island carrying capacity is saturated at time zero, which could generate unrealistic quantities of genetic diversity in the initial colonizing lineage. Continental islands are initially populated by making *K* independent random samples from the metacommunity proportional to their relative abundances. Here we are modelling a community of panmictic species that are simultaneously and instantaneously isolated on the island at time zero. Because we assume panmixia prior to isolation, the vector of colonization times (T_i^0) are initially identical

across the entire island community. Following subsequent local extinction and replacement by new colonizing species, the vector of colonization times T_i^j, becomes heterogeneous.

Quantifying equilibrium - Equilibrium is commonly defined as the dynamic balance between colonization and extinction rates that emerges over time, eventually leading to a stationary distribution where the two rates are expected to be equal (MacArthur & Wilson, 1967). However, under certain conditions, species richness and abundances may fail to equilibrate simultaneously, in which case the classic definition of equilibrium is insufficient (see Rosindell & Harmon 2013). To address the need for a more robust concept we follow Rosindell & Harmon (2013) in defining equilibrium as the point at which the starting conditions of the model are no longer detectable in the state of the system. In addition to colonization/extinction rate balance, this auxiliary definition guarantees that both richness and the SAD have reached their expected equilibrium values. Here we define a new term to measure the fraction of this equilibrium obtained by the community and treat it as a model pseudo-parameter that can be estimated by sampling the prior and posterior distribution, (Λ; Table 1.2). This quantity is defined as $\Lambda = (\sum_{i=1}^{K} E_i)/K$, where K is the carrying capacity and E is the boolean vector of length K such that E_i for $i = \{1, \ldots, K\}$ indicates the colonization status of each individual in the local community. When all individuals present in the local community are descended from a lineage that colonized during the simulation then $\sum_{i=1}^{K} E_i = K$ and $\Lambda = 1$. Our model of community assembly is inherently stochastic, so the amount of time for any given simulation to reach equilibrium is a random variable given the distribution under the model. For each forward time simulation we track

elapsed time, local community composition (both abundances and richness), and colonization times for all local species. We are interested in equilibrium and non-equilibrium dynamics, so we poll this information at regular intervals of arbitrary duration.

Summary statistics - At each time interval we extract the simulated sequences and calculate nucleotide diversity (π) within the local community for each sampled species given S_{local} (π_i= {π_1,...π_n}). We then construct a one dimensional histogram (Y) of local community genetic diversity such that:

$$S_{local} = \sum_{i=1}^{k} Y_i$$

where *k* is the number of bins (with *k*=10 for all simulation and empirical analyses), and bin width max(π_i)/*k*. We term this summary of local community diversity the one dimensional species genetic diversity distribution (1D-SGD). Next, we calculate absolute divergence (D_{xy}; (Masatoshi Nei, 1987) between each mainland-island sister pair (D_{xy_i} = {D_{xy_1}, … D_{xy_n}}). The values of π_i and D_{xy_i} are aggregated across all species-pairs sampled from the community within each time-point and summarized as a *k* x *k* joint frequency histogram (X) with equal-width bins such that:

$$S_{local} = \sum_{i,j=1}^{k} X_{i,j}$$

The upper bound for each dimension of the histogram is fixed to the maximum values of π and D_{xy} within a given simulation. We term this joint summary of community diversity/divergence as the two dimensional species genetic diversity distribution (2D-SGD). Additionally, at each time interval we record the rank abundance curve (RAC), the SAD, and Shannon's diversity index

calculated for the community (Boltzmann, 1872; Gorelick, 2006; Hill, 1973; Shannon, 1948). Given an observed sample of S_{local} species sampled from an empirical community, the simulated summary statistics are filtered to match the observed sampling configuration. As an additional method of comparison with the H' derived from the SAD, we also calculated the Shannon's index derived for both the 1D-SGD (π), and distribution of D_{xy} per sampling time point and notate this as H'_{π} and H'_{Dxy} respectively.

Simulation study design - To characterize the joint temporal dynamics of the SAD and 2D-SGD under non-equilibrium and equilibrium community assembly, we simulated assembly histories for both continental and volcanic islands, under a range of parameter values. These included varying local community sizes (K = 1000, 5000, 10000), colonization rates (c = 0.0001, 0.001, 0.01), and rate of post-colonization migration. We generated 10,000 replicated simulations for each combination of origin type, local community size, and colonization rate, resulting in 180,000 total simulated community histories. All forward time simulations were run for twice the mean time to turnover equilibrium (Λ) for the largest island with the smallest colonization rate ($5x10^9$ generations). We then summarized the temporal changes in H', π, D_{xy}, H'_{π}, and H'_{Dxy} by calculating the mean and standard deviation of each of these metrics for each parameterization across replicate sets of simulations at five values of Λ (0.1, 0.25, 0.5, 0.75, 1). For this initial set of exploratory simulation experiments, we calculated H' on the entire set of species while π, D_{xy}, H'_{π}, and H'_{Dxy} were likewise calculated on this entire set of S_{local} species given samples of 10 individuals per species in the local community and associated metacommunity source populations.

Bias and accuracy in estimating parameters - Next, we evaluated the suitability of H' and the relative bin magnitudes of the SGD as summary statistics for parameter estimation using

ABC by conducting a battery of leave-one-out simulation experiments under various ABC configurations (Table 1.3). We focus on evaluating accuracy and precision in estimating the following community-wide model parameters and pseudo-parameters: local community size (K), parameterized colonization rate (c), fraction of equilibrium (Λ), realized colonization rate (c'), extinction rate (†), and Shannon's diversity index (H'). We additionally explored estimation of community-wide parameters given various sequence and abundance data availability configurations (see Table 1.3). For example, given only the DNA sequence data sampled from a focal local community, the relative bin magnitudes of the observed 1D-SGD can be used as the summary statistic vector and both H' and Λ can be estimated, along with the other model parameters such as c, and † (ABC configuration M_1; Table 1.3).

To construct the reference table for the cross-validation analyses, we performed 1,000,000 community assembly simulations, sampling parameter values of K, c, and Λ according to uniform prior distributions (K = ~U(1,000-10,000), c = ~U(0.0001-0.01), and Λ = ~U[0, 1); see Table 1.1 for all simulation parameters). We then conducted ABC leave-one-out cross-validation using the *cv4abc* function of the *abc* R package (Beaumont et al., 2002; Csilléry et al., 2012; Lintusaari et al., 2017). For the ABC procedure we used simple rejection sampling and a tolerance sufficient to retain 1000 samples from the prior to construct the posterior estimate for each parameter of interest. We performed 100 leave-one-out cross-validation replicates per data configuration for each estimated parameter, and quantified accuracy of parameter estimation by calculating root-mean-square error (RMSE) and the coefficient of determination (R^2) for sampled and estimated parameter values.

Empirical application - Following our simulation experiments demonstrating that the ABC model can effectively estimate parameters, we perform an empirical analysis on a published

dataset from a community of spiders from the island of Réunion, an overseas department of France located in the Indian Ocean approximately 900 km east of Madagascar. In the original study, spiders were sampled from 10 lowland rainforest plots distributed across the island and sorted into 57 presumed biological species using a protocol combining morphological sorting and mtDNA sequencing (570bp Cytochrome Oxidase c Subunit I; Emerson et al., 2017). The dense sampling allows us to use both the H' calculated from the observed SAD as well as the 1D-SGD calculated from the observed sequence data for estimating assembly model parameters. Therefore, we use model configuration M_I to estimate H', and M_A, M_I, and $M_{AI}$to alternatively estimate Λ (Table 1.1). Under all model configurations we estimate parameters c' and †. For the ABC inference procedure, we simulated 1,000,000 samples by drawing parameter values from the same prior distribution used for the cross-validation analysis, and used the same rejection method to accept the closest 1,000 data sets to sample from the posterior distribution. When calculating π for each island taxon we used sample sizes with respect to numbers of individuals matching the observed spider data exactly with respect to numbers of individuals and length of DNA sequence.

We evaluated the overall goodness of fit of our posterior estimate to the observed data in two ways. First, we quantified the absolute Euclidean distances between the retained and observed summary statistics. Additionally, we performed a prior predictive check by projecting the retained simulated SGD, along with the observed SGD into principal component (PC) space. A good fit of the model to the data should generate simulated summary statistics sufficiently similar to those of the observed data as to be indistinguishable in the PC analysis.

Results

The joint SAD and SGD through time - The classically lognormal-like shape of the

SAD, with most species being of low abundance, is mirrored by a similar distribution of genetic diversities (Fig. 1.1). The shape of the joint spectrum of community genetic diversity (π) and genetic divergence (D_{xy}) generally widens over time as richness increases, while the corresponding H' of the SAD generally increases over the same time intervals (Figs. 1.1 & 1.2). We find that most species display low amounts of standing genetic diversity, as characterized by average pairwise differences (π), although there are important temporal dependencies as these characteristics only accrue with time as Λ progresses. On the other hand, time has a reduced impact on the distribution of D_{xy}, which obtains the lognormal-like shape even at very early stages of assembly, although with greater variability, as expected given that the final waiting times in the larger ancestral population will predict a large variance in this summary statistic, regardless of colonization time (Takahata & Nei, 1985).

Varying community-wide colonization rate (*c*), community size (*K*), and island origin (volcanic vs continental) also have characteristic impacts on components of the SGD and H'. Lower *c* resulted in a greater change in H' over time which was most apparent at lower *K* values and volcanic island settings, yet H' had the reverse trend under continental island settings. The mean values of π and D_{xy} tended to increase over time under the continental island settings whereas only the former tended to increase under the volcanic island setting. However, Shannon's index calculated on these two distributions of genetic diversity (H'_{π} .and H'_{Dxy}) both tended to increase over time under both island settings. Likewise, the mean values of π, D_{xy}, H'_{π} .and H'_{Dxy} tended to all increase over time regardless of the colonization rate (*c*) or community size (*K*), although the magnitude of change depended on these parameter values and island setting.

In the early stages of island community assembly, the 2D-SGDs from volcanic islands differ substantially from those of continental islands. There is a priority effect on volcanic islands

whereby the initial colonizing species quickly consumes all available niche-space as early arriving populations saturate the local carrying capacity. In this case the initial colonizing species have elevated π as well as high D_{xy}. However, this genetic signature of the early stage of community assembly quickly erodes as more species gain a foothold on the island, and as Λ approaches 1.0 a characteristic distribution of π and D_{xy} emerges. In contrast, the early stages of assembly in continental islands are characterized by uniformly higher values of π and D_{xy} which tend to decrease as Λ approaches 1.0. As Λ approaches 1.0, the SAD and 2D-SGD for both island origin models become indistinguishable.

Different colonization rates and local community sizes leave different signatures through time on the both the SAD and the 2D-SGD. Overall, higher colonization rates tend to increase the species richness in the community, predominantly by increasing the proportion of rare species, as well as species with lower π. Higher colonization also increases local extinction rates, and this increase in turnover decreases average divergence times, with a subsequent reduction in both π and D_{xy} in a higher proportion of sampled species. In a similar fashion, under reduced colonization rates, turnover is lower, the proportion of rare species is reduced, divergence times are longer on average, and π is increased on average.

Bias and accuracy in estimating parameters - Broadly speaking, cross-validation indicated reasonable accuracy and limited bias in estimating all parameters under all ABC model configurations, with the notable exception being ABC configuration $M_A$as well as attempting to estimate *K* under all ABC configurations. Under ABC model configuration M_I, ABC cross-validation indicated a strong signal in the data for estimating H' using only the 1D-SGD bin values as the summary statistic vector (Fig. 1.3; RMSE=0.26, R^2=0.96), with little added value when additionally including D_{xy} under M_{AI} (RMSE=0.27, R^2=0.95). Likewise, Λ could be

estimated well using ABC model configuration M_I and M_{AI} (R^2=0.68-72), yet using only H' as the lone summary statistic (M_A) resulted in poor conditions for estimating Λ (RMSE=0.28, R^2=0.05). Our joint framework additionally demonstrated accurate estimation of other ecologically important parameters governing assembly such as community-wide extinction rate (†), and effective colonization rate (*c'*), with R^2 between estimated and true values ranging from 0.61-0.89 under ABC model configurations M_I, M_{AI}, M_{MI} and M_{AMI}.

Estimating parameters for the Réunion spider community - For an empirical application, we chose to use only the 1D-SGD as observations to estimate H' calculated from the observed SAD (M_I). In this configuration the bin magnitudes of the 1D-SGD are treated as the summary statistic vector, and H' is treated as the parameter to be estimated. However, we also have the observed H' calculated from the samples for direct comparison to the estimate of H' under the ABC configuration M_I. In this case, our ABC mode estimate of H' = 1.816 (Fig. 1.4a; 95% HPD: 1.171-2.822) came remarkably close to the observed H' of 2.246 calculated from the sampled abundance data. This good fit of the posterior estimate to the observed H' indicates that the observed distribution of genetic diversity contains sufficient information about the community history of effective population size trajectories across island species with regards to predictions of the contemporary SAD under a neutral model of assembly (Fig. 1.3). Our simulation study demonstrates this possible dynamic as both H' and the SGD are predicted to increase over time under most conditions, such that our ABC model could potentially estimate the former with the latter given the strongly temporal features of our assembly model. Given the coupled dynamic of H' and the SGD as a progressive function of time in our simulation study, it follows that our ABC procedure has potential to estimate the degree of equilibrium parameter, Λ as shown in our cross-validation experiments. We estimated Λ for the spider community using three different ABC

model configurations configurations representing different combinations of H' and the 1D-SGD as summary statistics (M_A, M_I, and M_{AI}). Given M_A the mode estimate of Λ was 0.51 but with a diffuse posterior distribution (Fig. 1.4b; 95% HPD: 0.05-0.93). In sharp contrast, ABC configurations M_{AI} and M_I yielded mode estimates and HPDs that were both relatively clustered around high values of Λ (Fig. 1.4c & 1.4d; posterior mean 0.89; 95% HPD: 0.69-1). Additionally, ABC estimates of c' (Fig. 1.4e; posterior mean 0.001; 95% HPD: 0.0007-0.0017) and † (Fig. 1.4f; posterior mean 0.001; 95% HPD: 0.0009-0.0012) under model M_{AI}, were broadly concordant. More formal goodness-of-fit analysis with both the prior predictive check with principal components and Euclidean distances between retained and observed summary statistics corroborate the good fit of the model.

Discussion

The ETIB and its extension, the UNTB (MacArthur & Wilson 1963; Hubbell 2001) have a history of success predicting regional patterns of abundance distribution curves (Chust, Irigoien, Chave, & Harris, 2013; McGill, 2003), beta diversity (Chase, 2010; Chave & Leigh, 2002; Condit et al., 2002), phylogenetic patterns (Burbrink, McKelvy, Alexander Pyron, & Myers, 2015; Graham & Fine, 2008; Franck Jabot & Chave, 2009), and spatial structure in populations (Jordan, Barraclough, & Rosindell, 2016; Rosindell & Cornell, 2007). In recent years significant advances have been made to explicitly incorporate spatial relationships (Azaele et al., 2015; Gascuel, Laroche, Bonnet-Lebrun, & Rodrigues, 2016; O'Dwyer & Green, 2010), temporal dynamics (Engen, Solbu, & Sæther, 2017), species-area relationships (O'Dwyer & Cornell, 2017), and phylogenetic information (Cavender-Bares, Kozak, Fine, & Kembel, 2009; T. J. Davies, Allen, Borda-de-Água, Regetz, & Melián, 2011; Manceau, Lambert, & Morlon, 2015; Morlon, 2014; Webb, Ackerly, McPeek, & Donoghue, 2002). Here we have introduced a flexible framework

that brings community-level population genetic or comparative phylogeographic data into the realm of biogeographic community assembly. Reciprocally, our joint approach provides a way to ground comparative phylogeographic models using ecological and biogeographic neutral theory (Rosindell et al. 2011) rather than focusing on generic models of concordance and discordance (Papadopoulou & Knowles 2016). This approach of using ecological neutral theory to derive a testable null model with associated predictions of colonization times was recently explored in the context of biogeographic assembly of a gall wasp and parasitoid community in the western Palearctic (Bunnefeld et al. 2018).

From the perspective of community ecology, important progress has been made toward linking community ecology models with population genetics (Baselga et al., 2013; Baselga, Gómez-Rodríguez, & Vogler, 2015; Vellend, 2005)), with forthcoming opportunities for ecological theory to further incorporate the potentially powerful dimension of flexible comparative phylogeographic models (McGaughran 2015; Satler & Carstens 2017; Xue & Hickerson 2017). This should be facilitated by the increasing availability of genome-scale phylogeographic data that allows exploration of evolutionary models of increasing complexity and explanatory power (Schraiber & Akey, 2015), yet such approaches have seen limited use to infer the temporal and spatial dynamics at play at the community level (but see (Bunnefeld *et al*. 2018)). On the other hand, while many classic comparative phylogeographic studies attempted to infer histories of Pleistocene community assembly and diversification (Bermingham & Moritz 1998; Bernatchez & Wilson 1998; Hewitt 2000; Brunsfeld *et al*. 2001) by examining combined results of multiple single-taxon phylogeographic studies within a region (Emerson & Hewitt 2005; Emerson *et al*. 2011), most of these endeavors were not explicitly grounded in ecological assembly theory.

Even the explicitly comparative phylogeographic models that globally operate at the

assemblage level have yet to be grounded in ecological theory that can account for stochastic and deterministic forces underlying community assembly (Satler & Carstens 2016; Prates *et al*. 2016; Gehara *et al*. 2017). Fortunately, the community assembly models that generate expectations for temporally dynamic SADs (Missa, Dytham, & Morlon, 2016) and speciation/colonization rates (Rosindell & Harmon, 2013) could have an identifiable relationship with population genetic parameters like divergence times, admixture, expansion, colonization times, and changes in effective population sizes. Unifying the parameters of these two modeling frameworks could provide a new way of testing an array of competing assembly models with genetic data as well as estimating the relative strength of various deterministic forces underlying the assembly models such as niche filtering and competition. By explicitly linking ecological and micro-evolutionary processes whose dynamics and equilibrium expectations can occur on different time-scales, our new joint approach potentially allows for improved resolution and statistical power for estimating parameters as well as improved potential for and testing and fitting a number of different various neutral and non-neutral community assembly models (Vellend 2010). Likewise, understanding whether or not communities tend toward stable equilibria remains an unanswered question (Harmon & Harrison, 2015; Rabosky & Hurlbert, 2015; Valente, Etienne, & Dávalos, 2017; Valente, Phillimore, & Etienne, 2015) that can now be addressed with our joint approach that makes generative predictions of richness, abundance, and the spectrum of genetic diversity under both ecological and evolutionary time scales.

Assembly of the Réunion spider community - The joint data of mitochondria polymorphism and abundance structure from > 50 spider species on the volcanic island of Réunion affords us the opportunity to compare the estimate of the Shannon's index (H') using only the genetic data (i.e. ABC model configuration M_I) with the H' calculated from the observed abundance distribution. In this case, the posterior distribution of H' under M_I was able to

successfully recover the observed H'. If this is a general feature of our approach it would be encouraging given that estimating species abundances directly from field surveys can be difficult and problematic for some taxa (Kunin et al. 2000; Petrovskaya et al. 2012).

Using the distributions of abundance and genetic diversity jointly (M_{AI}) also allowed us to gain insight into the stage of progression towards equilibrium of this spider assemblage under our ecologically neutral model, yet the distribution of genetic diversity alone may have been sufficient (M_I). This was not the case of using the distribution of abundances alone (M_A), as the 95% HPD of the sampled posterior distribution for Λ under ABC configuration M_A was very wide, and heavily influenced by the prior (Fig. 1.4b), indicating there is little information about equilibrium state Λ solely from in H'. In contrast, the 95% HPD for both ABC configurations including the SGD (M_I amd M_{AI}) were significantly narrower and less influenced by the prior (Figs. 1.4c & 1.4d). This result is in agreement with the ABC cross-validation findings suggesting that estimation under ABC configuration M_{AI} improves accuracy and reduces bias in the estimation of Λ. It is notable that both ABC configurations including island genetic data (M_I amd M_{AI}) strongly indicate that this isolated spider community is consistent with an ecologically neutral assembly that is approaching or has reached equilibrium. Additionally, this assessment is supported by the similar mode estimates and largely overlapping HPD of *c'* (mode: 0.001, 95% HPD: 0.0007-0.0017) and † (mode: 0.001, 95% HPD: 0.0008-0.0012) which hews to the more traditional consideration of equilibrium as the dynamic balance of colonization and extinction. Indeed, Réunion island emerged from a classic volcanic hotspot formation approximately five million years ago (Gillot et al. 1994; Lénat et al. 2001), and this is likely sufficient time for equilibrium expectations of species richness, and community wide distributions of abundance and genetic diversity to have accumulated.

Although we do not sample any of the source sister species or sister populations from the mainland, the parameterization of the source meta-community remains under all ABC configurations. A related feature is that our model does not include *in situ* speciation in the local island community, yet because we do not collect data from the source species and do not use any phylogenetic information, *in situ* speciation is perfectly accommodated whereby the formation of new island species from pre-existing island species is parameterized as colonization from the source meta-community.

Outlook - The simple neutral model we introduce can be used as a candidate null hypothesis against which to test comparative population genomic/phylogeographic data, while the flexibility of the framework can be extended to accommodate various particular ecological contexts. For example, the model could explicitly incorporate *in situ* local speciation either as instantaneous events or as a protracted process (Rosindell, Cornell, Hubbell, & Etienne, 2010). Furthermore, it could incorporate non-neutral processes by including trait parameters for differential niche-filtering or dispersal limitation across species that result in variable colonization rates. In this case variation in colonization probabilities would be a proxy for non-neutral processes such as trait-dependent environmental filtering (Pigot & Etienne, 2015). Along these lines, the model could also accommodate deterministic processes such as resource-limited colonization probabilities or priority effects while retaining the stochastic dynamics of ecological drift underlying our joint model in the spirit of stochastic assembly theory (Tilman, 2004). In this case the magnitude of deviation from neutral expectations of colonization time, abundances, and genetic diversities could be modeled as a free parameter within our joint assembly model.

The increased complexity of these different modelling strategies would all benefit from the increased information content of higher resolution data types such as RADseq (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016), UCEs (Faircloth et al., 2012) or even whole genomes

(Bunnefeld et al. 2018). Thanks to algorithmic improvements, the coalescent simulator we incorporate here (*msPrime*) is capable of efficiently generating genomic-scale complex and arbitrary demographic histories under the full ancestral process of coalescence and recombination. Further, new developments in obtaining spectral summaries of short read SNP data across species (Xue & Hickerson 2017) or longer block-wise data to better accommodate recombination (Reddy et al. 2017) could be used within the ABC approach we develop here for parameter estimation or extended into a supervised machine learning for robust discrimination of complex neutral and non-neutral models (Schrider & Kern 2018).

For the joint modeling ABC approach we present here, we take the alternative approach to sampling deeply across the genome at the expense of only being able to accommodate limited numbers of taxa. While this results in more uncertainty due to only using single draws from the highly stochastic coalescent process per species, this noise is incorporated into our posterior estimates while gaining the borrowing strength from sampling potentially large numbers of taxa (Beaumont 2010) The widespread availability of mitochondrial and environmental DNA data also makes our approach amenable to model the assembly of complex microbial systems (Venkataraman et al., 2015) with time-series information (Capo, Debroas, Arnaud, & Guillemot, 2016; Ridenhour et al., 2017). Such time series data could introduce an additional axis of information allowing increased power to test hypotheses about the process of community assembly within a historical perspective.

From a practical standpoint, our model makes it possible to fit assembly models and estimate abundances from a small genetic sample of the community. An obvious advantage is that obtaining comparable DNA sequence data for a community of species can be logistically less challenging than obtaining reliably comparable abundance data. Taxa with high dispersal potential such as spiders are ideally suited for the estimation of SADs because their genetic

samples are more likely to have arisen from a panmictic coalescent process. While taxa with elevated levels of population structure might be more challenging for parameter estimation under our simple model, it could potentially be extended to accommodate in situ speciation within the local community, as well as explicitly modelling spatial processes (Haller & Messer 2017). Our model thus provides a flexible framework that can, even in the absence of comparable species abundance data, allow researchers to use the vast amounts of available mitochondrial DNA sequence data to test among competing models of island community assembly.

Chapter 2: Unifying the study of ecological communities across timescale

Introduction

Biodiversity in ecological communities accumulates in a hierarchical fashion across spatial and temporal scales (Leibold and Chase 2019). Fluctuations of species abundances within these communities operate on rapid ecological timescales, with periods of relative stability obtained over handfuls or tens of generations. Population genetic variation, by contrast, accumulates and degrades over timescales of tens to tens of thousands of generations, while phylogenetic and functional diversity accumulate even more unhurriedly, on the order of thousands to millions of generations (Uyeda, Hansen, Arnold, & Pienaar, 2011). Over time, various fields have emerged to investigate processes within individual levels of organization (macroecology, comparative population genetics, macroevolution), but only recently have inroads been made to combine theory across multiple levels of organization. Complicating matters, there is little consensus over whether, and to what degree, ecological interactions contribute to the structuring of ecological communities. Likewise, the relative contributions of colonization and *in situ* speciation to the composition of community structure remains an open question. Feedbacks across biological levels of organization are well known, yet we continue to lack a unified model of community assembly that accounts for such feedbacks, while incorporating the possibility of variable strengths of ecological interaction, as well as the continuum of the contribution of colonization and speciation to the accumulation of biodiversity.

Historically there have been two methods to investigate the impacts of evolutionary history and ecological assembly processes on community dynamics and macroecological patterns: 1) idealized complex simulation models that generate hypotheses about idealized community (Chesson, 2000; Gavrilets & Vose, 2005; Hubbell, 2001; MacArthur & Wilson, 1967;

Marquet et al., 2014; Tilman, 2004); and 2) empirical data investigated in a descriptive fashion that reveal aggregate differences in macroecological patterns from real world systems across a range of spatial and temporal scales (Craven, Knight, Barton, Bialic-Murphy, & Chase, 2019; Keil & Chase, 2019; R. E. Ricklefs & Bermingham, 2001; Rominger et al., 2016; Wagner, Harmon, & Seehausen, 2014). Recent advances in simulation-based inference under increasingly complex models provides a third option of unifying multiple processes and multiple data categories across different scales - we can use real multi-axis data to fit and compare competing models representing modes of community formation, from evolved to dispersal assembled, via various pathways. Several studies have recently shown that complex biological models and resultant high-dimensional data can be tractable within a machine learning framework (Schrider & Kern, 2018; Sheehan & Song, 2016), providing a robust inference procedure for simulation-based interrogation of empirical data.

Whether there are universal rules that structure ecological communities is a question of great interest, and there have been many previous efforts to investigate this. Inasmuch as one might subscribe to our formalization of the accumulation of biodiversity as a hierarchical process across timescales, previous approaches have tended to focus on one or at most two of these timescales (Leidinger & Cabral, 2017). For example, ecological models inspired by the Neutral Theory of Biodiversity and Biogeography (Hubbell 2001) have primarily focused on predicting the shape of the local species abundance distribution (SAD) under the assumptions of community equilibrium and/or stationarity. As central as the SAD is to macroecology and community ecology, it is often not sufficient to distinguish among different models of community assembly (Chave, Muller-Landau, & Levin, 2002; McGill et al., 2007). A great deal of work has been done to incorporate phylogenetic information with abundance data to make inference about community assembly processes (Webb et al. 2002, Jabot & Chave 2009). While such approaches make useful

predictions, they rely heavily on an assumption of equilibrium within the local community. Along another axis, recent important progress has been made toward linking community ecology models with population genetics (Baselga et al., 2013, 2015; Vellend, 2005); however, current theory either lacks an explicitly population genetic foundation (Vellend 2005), or considers genetic variation only of a focal taxon (e.g. Laroche *et al.* 2015). There have been other efforts to unify different time-scales with mechanistic eco-evolutionary models. For example (Cabral, Wiegand, & Kreft, 2019) unify population-level and evolutionary timescales to investigate the dynamic relationship between community age, competition, and local richness. Likewise, (Pontarp, Brännström, & Petchey, 2019) devise a trait-based, spatially explicit eco-evolutionary model to make inferences about prey and predator niche width with potentially diverse data types.

The shape of the species abundance distribution (SAD), as central as it is to macroecology and community ecology, is not sufficient to distinguish among different models of community assembly, even at equilibrium (Chave et al., 2002; McGill et al., 2007). As massive multi-dimensional datasets continue to emerge from next-generation biodiversity monitoring efforts applying community-wide surveying techniques such as eDNA (Deiner et al., 2017), metabarcoding (Andújar, Arribas, Yu, Vogler, & Emerson, 2018; Dopheide et al., 2019) and remote-sensing technologies that can directly infer trait data (Cavender-Bares et al., 2017), the challenges associated with moving beyond descriptive approaches of interpretation and inference have limited broader understanding of processes generating biodiversity patterns (but see Bohan et al., 2017; Derocles et al., 2018).

Incorporating temporal dynamics can help to distinguish among processes (Azaele, Pigolotti, Banavar, & Maritan, 2006; Chisholm & O'Dwyer, 2014; F. Jabot, Laroche, Massol, Arthaud, & Crabot, 2018; Kalyuzhny, Kadmon, & Shnerb, 2015; Nee, 2005; Robert E. Ricklefs,

2006), yet current theory fails to generalize across levels of biological organization. Modern high-throughput sequencing technology which facilitates community-scale metabarcoding efforts (Andújar et al., 2018; Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012), combined with hierarchical population genetic models of aggregate demographic histories, provide insights into temporal dynamics of the community assembly process (Overcast et al 2019a). Likewise, species trait data have been shown to be key in distinguishing ecological drivers of local diversity (Kunstler et al., 2012; McGill et al., 2007) and furthermore constitute the bedrock of comparative evolutionary (Harmon, Weir, Brock, Glor, & Challenger, 2008; Pennell & Harmon, 2013) and community phylogenetic (Ruffley et al 2019) approaches. Thus, a full model incorporating the totality of ecologically and evolutionarily relevant data axes remains to be described.

Here we introduce a mechanistic eco-evolutionary model of community assembly that builds upon classic community ecology theory (Hubbell, 2001; Leibold & Chase, 2017; MacArthur & Wilson, 1967; Vellend, 2016) to make historically dynamic joint predictions for observed data along three biodiversity axes, including species richness and relative species abundance (Rosindell & Harmon, 2013), genetic diversity and divergence (Overcast et al 2019), and trait evolution in a phylogenetic context (Ruffley *et al.* 2019). Specifically, we integrate ecological models of community biodiversity, comparative phylogeography, and community phylogenetics, with an explicit focus on incorporating microevolution and ecological interaction processes, which are often underrepresented in mechanistic models (Leidinger & Cabral, 2017). We combine summary statistics from these massive eco-evolutionary synthesis simulations (MESS) with supervised machine learning to test competing models spanning a continuum of community assembly and evolution (niche versus neutral and evolved versus assembled) and to estimate model parameters relevant to understand complex histories of community assembly and evolution. We perform extensive simulation-based cross-validation analyses to explore precision

and accuracy of model inference. Finally, we apply the model to four empirical datasets representing different temporal and spatial scales: weevils from the islands of Mauritius and Reunion (Kitson, Warren, Thébaud, Strasberg, & Emerson, 2018); spiders from Reunion (Emerson et al., 2017); trees from south-eastern Australia (Rossetto et al., 2015); and snails from the Galapagos Islands (Kraemer, Philip, Rankin, & Parent, 2019; Triantis et al., 2016). We find that distributions of genetic variation are more even than abundance distributions, and that, as communities approach equilibrium, the correlation between abundance and genetic diversity increases. Both of these phenomena are direct outcomes of the different timescales these diversity axes operate on.

Methods

Metacommunity composition - The MESS model comprises three components summarised in Figure 2.1. The metacommunity component consists of a metacommunity phylogeny relating all species, along with species abundances, and trait values evolved along the phylogeny. The global phylogeny is produced by simulating a constant birth-death process with fixed speciation (λ) and extinction ($\lambda \cdot \epsilon$) parameters, until the desired number of species (S_M) is reached (*TreeSim* v2.4; Stadler, 2019). Next, we simulate a Brownian motion model of trait evolution on the phylogeny with a root value of 0 and a rate of σ^2_M (*ape* v5.3; Paradis, Claude, & Strimmer, 2004). Traits evolve following a Brownian motion process in the metacommunity, rather than an Ornstein–Uhlenbeck process (Butler & King, 2004), because species in the metacommunity are not exposed to constraints imposed by the local environmental conditions. Additionally, we assume no trait variation among conspecific individuals in the metacommunity. Finally, the abundances of each species are sampled from a log-series distribution parameterized by the total number of species (S_M) and the total metacommunity size (J_M).

Local community dynamics - The foundations of the community dynamics underlying MESS are based on the joint neutral model of abundance and genetic diversity described in Overcast et al. (2019a). Briefly, we simulate an individual based model of community assembly inspired by the ecological neutral theory of Hubbell (2001), with assembly in a local community proceeding by a process of birth, death, and colonization from the metacommunity. Departing from the previous model, MESS local community dynamics can range from fully neutral (species traits have no effect), to various degrees of non-neutrality determined by the magnitude that species traits influence individual death probability (δ) through competition or environmental filtering. Following Ruffley et al. (2019), we based our environmental filtering and competition models on a functional relationship common in coevolutionary models which relates trait interactions with the probability of persistence in a community, scaled by the ecological strength (s_E; Lande 1976; Nuismer and Harmon 2015; Andreazzi et al. 2017). Calculated death rates per species are normalised to provide a vector of death probabilities that weight the random sampling of which individual will die in each time step according to a multinomial distribution.

As a first approximation, we implement a point mutation speciation process (Hubbell 2001), although other modes could be incorporated in future versions of the model (Rosindell et al., 2010). Speciation is implemented phenomenologically and takes place with probability ν upon each birth event. Upon each speciation event, the new individual is assigned a unique species identity, and its prior species identity is recorded as the parental for purposes of building the local phylogeny. The offspring species receives a new trait value sampled from a normal distribution centered on the parent species' trait value and with variance equal to $\sigma^2_M/(\lambda+\lambda \cdot \epsilon)$.

Population genetics component - Following Overcast et al. (2019), the forward-time histories of colonization and abundance changes through time per species are used to parameterize backward-time coalescent models with immigration for each species to generate

sampled local nucleotide diversities (π; Nei & Li, 1979). For reasons of computational efficiency, and to achieve a realistic scale in terms of numbers of individual organisms, we use a scaling parameter (α) to specify the number of individuals per deme, thus the total number of organisms in the local community is given by $J \cdot \alpha$. This notion of demes (or 'cohorts', groups of individuals that perform the same actions at the same time, see Harfoot et al., 2014) is conceptually similar to that of propagules from MacArthur and Wilson (1963), which they defined as "the minimum number of individuals of a given species needed to achieve colonization". We use the forward-time frequency of colonization events (scaled to number of colonizations per generation) for each species to parameterize the migration probability in the coalescent of colonization/divergence with ongoing immigration. Given an observed dataset, coalescent simulations match the observed sample sizes of each species for which DNA sequence data was obtained with regards to numbers of individuals per taxon and length of sequence.

Summary Statistics - We specify a hierarchical structure of summary statistics for each of the target data axes: species abundances, population genetic variation, and trait values. First, several relevant summary statistics are calculated per species, for each of the data axes. Next, each species-level statistic is aggregated and community-scale summary statistics are calculated per axis of data, capturing information about the distribution of the statistic across the community. We include as summaries the first four moments of each community-wide distribution, as well as pairwise Spearman rank correlations among all data axes. For correlations involving the trait axis, we consider the absolute value of the difference between the species trait and the local trait mean as the trait variable. We also calculate the differences between regional and local values of trait mean and standard deviation (Δ_{μ}^{trait} and $\Delta_{\sigma^2_{\square}}^{trait}$ respectively). Additionally, we utilize a framework of generalized Hill numbers as community-scale summary statistics, to

quantify the shape of each distribution (Chao, Chiu, & Jost, 2014; Gaggiotti et al., 2018). In order to distinguish between these diversity metrics when calculated on distributions of different data axes we will refer to the Hill number of order q for abundance data as ${}^{q}D$, for trait (functional) data as ${}^{q}FD$, and for genetic data as ${}^{q}GD$. For simplicity, throughout the manuscript we will refer to Hill numbers calculated on distributions of each data axis as abundance, π, and trait Hill numbers.

Model behavior - In an effort to investigate the behavior of the MESS model, we undertook a series of exploratory simulation experiments, with the aim of understanding how varying parameters of the model affect the distributions of community-scale data, and whether the chosen summary statistics capture information that could distinguish the degree to which differences in species traits (i.e., non-neutral processes) influence the structure of the community. Given that time is integral to the dynamics underlying the MESS model, we aimed to control for this in the first suite of simulations, with the goal of evaluating variability and overlap of summary statistics across assembly models at a fixed point in time. Temporal approach to equilibrium (Λ) is measured as the fraction of information about the initial state of the local community which is no longer present in the current state (see Overcast et al. 2019a) for a full treatment of this parameter). To control for temporal variation of summary statistics, Λ was fixed at 0.75 and we allowed ν to take one of three values corresponding to no-, low- and high-speciation (0, 5.10^{-4}, and 5.10^{-3} respectively). We generated 10,000 simulations for each assembly model (neutral/filtering/competition) using fixed parameter values of intermediate magnitude.

We were additionally interested in how summary statistics of different assembly model types vary through time (e.g. from early-, to middle-, and late-stage community assembly). To investigate this, we generated 10,000 simulations for each assembly model using fixed parameters of intermediate magnitude, allowing only ν to vary (taking one of three values as

above) and sampling communities at different stages of the assembly process ($\Lambda \sim U[0,1]$). Given the complexity and volume of data generated by these simulations, we summarize the results by plotting fitted least squares polynomial functions on all of the summary statistics for each model independently through time.

Machine learning inference power analysis and cross-validation - The MESS package includes an automated multi-stage machine learning (ML) inference procedure. Briefly, the MESS ML classification and regression procedures can be performed with a number of ensemble learning strategies including random forest (Breiman, 2001) and gradient boosting (Friedman, 2001). We quantify model uncertainty on parameter estimates as prediction intervals (PIs) using a quantile regression approach (Meinshausen 2006), and we implement posterior predictive simulations to assess the goodness of fit of the model to the observed data (Gelman 2003). Unless otherwise indicated, all ML algorithms are implemented in python using the architecture of *scikit-learn* (v0.20.3, Pedregosa et al., 2011).

We explored the power, accuracy, and bias of the ML inference procedure to classify community assembly models and estimate parameters using simulation experiments and cross-validation (CV). For assembly model classification, we generated 10,000 simulations per model class (i.e. neutral/filtering/competition) and fixed all MESS parameters at intermediate values, varying only the size of the local community (J) and the local speciation probability (ν). For quantifying the accuracy and bias of MESS parameter estimation utilizing an ML ensemble method regression framework, we generated 10,000 community simulations per assembly model class while varying several parameters of interest (α, J, s_E, m, ν, and Λ) using log-uniform or uniform prior distributions. ML estimator performance was then investigated using a K-fold CV procedure whereby simulations were split into training and testing sets, with the model being iteratively trained on each K-fold and performance being evaluated as minimized CV prediction

error on the held out training set. Classifier model adequacy was quantified by the percent error rate of misclassification, and regression model accuracy was quantified by the explained variance and R^2 (coefficient of determination) regression scores.

Empirical analysis - As a demonstration of the model, we selected four sets of local communities that are assumed to occupy different locations on the continuum of evolved/assembled and neutral/non-neutral assembly. Each system has some combination of community-scale data available for two of the three axes which can be considered by the model. In this way we hope to demonstrate the power of MESS across taxonomic and spatial scales, using data availability scenarios that might be encountered by empirical biologists in the present or very near future. Our empirical analyses include: 1) the spider community from Réunion island with a standardized sampling of ten 50 m x 50 m plots and 1282 individuals sequenced for one ~500bp mtDNA region (COI) (Emerson et al., 2017); 2) weevil communities from two Mascarene islands (Réunion and Mauritius) which have been densely sampled for abundance and sequenced for one mtDNA region (~600bp COI) at the community-scale (Kitson et al., 2018); 3) three subtropical rain forest tree communities scored for multiple continuous traits and shotgun sequenced for whole cpDNA (Rossetto et al., 2015) and ; 4) Galapagos snail communities collected from all major islands, sampled for one mtDNA region (~500bp COI; Kraemer et al., 2019) and scored for two continuous traits (Triantis et al., 2016). For each empirical dataset we conducted 10,000 simulations of each assembly model class and generated abundances, trait values, and genetic variation corresponding to genomic regions with identical numbers of base pairs under an infinite-sites model at a rate sufficient to generate diversity similar to the empirical data. We then conducted a round of ML model selection, parameter estimation, and quantile regression to generate parameter estimates and PIs. Finally, we implemented posterior predictive simulations to assess goodness of fit of the selected model and parameters to each of the observed

datasets.

Results

Model behavior and power analysis - Simulations generated under different community assembly models produced markedly different distributions of community-scale data which translates into perceivable differences captured by the summary statistics. In the first case we considered the behavior of the model by comparing summary statistics from 10,000 simulations under each community assembly model using fixed parameter values and sampling Λ at exactly 0.75 (Fig. 2.2). Neutral simulations generated communities with higher species richness, more even distributions of abundance as summarized by the normalized ${}^{q}D$ values, and higher mean and standard deviation of π values. Filtering and competition models were largely indistinguishable in terms of abundance and genetic diversity, with distributions of species richness, and mean and standard deviation of the population genetic statistics broadly overlapping (Fig. 2.2).

Distributions of statistics related to trait values showed more nuanced and variable behavior, obtaining characteristics that differ between the three models. There was little difference among distributions of Δ_{μ}^{trait}, with the exception that filtering models produced more variable results. However, trait variance distributions ($\Delta_{\sigma_{\square}^2}^{trait}$) varied considerably among models, with competition tending to yield negative values (more variation locally than regionally), filtering producing positive values (less variation locally than regionally), and neutral models producing values centered on zero. The trait Hill numbers (${}^{q}FD$) tended to be higher for neutral models, though the differences among models were more subtle. Looking at correlations between pairs of data axes provides further information. For example, abundance and trait values were strongly negatively correlated, indicating that species with traits far from the local optimum

tended to have low abundance. Also, abundance and π were positively correlated in the large part, indicating the tendency of more abundant species to harbor more genetic diversity, though there was a strong temporal dependency to this correlation.

Next, we investigated the temporal dynamics of MESS community histories by comparing 10,000 simulations for each assembly model using fixed parameters of intermediate magnitude and sampling a random Λ per simulation (Fig. 2.3). In agreement with the first round of simulations, species richness in neutral models tended to exceed that of the non-neutral models throughout the entire community assembly process. In general, a low rate of local speciation produced a slight increase in richness and Hill numbers for neutral simulations, whereas a high rate produced dramatic increases in these metrics for all simulation scenarios. Between non-neutral models, richness and Hill numbers for competition were, on average, always greater than those of filtering models across all timepoints, with differences increasing with v. For neutral models, ${}^{q}D$ tended to slowly increase monotonically through time, whereas ${}^{q}GD$ initially increased quickly with community-scale genetic diversity accumulating more slowly in later stages of assembly. Increasing v increased the average maximum ${}^{q}GD$ for non-neutral models, but in these simulations this maximum value tended to saturate very early, with little change through time. ${}^{q}FD$ demonstrated a more dynamic temporal trajectory. Broadly, the relationships among the tTrait Hill numbers (${}^{q}FD$) mirrored those of the abundance and π Hill numbers, with neutral models obtaining the highest, filtering the lowest, and competition somewhat intermediate values, and a trend of increasing values through time. However, one key difference in ${}^{q}FD$ is that early-stage communities display relatively high values, with values decreasing as Λ increases from 0 to ~0.2, and then showing an increasing trend as Λ proceeds from 0.2 to 1.

Model selection ML cross-validation - ML model classification prediction error reached a minimum value with J of 1000 for all model classes and all evaluated feature sets (Fig. 2.4;

mean error rate 0.16). Prediction error was slightly higher for small J (mean error rate 0.19), and did not improve dramatically when increasing J from 1000 to 2000 (mean change in error rate -0.02). Neutral simulations were more accurately classified than non-neutral simulations across all feature sets and ν values (mean error rate 0.05 and 0.18 respectively). ML classifiers trained using only summary statistics related to abundance and π produced highly accurate classification of neutral simulations (mean error rate 0.05), but failed to distinguish the two non-neutral models (error rate > 0.4). Importantly, in this condition the predicted model class for non-neutral simulations was overwhelmingly the alternative non-neutral model and rarely the neutral model. For example simulations under a competition model were misclassified as filtering (0.35) with a much higher rate than neutral (0.08). Including trait information along with one other data axis (either π or abundance) produced classification error rates approximately equal to error rates produced by models trained on the full suite of summary statistics.

Parameter estimation ML cross-validation - Cross-validation explained variance and R^2 regression scores for model parameter (α, J, s_E, s_C, m, ν, and Λ) estimation were broadly congruent and positive in almost all cases, indicating simulated and estimated parameter values were correlated (in some cases highly so). For neutral simulations Λ had the highest R^2 (0.963) and s_E the lowest (-0.037), with most parameters having moderate R^2 values (e.g. α = 0.567; m = 0.685; Fig. 2.5). The small R^2 for s_E is expected given that neutral simulations should have no information about strength of environmental interactions. Estimates of small to moderate values of m and ν were accurate, but larger values tended to be underestimated. ML parameter estimation for simulations of filtering and competition models obtained improved accuracy to estimate s_E (R^2 = 0.146 and R^2 = 0.287, respectively); however, R^2 values for other parameters were reduced with respect to the neutral simulations. Both non-neutral models produced diffuse estimates of α (R^2 = 0.205 and R^2 = 0.258) and J (R^2 = 0.398 and R^2 = 0.448). The most significant

difference between the non-neutral models concerned estimates of Λ. Under competition scenarios, Λ estimates were precise but upwardly biased between $R^2 = 0$ and 0.5, with increasing variance between Λ = 0.75 and 1. Under filtering scenarios, Λ estimates were only accurate for values close to Λ = 0.5, with decreasing accuracy as Λ moved away from this value in either direction.

Empirical Examples - We used three empirical examples to demonstrate how community assembly processes can be characterized and model parameters estimated using multiple axes of community-scale data. The ML classification procedure identified the neutral model as the most probable for all three Mascarene arthropod communities (Fig. 2.6a), with considerable support for neutrality of the Reunion spider community (predicted class probability 0.939), and more equivocal class probabilities for Mauritius and Réunion weevil communities (0.566 and 0.53, respectively). For the classification of communities which included data axes of abundance and genetic variation, the most important features for classification were ${}^{1}D$, standard deviation and mean of π, ${}^{2}D$, and ${}^{4}D$ (accounting for 44% of relative importance of all retained features).

The ML classification procedure identified environmental filtering as the most probable model for all tree and snail communities, with higher support for the snails (mean predicted class probability 0.698), and weak support for the trees (mean probability 0.440). Combining filtering and competition predicted class probabilities indicated the average probability of non-neutrality for the trees was 0.633, and for the snails was 0.865. Feature importance values for classification using axes of trait and genetic data were broadly diffuse across the retained summary statistics, with $\Delta^{trait}_{\sigma^2_{\square}}$ accounting for 11% of relative importance of all retained features, and the remainder accounting for 5% or less.

The ML regression procedure for parameter estimation indicated that the selected

empirical datasets occupied a broad swath of parameter space (Fig. 2.6b). Empirical PIs were quite varied, with some parameter estimate PIs spanning the width of the prior, while the PI of other parameters were narrow, a result which is consistent with CV results. The tree communities had small estimated α and narrow PIs (mean α = 1423; 1019-2481 95% PI), when compared to the arthropod and snail communities, which had larger estimated α (e.g. Mauritius weevil α = 7107; 3497-9831 95% PI). ML estimates of Λ were more varied, with the weevil and spider communities approaching or reaching Λ = 1, snail communities having more intermediate Λ, and tree communities having the lowest values (< 0.4 in all cases). Estimates of m and $v$displayed an idiosyncratic pattern, with spider and snail communities having low estimated values for both, and weevil and tree communities having high estimated values for both, with the exception of the Nightcap trees, which had high v and low m. Ecological strength (s_E) was the most difficult parameter to estimate, in the sense that all estimates were close to the mean of the prior, and PIs spanned the majority of the prior range. Posterior predictive simulations indicated a good fit of the estimated parameters to all empirical datasets.

Discussion

We have described an individual-based mechanistic model of community assembly, the MESS model, that unifies the key processes underlying the dynamics of local biodiversity across multiple timescales: 1.) dispersal; 2.) stochastic drift; 3.) deterministic competition/filtering; and 4.) speciation (Vellend 2010, 2016).The MESS model integrates these processes in The MESS model implements an hierarchical framework to make local multi-dimensional predictions of summary statistics that capture information both within and among the various axes of data. Generalized Hill numbers provide the unifying framework within which ${}^{q}D$, ${}^{q}FD$, and ${}^{q}GD$ values are comparable across communities of different spatial and/or temporal scales. Simulation experiments show that neutral models have elevated S, ${}^{q}D$, ${}^{q}FD$, and ${}^{q}GD$ with respect to filtering

and competition models, across all except the earliest timepoints (Fig. 2.3), a direct result of the ecological equivalence of individuals in neutral models generating communities with lower species dominance. In a similar fashion, for non-neutral models, species that are more fit survive preferentially and increase in abundance, reducing evenness in the community and causing ${}^{1}D$ to plateau at a low level. Increased speciation rate has little impact on ${}^{1}D$ in the neutral case because ecological equivalence confers no cost or benefit to offspring species, whereas in non-neutral models new species inherit the trait value of their parent (with small perturbation). In these conditions increasing speciation rate increasingly favors the evolution and accumulation of small clades of species that have ecological advantage, causing a concurrent reduction in ${}^{1}D$. ${}^{q}GD$ and ${}^{q}FD$ obtain broadly similar temporal dynamics.

Overall, we find that any two of the three data axes are sufficient to accurately identify the relative strength of deterministic versus stochastic processes in local community assembly, and that including trait information allows descrimination between which of the non-neutral processes are more important in driving the local patterns of biodiversity (Fig. 2.4). Additionally, using any two data axes always resulted in improved classification accuracy when compared to using a single axis alone. These results highlight the flexibility of MESS to mask unobserved summary statistics such that inference can be made from a wide variety of high-throughput biodiversity surveys across different spatial scales and data availabilities. MESS will perform best when provided data for all three axes, but it was designed to allow for incomplete and heterogeneous sampling (with some decrease in accuracy; Fig. 4), recognizing that some data axes are more or less difficult to obtain given different focal communities.

The empirical communities we chose to evaluate represent both a variety of available data axes, and a range of perceived dispersal limitation, with Galapagos snails being the most dispersal-limited, the Australian trees being least limited, and the Mascarene spiders and weevils

somewhat intermediate. As we assume the Reunion spider community is well mixed (i.e. panmictic within the island), the high probability of classification as neutral, and estimate of Λ approaching 1, along with the relatively high *m* and low *v*, are concordant with a late-stage community that is structured primarily by colonization and ecological drift (Barabás, D'Andrea, Rael, Meszéna, & Ostling, 2013; Vergnon, van Nes, & Scheffer, 2012). Both weevil communities had similar estimates of Λ, but higher estimated *v*, and more equivocal classification as neutrally evolving. The elevated *v* and partial weight of non-neutral classification could be a strong indication of cryptic diversity, which is in line with the expectation that the weevils are less dispersive compared to spiders. The snail communities were classified as being structured by environmental filtering, with low estimated *m* aligning with expectations. However, the low estimates of *v* and s_E are somewhat surprising, given their poor dispersal ability and documented pattern of single-island endemism (Parent & Crespi, 2006). In this case, unmodeled habitat heterogeneity, which is known to be an important predictor of snail diversity (Parent & Crespi, 2006), could easily artificially deflate estimates of *v* and s_E. Finally, because the Australian tree communities are sampled from semi-isolated habitat patches we expect their behavior to deviate from that of truly isolated communities. This is in agreement with the finding that these tree communities are all far from equilibrium, though the moderate *m* and high *v* and s_E estimates indicate that local turnover, in the context of a selective environment, is important and ongoing. Additionally, considering the fit of the tree data to a smaller α, the sample abundance in the scaled model and the (unobserved) 'true' abundance that better reflects the effective population size are more similar for trees than for the other datasets. More simply this could mean that sample abundance is closer to true abundance even though the former is unobserved.

The MESS model is an individual-based mechanistic model of community assembly that unifies processes relevant to the accumulation of biodiversity across ecological and evolutionary

timescales, incorporating dispersal, stochastic drift, deterministic competition/filtering, and speciation to generate joint predictions of abundances, population genetic diversities, and trait variation in a phylogenetic context. The model generates explicit temporal predictions of community-scale data across these three axes, spanning equilibrium and non-equilibrium conditions, and allowing for stochasticity along a continuum of scenarios ranging from pure ecological neutrality, to strong ecological interactions and/or environmental filtering. To complement the simulation framework of the MESS model, our implementation includes an extensive suite of ML tools for performing model selection and parameter estimation from observed data, and plotting routines for visualizing and evaluating results. This unified mechanistic model provides a general framework for hypothesis testing and biodiversity data synthesis, enabling scientists to generate multi-dimensional forecasts and test parameterized hypotheses about the historical and future processes driving biodiversity patterns from small-scale intensively sampled plots, to islands *sensu lato*, to regional and sub-continental scales.

Chapter 3: The spatial distribution of genetic variation in ecological communities

Introduction

The rapid development of remote sensing techniques which generate high-resolution environmental data along with the increasing spatial and taxonomic scale of integrated multi-dimensional data from high-throughput ecological surveys have transformed our ability to monitor the biosphere and understand the processes that underlie how communities are formed. Despite an accelerating data revolution driven by widespread deployment of technologies that can obtain community-scale data ranging from biodiversity metrics that quantify genes, traits, abundances, and ecosystem function at various levels of spatial granularity from local to global-scale, two fundamental bottlenecks have limited our ability to develop a more integrated understanding of how biodiversity accumulates within regional biotas: 1) lack of a mechanistic model for generating process based hypotheses of biodiversity structure (chapter 2); and 2) lack of a predictive model for extrapolating biodiversity structure from a limited sample to unsampled locations across the landscape. Here we will develop a framework to address the second challenge by using multiple heterogeneous data types collected locally at fine spatial scales to make spatial predictions of different axes of biodiversity. This will allow for a better understanding of how geophysical, climatic, and oceanographic features correlate with community structure at the regional scale. In conjunction with a mechanistic model that uses the same multi-axis biodiversity metrics to infer community assembly processes (chapters 1 & 2), we hope to be able to make broader inference of ecological processes across the planet.

Species distribution modeling (SDM) has been a highly successful endeavor to better understand abiotic and biotic determinants of single species ranges, how ranges change in the context of historical and/or future changes in climate and landscape as well as filling in the "Wallacean shortfall", (i.e. the incomplete information on species distributions; (Lomolino,

2004)) by means of SDM-based spatial predictions on the basis of correlations of known occurrences with environmental variables (Lozier, Aniello, & Hickerson, 2009; Phillips, Dudík, & Schapire, 2004). Extending this general correlative strategy to make global spatial predictions at the community or assemblage-level have made significant strides on several fronts. Species richness is one of the most well characterized biodiversity metrics with many examples of predictive models that correlate spatially explicit abiotic and biotic variables with richness at different spatial scales (Jetz & Rahbek, 2002; Kerr & Packer, 1997; Zellweger et al., 2016). Likewise, local richness can be predicted with correlative models similar to single species SDMs by way of stacked SDMs (D'Amen et al., 2015; Distler, Schuetz, Velásquez-Tibatá, & Langham, 2015) and joint SDMs (Harris, 2015; Ovaskainen, Roy, Fox, & Anderson, 2016), while local species abundance distributions can be similarly modeled and predicted by relating environmental variables with observed rank abundance distributions (Ellis, Smith, & Pitcher, 2012; McCarthy, Mokany, Ferrier, & Dwyer, 2018).

Along with the increasing availability of high-resolution spatial data, and the increasing complexity of spatial modelling tools, the availability of genetic data has been increasing along multiple axes, including greater sampling of loci within individuals, and greater sampling of individuals within populations (Taberlet et al., 2012). As sequence data continues to be more and more widely available, recent efforts have been made to move beyond local prediction from abundance or genetic samples, to global predictions of biodiversity structure (Miraldo et al., 2016; Pelletier & Carstens, 2018; Smith, Seeholzer, Harvey, Cuervo, & Brumfield, 2017). The advent of large-scale, curated databases of sequences (Genomic Observatories; N. Davies et al., 2014; Deck et al., 2017) allow for the possibility of making spatial and temporal predictions of occurrence and abundance for whole assemblages at regional or global scales (Crandall et al., 2019; Gratton et al., 2017). Additionally, large scale ecological monitoring projects, such as the

National Ecological Observatory Network (NEON) sites, offer the potential for replicated, high-throughput ecological surveying, which can help ground-truth methods developed for community-scale inference. Finally, the distribution of genetic variation at the scale of the entire community is becoming possible not only to obtain, but also to model.

As the biodiversity science and observation community is now moving to conceptualize and formulate an expanded set of essential biodiversity variables (EBVs) to enable better integrated and effective biodiversity monitoring, prediction, and inference (Jetz et al., 2019), a general approach to make spatial prediction across much of the planet that is logistically unreachable for intensive sampling is needed. The challenge here is to move beyond description and quantification of sampled biodiversity to actually enable prediction of biodiversity structure across the landscape. The inherent heterogeneity and sparseness of raw biodiversity data can be overcome by the use of models and remotely sensed covariates to inform predictions that are contiguous in space and time. The increasing availability of high-resolution spatial data, along with the increasing complexity of spatial modelling tools, and the increasing availability of population genetic sampling of sequence data at the community scale suggests the potential for a whole new kind of inference.

To this end, we offer here a novel approach that uses supervised machine learning to model the spatial relationships between suites of biotic and abiotic environmental variables (Title & Bemmels, 2018) and the structure of local distributions of species abundances and genetic diversities as summarized by a framework of generalized Hill numbers (Chao et al., 2014). This approach will have general applicability with the emerging efforts to advance remote sensing of biodiversity (Pettorelli et al., 2016; Turner et al., 2003) from the sky as well as on the ground intensive genetic biodiversity surveys (Porter & Hajibabaei, 2018; Valentini et al., 2016) that obtain ground-truthed assemblage-level data across multiple axes of the EBV hypercube (Miller,

1994). To demonstrate this approach, we use DNA sequence and abundance data of local decopod communities intensively sampled across the Indo-Pacific Coral Triangle (Al Malik et al., 2018; Kholilah et al., 2018; Knowlton & Leray, 2015; Pertiwi, Malik, & Kholilah, 2018) and abiotic (MARSPEC; Sbrocco & Barber, 2013) and biotic (Bio-Oracle; Assis, Tyberghein, & Bosch, 2018; Tyberghein et al., 2012) data layers for fine scale resolution of environmental variables for marine systems. We use the supervised machine learning (ML) via random forest (Breiman, 2001; Prasad, Iverson, & Liaw, 2006) to make spatial predictions of these two categories of Hill numbers across the coral seascape, as well as quantify pairwise site dissimilarity, and fit different non-equilibrium models of local community assembly that quantify levels of dispersal, speciation and magnitudes of ecological equivalency with regards to competition and environmental filtering (Overcast et al., 2019a; Overcast et al., 2019b).

Methods

Sampling design - The Coral Triangle is a volcanically and tectonically active region spanning 6 million km^2 in Southeast Asia and is a global hotspot of marine biodiversity (Bellwood, Renema, & Rosen, 2012; Hoeksema, 2007; Myers, Mittermeier, Mittermeier, da Fonseca, & Kent, 2000). Decapod communities were sampled from 136 dead branching corals of similar size collected at 10 sites widely distributed across the Coral Triangle (Table 3.1; Fig. 3.1). In order to reduce decapod community sampling variance we specifically targeted sampling from *Pocillopora* species. At each site between 5 and 32 (mean 12.7) dead coral colonies were sampled at approximately 10 m depth. Macro-organisms inhabiting each coral colony were removed and sorted following well described sampling protocols (Head et al., 2018; Plaisance, Knowlton, Paulay, & Meyer, 2009). Coral heads with fewer than 10 decapod samples were removed from the study prior to sequencing.

DNA barcoding, OTU clustering and calculation of summary statistics - Samples

were sequenced for one 660 base pair region of the Cytochrome Oxidase Subunit I (COI). Sequences were clustered at 5% similarity using MOTHUR (Schloss et al., 2009), to establish working hypotheses for operational taxonomic units (OTUs). Samples were sorted by hand and assigned to higher taxonomic rank (infraorders Anomura, Brachyura or Caridea), and where possible were identified to species level based on the results of BLAST searches of genbank. We constructed rarefaction curves for each sampling site in order to ensure approximately equal sampling effort, given probable differences in richness per site. As we are interested in the abundance and genetic diversity structure of communities, we aggregated abundances and calculated nucleotide diversity (π; Nei & Li, 1979) for each species per site. All downstream analyses were performed for each infraorder independently, as well as for the combined dataset, pooling all infraorders together. As all individuals of each infraorder within a site are presumed to compose an ecological community, results are reported primarily for the combined data, with key infraorder-specific results provided when relevant. All downstream analyses were performed on the full dataset, as well as on a subset of data with sampling rarefied to the site with the smallest number of samples. As results did not qualitatively change under rarefaction, here we report results only of the full data.

Characterizing diversity: Hill numbers for genetics and abundance - We quantify community structure using a framework of generalized Hill numbers, following a growing body of literature indicating their usefulness as a summary of high dimensional community data (Chao et al., 2014; Gaggiotti et al., 2018). The attribute diversity component of generalized Hill numbers have the form:

$$ {}^{q}AD = \left[\sum_{u \in S} v_u \left(\frac{a_u}{\sum_{h \in S} v_h a_h} \right)^q \right]^{1/(1-q)} \qquad \text{(Eq 3.1)} $$

where S is species richness, v_u is the attribute value for species u, a_u is the abundance of species u, and q is the order of the equation. ${}^{q}AD$ quantifies the relative frequency of species attribute values (in this case abundance or π) and is undefined for order 1, though a limit exists as q approaches 1 (see Chao et al. 2014). The ${}^{q}AD$ value is difficult to interpret directly and is not comparable across different data types, but it can be converted into an effective number of species or species equivalents:

$$ {}^{q}D = \left[\frac{{}^{q}AD\left(\sum_{u \in S} v_u a_u\right)}{\sum_{u \in S} v_u a_u} \right]^{1/\varphi} \qquad \text{(Eq 3.2)} $$

where $\varphi = 1$ for species diversity and genetic diversity. Hill numbers calculated in this way are not directly comparable across sampling locations, as different S will change their interpretation. To account for this, all Hill number values for all data types are additionally normalized by dividing by S, converting them to percentages and allowing for comparability across communities of differing richness. For simplicity, and to allow distinguishing between Hill numbers calculated on different data axes, we refer to Hill numbers calculated on abundance distributions as ${}^{q}D$, and on genetic diversity distributions as ${}^{q}GD$, for given values (q) of the order of the function.

Characterizing site dissimilarity - As a first exploration of community turnover across the landscape we calculated pairwise dissimilarity among sites for abundance distributions. For dissimilarity analysis we selected two different metrics, Bray-Curtis dissimilarity (Bray & Curtis, 1957) and cosine distance (Smith, Pontasch, & Cairns, 1990), as these capture different aspects of the data. Bray-Curtis (BC) is commonly used to quantify compositional dissimilarity of ecological communities, accounting for both species composition and abundance structure. Cosine distance is similar in spirit to Euclidean distance, but is insensitive to magnitude, and so provides a simple measure of species turnover that does consider differential abundances. For the

genetic data we quantified pairwise dissimilarity as the mean ($^{\mu}D_{xy}$) and standard deviation ($^{\sigma}D_{xy}$) of D_{xy} (Nei, 1987; Nei & Li, 1979) between all population pairs present in both sites under consideration. For example, to calculate $^{\mu}D_{xy}$ between Aceh and Solor we calculate D_{xy} between sampled species present in both sites, then take the average. In a marine environment, complex hydrological regimes may distort the classic distance-decay relationship among sites (Soininen, McDonald, & Hillebrand, 2007), therefore we investigated the relationship between geographic distance and community dissimilarity using *vegan* (Oksanen et al., 2010). Distance-decay analyses were performed using both abundance information, and presence/absence data. We also investigated the proportional contributions of turnover and nestedness to community composition using *betapart* (Baselga & Orme, 2012).

Abundance genetic diversity correlation - We investigated the abundance genetic diversity correlation (AGDC) by examining R^2 values of linear regressions between these data axes across different sites. As most species are rare, many have been sampled and sequenced for only a handful of individuals, complicating the calculation of π. To ensure our regression R^2 values were not impacted by the increased variance of small sample size, we calculated correlations for both the full data at each site, and for the subset of species for which there were more than 4 individuals sequenced. Additionally, as abundance and genetic diversity are on different scales, and are not normally distributed, we calculated correlations after log-transforming both axes, and also after rescaling all values into proportions.

Spatial environmental variables - We focus on the current community standards for global data sources with fine scale resolution for marine systems for abiotic (MARSPEC; Sbrocco & Barber, 2013) and biotic (Bio-Oracle; Assis et al., 2018; Tyberghein et al., 2012) data layers. MARSPEC comprises geophysical and bioclimatic data layers (e.g. average sea surface temperature, north/south aspect, or depth of the seafloor) at ~30 arcsecond (1km) resolution. Bio-

ORACLE provides layers at 5 arcminute resolution (~10 km) related to nutrient concentration, primary production, phytoplankton biomass, and several additional abiotic layers which overlap with MARSPEC. We additionally obtained bathymetric layers projected to the last glacial maximum provided by Paleo-MARSPEC (Sbrocco, 2014). All marine data layers were downloaded using the *sdmpredictors* R package (Bosch, Tyberghein, & De Clerck, 2017). We clipped data layers to a bounding box around the sampling locations with a 5 degree buffer in all cardinal directions (approximately 500 km). We further masked the climatic/environmental data using the Global Distribution of Coral Reefs (GDCR) shape data (https://data.unep-wcmc.org/datasets/1) provided by the UN Environment World Conservation Monitoring Centre's Ocean Data Viewer project (UNEP-WCMC, WorldFish Centre, WRI, TNC, 2010). Finally, we performed a principal component (PC) analysis, as a first exploration of the regions of environmental space that each sampling site occupies. We extracted bioclimatic and geophysical data for each sampling site, as well as for 1000 random background points at greater than -20m depth, projected these into PC space and plotted the first two PCs. We also plotted loading for each data layer, to evaluate the correlations between data layers, and how the layers contribute to environmental variation across the region. As a further exploration of spatial environmental variation we extracted environmental data for all GDCR sites, projected these into PC space and plotted the PC values on a map of the region, parameterizing the color of each GDCR site by its location in PC space.

Estimation of community neutrality and proximity to turnover equilibrium - We estimated the degree of ecological neutrality and the proximity to turnover equilibrium of communities at each sampling location using computer simulations and the ML infrastructure of the MESS package (Overcast et al. 2019b). We chose prior ranges on parameters for the MESS community simulations which were sufficient to reproduce patterns of richness, abundance, and

genetic diversity observed in the empirical data. We generated 10,000 simulations for each of three community assembly models representing pure ecological equivalence (the 'neutral' model), environmental filtering ('filtering'), and competitive exclusion ('competition'). We pooled all simulations and performed a cross-validation procedure to evaluate precision and recall of model classification given our priors and the simulated summary statistics. We evaluated the accuracy of the final trained ML classifier by plotting a confusion matrix of simulated versus predicted class labels for a hold-out set of test simulations. Finally, we used the trained classifier to predict assembly model class probabilities for each of the 10 sampling sites.

After establishing the most probable assembly model for each site, we undertook an ML regression procedure to estimate the community equilibrium state, and migration rate into the local community (the Λ and m terms in MESS, respectively). Again, we performed a cross-validation procedure to evaluate ML regression accuracy, using R^2 and explained variance scores as metrics. Finally, we split simulations into training and testing sets, trained a random forest regressor, and used it to predict Λ and m per site. Prediction intervals (PI) were constructed using a random forest quantile regression approach (Meinshausen, 2006). As sites may differ in their most probable assembly model, we performed cross-validation and trained ML ensemble regressors on simulations for each model class independently. All computer simulations were performed within the MESS framework, and custom ML architecture was constructed in python using *scikit-learn* (v0.20.3; Pedregosa et al., 2011). Machine learning classification and regression procedures for MESS simulations were performed with both random forest (RF; Breiman, 2001) and gradient boosting (Friedman, 2001) algorithms, and python code and jupyter notebooks are provided for both methods in the github repository. Results are reported only for RF methods, as we found these to produce more accurate classification, and higher cross-validation R^2 scores.

Predicting community structure across the landscape - With the dual goal of 1) learning associations between climatic/geophysical variables and community-level metrics at two different axes of biodiversity (species abundances and genetic diversities) and 2) making predictions of these biodiversity metrics in unsampled locations, we undertook a multi-stage random forest supervised ML parameterization, training, and validation procedure using climatic/ geophysical metrics as predictor variables and S, ^{1}D & ^{1}GD as response variables. Care must be taken when training a predictive model on small sample size data (Kirpich et al., 2018) as highly complex models will tend to overfit the data (Bzdok, Krzywinski, & Altman, 2017; Lever, Krzywinski, & Altman, 2016). As a first step we perform a feature selection procedure (Degenhardt, Seifert, & Szymczak, 2019) to retain only those variables that contain the greatest amount of information about the prediction targets, and to remove variables that are invariant, correlated, or uninformative (*boruta_py* v0.1.5; Kursa & Rudnicki, 2010; Speiser, Miller, Tooze, & Ip, 2019). Latitudes and longitudes were included as potential covariates, along with all MARSPEC and Bio-Oracle data layers. Next, we explored random forest regressor hyper-parameter space with cross-validation and a randomized search strategy to identify model parameters that maximized prediction accuracy. We evaluated estimator performance given our available data using 4-fold cross-validation with R^2 and mean absolute error (MAE) scores as the evaluation metrics (Fushiki, 2011; Kohavi, 1995). Finally, we generated a prediction set by sampling 1000 random sites as latitudes and longitudes falling within the GDCR mask. We extracted environmental and geophysical data as the estimator feature set, which we downsampled to retain only those features selected as above, and confronted with the trained ML estimator to predict ^{1}D & ^{1}GD for all 1000 sites in the prediction set. We constructed prediction intervals to quantify uncertainty around the most probable target values using a quantile regression approach (Meinshausen, 2006), and extracted feature importances to evaluate the

proportion of information contained within retained feature variables with respect to target variables of interest. Unless otherwise specified, all ML infrastructure was implemented in python, and built on top of *scikit-learn* (v0.20.3; Pedregosa et al., 2011).

Results

DNA barcoding, OTU clustering and calculation of summary statistics - The final dataset comprised 7572 decapod crustacean individuals sampled from 149 dead coral heads and sequenced for one 660bp COI region. After quality control, filtering, and OTU clustering the resulting database contained sequences from 685 OTUs across the infraorders Anomura, Brachyura, and Caridea. 57% of OTUs were identified to family level, 38% to genus level or below and the remaining ~5% identified to infraorder level. Coral head and individual decapod sampling effort were relatively even among sites, with on average 13.9 dead coral heads (8.7 SD), 146.1 OTUs (69.1 SD), and 702.9 individual samples per site (537.5 SD). Regional nucleotide diversity averaged 0.0028 (0.0007 SD). The number of individuals per site correlated with richness (Spearman rank correlation = 0.96; *p*-value = 7e-6), in agreement with theory. However, neither richness nor sampling effort correlated with average nucleotide diversity (p-value equal to 0.31 and 0.38, respectively), indicating a complex relationship between these data axes.

Hill numbers for genetics and abundance - As a very general trend, sites with larger ${}^{q}D$ tended to have larger ${}^{q}GD$ values, and the relationships among sites tended to remain the same across values of q (Table 3.2 & 3.3; Figs. 3.2a & 3.2b). For example, Pemuteran had the highest, and Karimunjawa the lowest, ${}^{q}D$ and ${}^{q}GD$ values across the spectrum of Hill numbers examined. Similarly, Kalimantan tended toward middling values for all q. Aceh provides an exceptional example, with very large values ${}^{0}D$ and ${}^{1}D$ but a precipitous reduction for larger values of ${}^{q}D$, yet retaining the second highest values of ${}^{q}GD$ for all q values. Another exception is Lembongan,

with middling to high values of ${}^{q}D$, yet the second lowest values across all ${}^{q}GD$. Normalized ${}^{q}D$ (Fig. 3.2c) and ${}^{q}GD$ (Fig. 3.2d) allow for a more direct comparison across different communities, as the values scale between 0 and 1, and represent a more direct notion of evenness (Tables 3.3 & 3.4). Normalized ${}^{q}GD$ were always higher than ${}^{q}D$ for all sites across all values of q, indicating a greater evenness in the genetic data. Several patterns are striking in the normalized data including Karimunjawa and Lembongan both having very high values for ${}^{q}D$, yet the lowest values for ${}^{q}GD$. Additionally, the disparity between abundance and genetic diversity at Aceh is clarified in the normalized Hill analysis.

Characterizing site dissimilarity - There was not a strong correlation between geographic proximity and either abundance or genetic dissimilarity. The geographic pattern of site dissimilarity was quite heterogeneous (Fig. 3.3). Compositionally there was high affinity among a cluster of eastern sites (Manado, RajaAmpat, and Solor), one central-southern site (Lombok) and Aceh, the most distant western site (cosine distances < 0.4). BarangLompo, though the most geographically central site, was the most compositionally distinct, with cosine distance > 0.7 for all sites except Kalimantan. Similarity within the eastern cluster was reduced, but still notable, when abundance was taken into account (mean BC = 0.55). Additionally, the affinity between Aceh and several of the eastern and central sites was reduced in the BC analysis, indicating that the abundance structures are more different than beta diversity would suggest. In terms of the genetic data axis, average ${}^{\mu}D_{xy}$ among all sites was 0.0083. Aceh was genetically most distinct with ${}^{\mu}D_{xy}$ values > 0.0114 in all cases except for Karimunjawa (${}^{\mu}D_{xy}$ = 0.0063). Karimujawa was also more genetically different than average (${}^{\mu}D_{xy}$ = 0.0095). Manado, Solor, Lombok, RajaAmpat, and Lembongan (and to a lesser extent Pemuteran) formed a cluster of genetic similarity, within which average ${}^{\mu}D_{xy}$ equaled 0.0055.

Abundance genetic diversity correlation - Correlation results were robust to various

subsampling regimes, for example selecting only the species with more than 4 individuals sequenced, therefore results are reported only for the full data. Correlation results were also robust to log-transformation and rescaling to proportions, so we report results with abundance and genetic diversity rescaled to proportions, as this eases the interpretation. Abundance and genetic diversity were positively correlated within all sites, considering both proportional and absolute values (Fig. 3.4; p-values all << 0.05). Correlations spanned a range of R^2 values across the sites, from 0.079 at Aceh, to 0.398 at Karimunajawa. In agreement with results of the Hill number analysis, we found far more rare species with greater proportional π than proportional abundance, and very few common species with proportional abundance greater than or equal to proportional π. There was not a clear correspondence between the AGDC within a site and its geographic location, though there was a conspicuous tendency for AGDC to increase with increasing proximity to the north east Java Sea/south Makassar strait, the geographic center of the Coral Triangle.

Spatial environmental variables - In total we used 45 environmental and geophysical data layers, including mean, minimum, and maximum contemporary bathymetry, average annual sea surface salinity, and profile curvature as a few examples. Additionally, we used the mean depth of sea floor at the last glacial maximum (LGM) from the Paleo-MARSPEC data. As all sample sites are ~6-10m below contemporary sea surface level, these were all well above sea level during the LGM, so we converted LGM bathymetric data into paleo-distance to sea shore, as a proxy for probable timing of recolonization history. Projecting environmental data from the observed sites along with random background sites from approximately equal depth shows the observed sites following a very strong gradient in PC space (Fig. 3.5). Inspecting the PC loadings showed the gradient was driven primarily by variables related to salinity, photosynthetically available radiation, concavity, and bathymetric slope, and to a lesser extent by distance to shore

and dissolved oxygen. Several variables contributed little to the variation in environmental space occupied by the observed sites, including those associated with sea surface temperature, chlorophyll A & calcite concentration. Sites within the Makassar strait (Kalimantan & BarangLompo) and the Java Sea (Karimunjawa) occupied more typical environmental space (more centrally clustered in PC space) whereas Eastern sites (Manado, RajaAmpat), and sites proximal to the lesser Sunda Islands (Solor, Lombok, Lembongan, Pemuteran) occupied more distal regions of PC space associated with increased salinity, pH, concavity, and bathymetric slope. Plotting GDCR environmental and geophysical PCs on a map of the region further illustrates gross regional landscape variation (Fig. 3.6), which largely recapitulates the results of the preliminary PC analysis. Several large geographic regions display environmental affinity, including the lesser Sunda islands with the eastern Banda sea, and the Mentawai Islands and Aceh with the Makassar strait. The north western Java sea, including Singapore and the islands of the western Riau Archipelago, occupy a unique and unsampled region of environmental space associated with increased chlorophyll concentration, diffuse attenuation at 490 nm, and annual variance in sea surface salinity.

Estimation of community neutrality and proximity to turnover equilibrium - Community assembly model classification with MESS computer simulations and random forest ML inference obtained several striking patterns (Fig. 3.7a). Evaluating feature importances for classification using all MESS summary statistics for both abundance and genetic diversity axes showed the majority of information with respect to model class was contained in ${}^{1}D$, ${}^{2}D$, ${}^{3}D$, ${}^{4}D$, and the standard deviation of π. Three eastern sites were classified as almost certainly neutral (Manado, Solor, and RajaAmpat; classification probability > 0.9). Aceh and BarangLompo were classified as most probably non-neutral (neutral probability < 0.2), with the bulk of probability assigned to the filtering model, but with some probability favoring competition. The rest of the

sites were classified with roughly equal probability for all three assembly model classes, resulting in equivocal support for all the models.

For estimation of the progress of each community toward turnover equilibrium (Λ), speciation rate (v), and migration rate into the local community (m), we subsampled the MESS simulations to retain only those simulations which belonged to the most probable assembly model from the classification step. For simulations with less than 50% support for one assembly model we retained all simulations during the parameter inference procedure. Feature importances for ML estimation of Λ indicated ${}^{2}D$, ${}^{3}D$, ${}^{4}D$ contained the most information with respect to this parameter, with all other summary statistics contributing less than 5% of feature importance. Feature importances for m were much more diffusely distributed among abundance and genetic diversity summary statistics, with significant weight placed on the correlation between abundance and genetic diversity (~20%), and approximately equal weight placed on the standard deviation of π, ${}^{4}GD$, ${}^{1}D$, ${}^{3}D$ and ${}^{4}D$ (~10% each). Aceh was predicted to have the smallest Λ (0.91), with BarangLompo and Karimunjawa also predicted to have low values (0.93 and 0.95 respectively). Eastern and Lesser Sunda Island sites were all predicted to have $\Lambda = 0.97$, indicating an increased proximity to turnover equilibrium. Most sites were predicted to have moderate migration rates (m), on the order of 0.007, with Karimunjawa having slightly lower m (0.006), indicating a largely homogenous migration regime region-wide.

Predicting community structure across the landscape - Random forest hyper-parameter tuning and feature selection identified east/west aspect, sea surface salinity variables, and distance to shore as the most relevant features for predicting ${}^{1}D$ and ${}^{1}GD$. Evaluating RF leave-one-out cross-validation for simultaneous estimation of normalized ${}^{1}D$ and ${}^{1}GD$ produced average mean absolute error of 0.05 (SD 0.03) and 0.026 (SD 0.016), respectively. Finally, we used abundance and genetic data from all sites to train a new model using previously selected

model hyper-parameters. We used this model to predict ${}^{1}D$ and ${}^{1}GD$ for all GDCR sites within the region, and plotted the predictions across the landscape (Fig. 3.7b & 3.7c, respectively), along with observed values from the sampled communities. As expected from inspection of the environmental PCs, north/south aspect, slope, depth of sea floor, and annual variance in sea surface salinity were the most important features for ML prediction. For the most part, values of ${}^{1}D$ from observed communities tended to closely match those of projected values of surrounding regions. Notable exceptions were Lembongan, with the highest scaled ${}^{1}D$ among samples (0.633), and Aceh with the lowest (0.375), both of which are surrounded by areas projected to have moderate ${}^{1}D$ values (0.45-0.5). Likewise, most sampling sites had values that were similar to those projected to sites in their near vicinity when considering ${}^{1}GD$, with the exception of Solor, which had very low ${}^{1}GD$ in an area projected to have higher values (~0.34 or higher). Plotting the difference between predicted ${}^{1}D$ and ${}^{1}GD$ showed a striking central/peripheral contrast (Fig. 3.8), with more similarity in evenness to the east in the Banda Sea and Banda Arc, and to the west in the central and northern Riau archipelago and the Mentawai Islands to the west of Sumatra. Regions with the greatest difference in evenness between ${}^{1}D$ and ${}^{1}GD$ tended to be peripheral (e.g. Manado and RajaAmpat), whereas reduced differences in evenness were concentrated around the Makassar strait, and the Bali and Flores Seas.

Discussion

Genetic diversity at the scale of the ecological community - Comparative phylogeography leverages the power of aggregated population genetic inferences of demographic history from multiple species in a community to answer fundamental questions about processes underlying community diversification, assembly, and macroecology (Hickerson et al., 2010; Papadopoulou & Knowles, 2016). Comparative phylogeography can be effective in the context of studies of community assembly because there are often large amounts of available data, along

with well-developed analytic pipelines to analyze these data. For example, mitochondrial data from > 100 neotropical avian taxon-pairs spanning several potential riverine and montane barriers was used to evaluate whether co-coversification was driven more by dispersal over pre-existing barriers than vicariance across historically emerging barriers (Smith et al. 2014). Next-generation DNA sequencing technology has also opened up new avenues of inquiry, such as a recent study that used genome-wide comparative phylogeographic data from a large number of desert lizard species to investigate whether estimates of effective population sizes correlated with observed abundances across species from this diverse taxonomic group (Grundler, Singhal, Cowan, & Rabosky, 2019). However, comparative phylogeographic methods typically lack a grounding in ecological theory, making it difficult to connect results with specific predictions of species interactions and coexistence. For instance, comparative phylogeographic methods have tended to focus on general models of shared demographic histories (Burbrink et al., 2016; Satler & Carstens, 2017; Stone et al., 2012), rather than models that are explicitly parameterized from ecological community assembly theory (but see Bunnefeld, Hearn, Stone, & Lohse, 2018).

On the other hand, the expanding field of community genetics has made some recent progress in this direction (Hersch-Green, Turley, & Johnson, 2011), and there have been some efforts to consider intraspecific genetic polymorphism within a dynamic non-equilibrium community assembly framework (Laroche, Jarne, Lamy, David, & Massol, 2015; Vellend et al., 2014), within statistical models of macroecology (Pelletier & Carstens, 2018; Smith et al., 2017), as well as characterizing the correlation between species diversity and the genetic diversity of a focal taxon in ecological communities (the species-genetic diversity correlation; Lamy, Laroche, David, Massol, & Jarne, 2017; Papadopoulou et al., 2011; Vellend, 2005) While positive species genetic diversity correlations are often expected, negative correlations are predicted from theory (Laroche et al., 2015) and have been observed in empirical systems (Marchesini, Vernesi,

Battisti, & Ficetola, 2018).

As emerging high-throughput ecological surveying technologies increasingly include ways of obtaining genetic information across entire local communities (Deiner et al. 2017, Krehenwinkel et al. 2018; Krehenwinkel et al. 2019a; Krehenwinkel et al. 2019b), these additional axes of biodiversity information can potentially improve efforts to infer the fundamental processes underlying community assembly and regional patterns of biodiversity (Vellend, 2010) across different spatial scales (Overcast et al., 2019a; Overcast et al. 2019b). However, even as these high-throughput surveys are deployed across large number of sampling plots with subsequent plans of following through as long term research studies (Lindenmayer et al., 2012), most of the range of any particular species will be missed. Therefore, even if techniques such as sample size adequacy and bootstrapping (Anderson & Santana-Garcon, 2015; DePatta Pillar, 1998) are used to quantify proper sampling regarding size and numbers of plots, logistic constraints and the spatial heterogeneity underlying any changes in species-specific attributes will complicate extrapolating aggregate biodiversity metrics into unsampled areas.

To help fulfill the extrapolative potential of high-throughput multi-dimensional ecological survey data, we describe here a way to make predictions of various assemblage-level biodiversity metrics across an unsampled or sparsely sampled landscape from small numbers of observed plots. This is accomplished using supervised machine learning the spatial relationships between suites of abiotic environmental variables (Title & Bemmels, 2018) and Hill numbers for local distributions of both species abundances and genetic diversities (Chao et al., 2014). This method is conceptually similar to gradient forests (Ellis et al, 2012), which have been previously used to identify environmental predictors that correlate with community composition turnover, or with turnover of genetic variation within species (Fitzpatrick & Keller, 2015). Here we unify both these approaches to make predictions of genetic turnover at the scale of the entire community. We

demonstrate this on regional-scale sampling of decapod crustacean communities from across the Coral Triangle. After describing the observed spatial patterns of local distributions of abundance and genetic diversities, and comparing sampled areas using dissimilarity metrics, we subsequently use the multi-dimensional data to fit different ecological assembly models before using our method to make spatial predictions into unsampled areas across the landscape.

Predicting community structure across the landscape - The Coral Triangle is a global biodiversity hotspot, yet little is known about the spatial and taxonomic distribution of the vast majority of macroinvertebrate diversity (Plaisance, Caley, Brainard, & Knowlton, 2011; Plaisance et al., 2009)**.** Previous work has shown that there is no relationship between abundance and species co-occurrence, and no evidence for covariation of species densities (Gotelli & Abele, 1983). Thus, it remains to be discovered whether abundance and genetic diversity scale with environment or whether they are decoupled in this system. Additionally, we do not know how or whether extreme hydrological regimes contribute to or impact community structure. Here we apply our new ML method to a sample of community-scale abundance and genetic data from Coral Triangle decapod communities to untangle the relationship between environment and community abundance and genetic diversity structure, and to make predictions for such data axes across the entire region.

Mean absolute error on predictions of abundance and genetic diversity Hill numbers of held out samples during leave-one-out cross-validation were quite low, indicating our ML model identified some signal in the data associated with environmental variation, even given such a small sample size. As environmental and geophysical characteristics of the region are quite dynamic, abundance and genetic diversity of decapod communities are distributed heterogeneously across space, yet are largely correlated with each other, with notable exceptions. Overall, variables that negatively correlated with proximity to deep water were the strongest

drivers of increased abundance and genetic diversity evenness, including reduced annual variance in sea surface temperature, increased sea surface salinity, and increased nutrient concentrations. We found a strong signal of increased evenness for both abundance and genetic diversity in the Java Sea and Makassar strait, potentially indicating an environmental buffering effect in regions proximal to the Sunda Shelf, and edge effects in peripheral regions, for example in the eastern region of the Banda Sea or the Mentawai Archipelago to the west.

Hill numbers for genetics and abundance - Our results demonstrate several striking findings. First, genetic diversity is more evenly distributed within ecological communities than is abundance. This finding is in line with simulations of local ecological communities from Overcast et al. (2019b) showing greater variance in abundances over time. This highlights the fact that abundance and genetic diversity accumulate over different timescales, and that these axes of data contain orthogonal information about the history of accumulation of biodiversity within ecological communities. The distributions of abundance and genetic diversity are broadly similar among sites. However, Aceh provides a striking and interesting counterexample. The abundance distribution within this community demonstrates strong dominance, with ${}^{q}D$ eroding precipitously with increasing q values, yet the distribution of genetic variation remains much more even, which is in line with our knowledge that Aceh has been recently invaded by one species which is dominating local community abundance (C. Meyer, personal communication). This finding demonstrates the power of examining multiple dimensions of biodiversity simultaneously.

The abundance genetic diversity correlation - If abundance and genetic diversity were partitioned in a linear fashion then we would expect all species within a site to fall on or close to the identity line in a plot of AGDC. Likewise, if abundance and genetic diversity were partitioned randomly then we would expect a fairly uniform scatterplot in AGDC space. However, what we observe in the decapod crustacean data is quite unique and distinct from either of these null

expectations. For Coral Triangle decapod communities we observe that most species are rare (5% of proportional abundance) and have proportionally low (< 5%) genetic diversity (Fig. 3.4). Species with more than 10% of proportional abundance tended to constitute less than 10% of proportional genetic diversity, and species with more than 10% of genetic diversity tended to constitute less than 10% of abundance. This is in line with theory and recent findings (Overcast et al 2019a; Overcast et al 2019b) which propose that rare species can obtain a complex abundance history, occupying the landscape for enough time to amass considerable genetic variation, yet remaining proportionally rare, and that abundant species may have only recently entered the community and risen to large numbers by chance, without having time to accumulate much genetic variation.

Estimation of community neutrality and proximity to turnover equilibrium - Model classification cross-validation has previously shown that neutral and non-neutral assembly model classes can be readily identified using genetic and abundance data (Overcast et al. 2019b), however reliably distinguishing between competition and filtering models requires information about species traits, which are presently unavailable for these decapod crustacean communities at this time. However, the strong estimation of neutrality for the eastern sites (Manado, RajaAmpat, Solor), while the rest are more equivocal is a very tantalizing outcome. Additionally, the finding of a low estimated Λ for Aceh may not be surprising, given what we know about its invasion history, and given the fact that MESS inference can not presently account for disturbance in estimation of Λ. In other words, it is difficult to distinguish between a scenario where a community has not naturally obtained equilibrium and one which has obtained equilibrium and then undergone a recent disturbance. Additionally, in general, patterns of estimated values of the MESS *m* and *v* parameters were broadly similar across sites, with *v* values much smaller than *m*, indicating a stronger contribution of dispersal than speciation to the structuring of the

communities.

Future directions - Taking into account the strict environmental tolerances of the coral host, prediction of decapod crustacean community diversity may be improved by increasing the resolution of spatial environmental and geophysical data along a number of axes. For example, constructing geomorphons from bathymetric data can capture more robust information about the shape of the sea floor, beyond simple aspect and slope (Jasiewicz & Stepinski, 2013). Higher resolution bathymetric data can also be used to increase the resolution of our current biotic and abiotic data layers through an interpolation procedure (e.g. ANUSPLIN; Hutchinson, 1998). Whereas in this study we only consider linear geographic distance, it may be critical to more explicitly model connectivity among sites by accounting for the regional hydrological regimes (Monismith, 2006; Werner, Cowen, & Paris, 2007). Finally, targeting undersampled regions of environmental space for future collecting efforts (e.g. the western Riau Archipelago) will allow for both improvement and validation of the model.

Our method will easily generalize to any system where community-scale abundance and genetic data are being gathered (Likens & Lindenmayer, 2018; Lindenmayer et al., 2012; Reinke, Miller, & Janzen, 2019). Indeed, terrestrial systems should be able to use our method, in concert with higher-resolution spatial data, and the variety of remote sensing products now becoming available, to make even more powerful predictions. Especially promising are intensive surveys that capitalize on eDNA techniques (Bálint et al., 2018; Taberlet et al., 2012) within which population genetic information can be gained along with species occurrence and abundance information (Adams et al., 2019; Deiner et al., 2017; Grummer et al., 2019) and even various biodiversity metrics from historical communities (Epp, 2019). Genetic diversity at the scale of the ecological community is a powerful axis of data which records the history of the community at a population genetic timescale, the investigation of which is still in its infancy (Overcast et al

2019a; Overcast et al 2019b). Here we contribute to the continuing investigation by demonstrating how the spatial partitioning of community-scale genetic variation can be decomposed into environmental correlates and used to make predictions across the landscape. Further investigations of this new axis of data will continue to shed new light on how communities assemble.

Table 1.1 gimmeSAD model input parameters
Model parameters, the definition of each parameter, and the values explored in the simulation analyses. Identical parameter values were applied during analysis of the empirical Réunion spider dataset. For the simulations, the sequence length and mutation rate (μ) were chosen to correspond with values for these parameters that are typical for arthropod mitochondrial DNA datasets.

Parameter	Definition	Values used in simulation experiments
K	Local community size	~ uniform(1000-10000)
c	Probability an empty deme is replaced by a colonizing individual sampled from the metacommunity (colonization rate)	~ log-uniform(0.0001-0.01)
μ	Mutation rate	.011 $base^{-1}$ $species^{-1}$ My^{-1}
σ	Abundance scaling factor	100
L	Simulated Sequence Length	570
S_{meta}	Number of species in the metacommunity	1000
A_{meta}	Abundances of species in the metacommunity	~ logseries(p=0.98)
n	Number of genetically sampled individuals per species at local and metacommunities for coalescent simulations	10

Table 1.2 gimmeSAD model response variables
Variable names, definitions, and the dimensions of each variable used in the framework.

Variable	Definition	Dimensions
S_{local}	Number of species in the local community (i.e. local richness)	Unbounded positive integer
A_i^j	Time-dependent abundances on the island community	$S_{local} \times \tau^j$matrix
$A_i^j \sigma = N_i^j$	Time-dependent effective population sizes	$S_{local} \times \tau^j$matrix
M_i	Post-colonization migrants	S_{local} vector
N_e	Harmonic mean of Time-dependent effective population sizes	S_{local} vector
T_i^j	Colonization time vector	S_{local} vector
A_{meta}	Abundances of species in the metacommunity	S_{meta} vector
Λ	Fraction of equilibrium obtained	Continuous [0, 1]
H'	Shannon's index of diversity	Continuous value > 0
$\dagger$	Effective Extinction rate	Continuous [0, 1]
c'	Effective colonization rate	Continuous [0, 1]

Table 1.3 gimmeSAD ABC model configurations
An overview of the five different model configurations explored, indicating summary statistics derived from available observed data (π, D_{xy}, and/or H') and the parameters to be estimated under our ABC framework. These scenarios arise from various combinations of observed abundances (A), community-scale nucleotide diversity (I), and island-mainland divergence (M). The Shannon's index (H') can be configured either as a summary statistic or as an estimated pseudo-parameter, depending on whether densely sampled abundances are available for the community of interest. Other pseudo-parameters ($c, c', K, \dagger, \Lambda$) can be estimated under all ABC configurations.

Model Configuration	Summary Statistic Vectors			Estimated Pseudo-Parameters
M_A			H'	$c, c', K, \dagger, \Lambda$
M_I	π			$c, c', K, \dagger$,H' and Λ
M_{AI}	π		H'	$c, c', K, \dagger, \Lambda$
M_{MI}	π	D_{xy}		$c, c', K, \dagger$,H' andΛ
`M_{AMI}	π	D_{xy}	H'	$c, c', K, \dagger, \Lambda$

Table 3.1 Coral Triangle sampling site geographical coordinates
Sampling site names used throughout this study and exact location information provided in degrees latitude and longitude.

Site	Latitude	Longitude
Pemuteran	-8.1159	114.62
Lembongan	-8.66159	115.461
Aceh	5.51136	95.16113
RajaAmpat	-0.59425	130.58079
Karimunjawa	-5.80239	110.37389
BarangLompo	-4.96478	119.28472
Solor	-8.4945	123.0762
Lombok	-8.73642	115.88083
Manado	1.60661	124.73667
Kalimantan	0.06514	117.55861

Table 3.2 Hill numbers for abundance

Hill numbers 0 through 6 for abundance distributions for each of the ten Coral Triangle decapod crustacean communities.

	Aceh	Barang Lompo	Kalimantan	Karimun jawa	Lembongan	Lombok	Manado	Pemuteran	Raja Ampat	Solor
0	208	92	118	63	102	129	105	315	169	153
1	78	50	66	34	65	69	58	143	89	68
2	29	25	43	22	42	41	34	77	49	38
3	17	16	33	17	32	30	25	52	34	28
4	13	13	28	15	27	25	21	40	28	24
5	11	11	25	14	24	23	19	34	25	22
6	10	10	23	13	22	21	18	30	23	21

Table 3.3 Hill numbers for genetic diversity

Hill numbers 0 through 6 for genetic diversity distributions for each of the ten Coral Triangle decapod crustacean communities.

	Aceh	Barang Lompo	Kalimantan	Karimunjawa	Lembongan	Lombok	Manado	Pemuteran	Raja Ampat	Solor
0	96	51	65	36	50	64	52	161	85	68
1	73	39	49	25	34	44	39	117	64	47
2	60	33	41	20	27	34	31	96	53	37
3	52	29	37	16	23	29	26	83	47	32
4	47	26	34	14	21	26	23	75	43	29
5	43	24	31	13	20	24	20	69	39	27
6	40	23	30	12	19	23	19	64	37	26

Table 3.4 Hill numbers for abundance scaled by species richness per site

Hill numbers 1 through 6 scaled by species richness for abundance distributions for each of the ten Coral Triangle decapod crustacean communities.

	Aceh	Barang Lompo	Kalimantan	Karimunjawa	Lembongan	Lombok	Manado	Pemuteran	Raja Ampat	Solor
1	0.375	0.54	0.563	0.543	0.634	0.533	0.555	0.455	0.526	0.444
2	0.142	0.275	0.365	0.342	0.416	0.315	0.326	0.245	0.29	0.247
3	0.081	0.176	0.28	0.271	0.315	0.233	0.242	0.165	0.204	0.186
4	0.061	0.137	0.238	0.24	0.266	0.197	0.205	0.128	0.167	0.16
5	0.053	0.119	0.213	0.223	0.238	0.177	0.184	0.107	0.147	0.146
6	0.048	0.108	0.198	0.212	0.22	0.164	0.171	0.095	0.135	0.13

Table 3.5 Hill numbers for genetic diversity scaled by species richness per site

Hill numbers 1 through 6 scaled by species richness for genetic diversity distributions for each of the ten Coral Triangle decapod crustacean communities.

	Aceh	Barang Lompo	Kalimantan	Karimun jawa	Lembongan	Lombok	Manado	Pemuteran	Raja Ampat	Solor
1	0.757	0.767	0.759	0.704	0.673	0.694	0.745	0.727	0.753	0.696
2	0.628	0.64	0.638	0.543	0.534	0.539	0.602	0.595	0.628	0.551
3	0.547	0.561	0.566	0.451	0.467	0.455	0.506	0.516	0.552	0.475
4	0.491	0.509	0.518	0.395	0.427	0.406	0.439	0.464	0.5	0.431
5	0.45	0.472	0.484	0.36	0.401	0.375	0.393	0.427	0.463	0.401
6	0.42	0.446	0.459	0.336	0.383	0.354	0.36	0.4	0.435	0.381

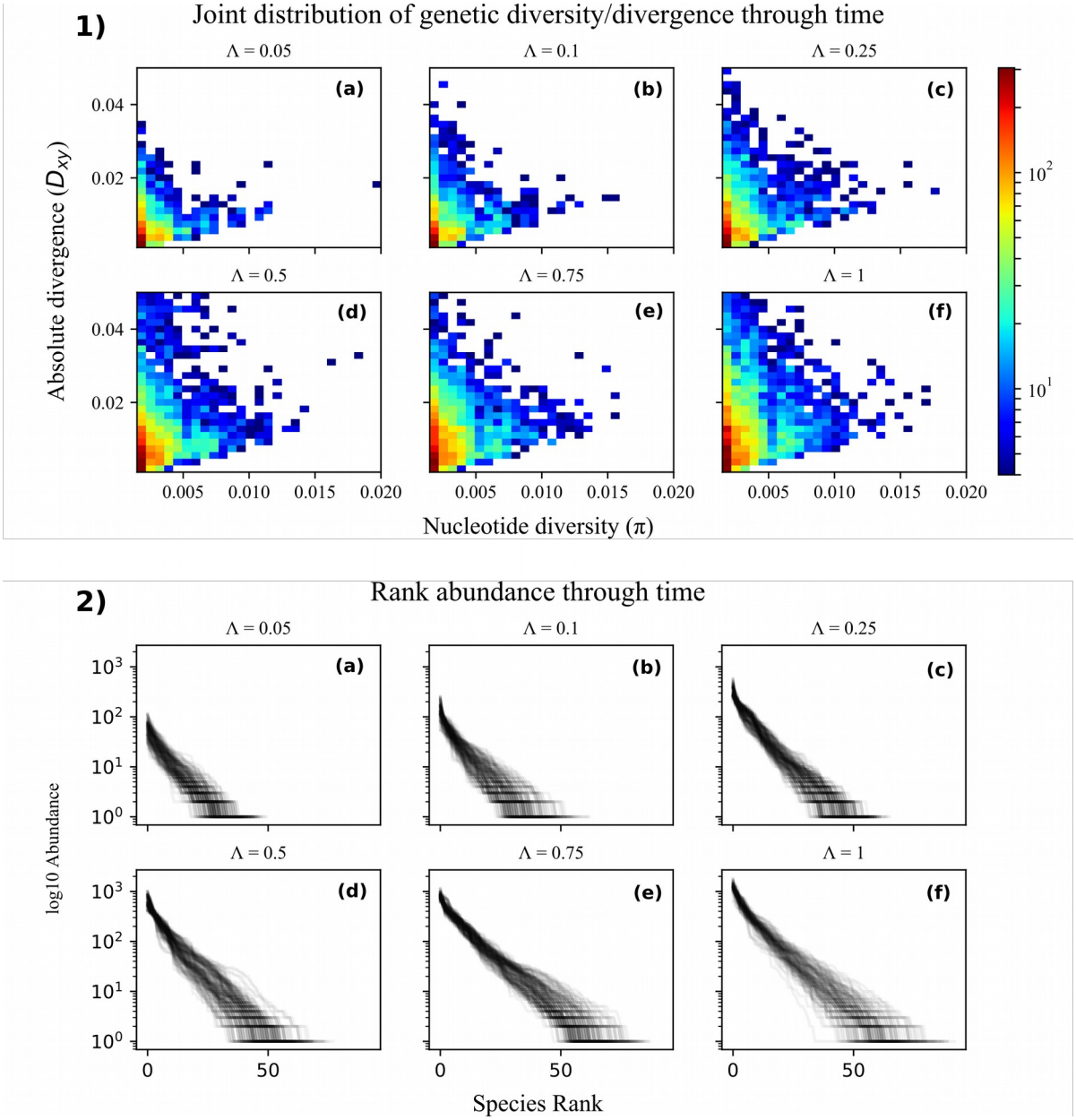

Figure 1.1 2D-SGD and corresponding rank abundance at varying stages of community assembly

Panel 1) Summed aggregations of the 2D-SGD across $1x10^4$ replicated simulations at varying stages of community assembly. All simulations were conducted with intermediate values of community size and colonization rate (K=5000, c=0.03). Each point in the plot is a joint frequency bin for values of local nucleotide diversity (π) and absolute genetic divergence (D_{xy}). The color of each bin indicates the number of species it contains, with cooler colors signifying fewer species and warmer colors signifying more species. Panel 2) Corresponding rank abundance plots of the $1x10^4$ simulated communities. Values of Λ depicted capture multiple stages of community assembly from early (0.05, 0.1), through middle (0.25, 0.5), to late (0.75, 1).

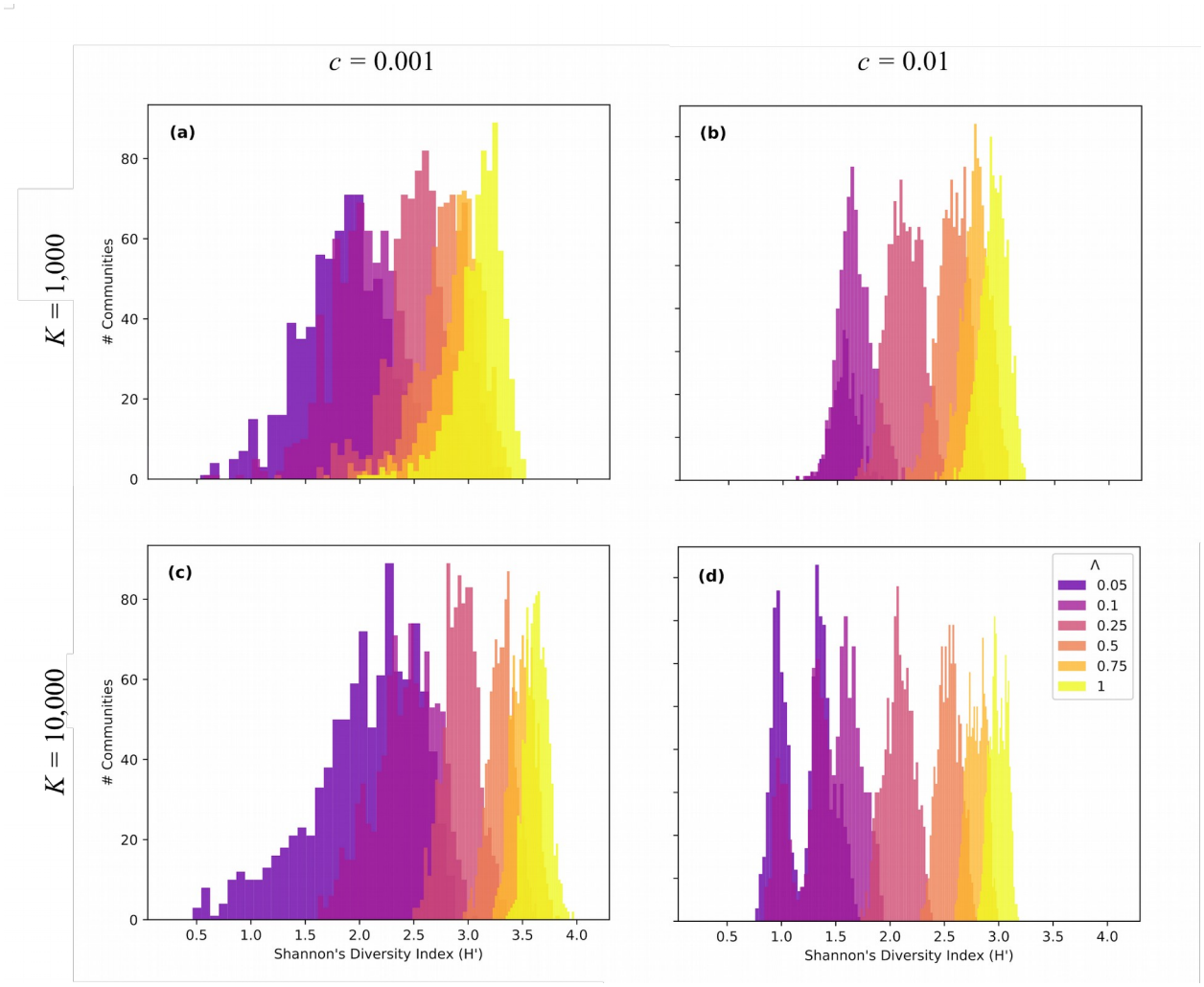

Figure 1.2 Shannon's diversity index at varying stages of community assembly
Histograms of Shannon's diversity index (H') for four different community parameterizations including low and high colonization, and small and large community sizes. $1x10^4$ independent simulations were performed for five Λ values for each parameter combination. Depicted are a) Low colonization rate, small community size; b) High colonization rate, small community size; c) Low colonization rate, large community size; d) High colonization rate, large community size. A range of Λ values were used to capture multiple stages of community assembly from early (0.05, 0.1), through middle (0.25, 0.5), to late (0.75, 1).

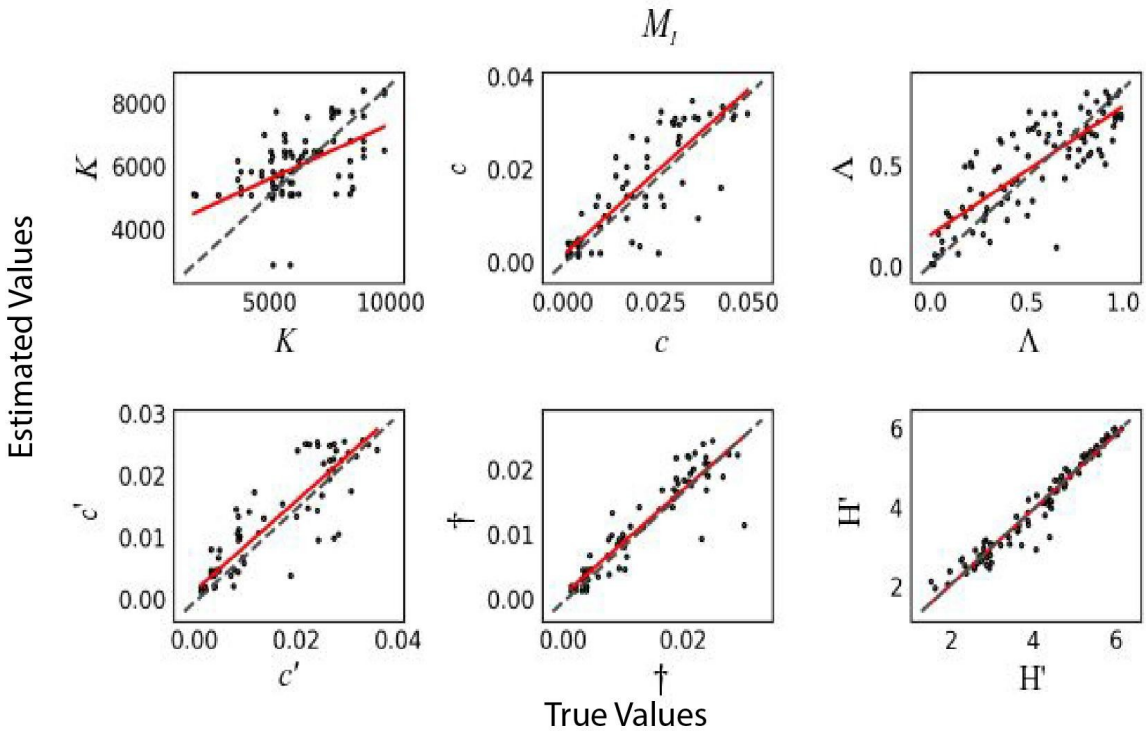

Figure 1.3 ABC cross-validation for model parameters
100 ABC cross-validation replicates for comparison of true vs estimated model parameters using only the 1D-SGD as data (M_1). The red line shows the linear least-squares regression between true and estimated values. Results are shown for estimating carrying capacity (K), colonization rate (c), fraction of equilibrium (Λ), effective colonization rate (c'), extinction rate (†), and Shannon's diversity index (H').

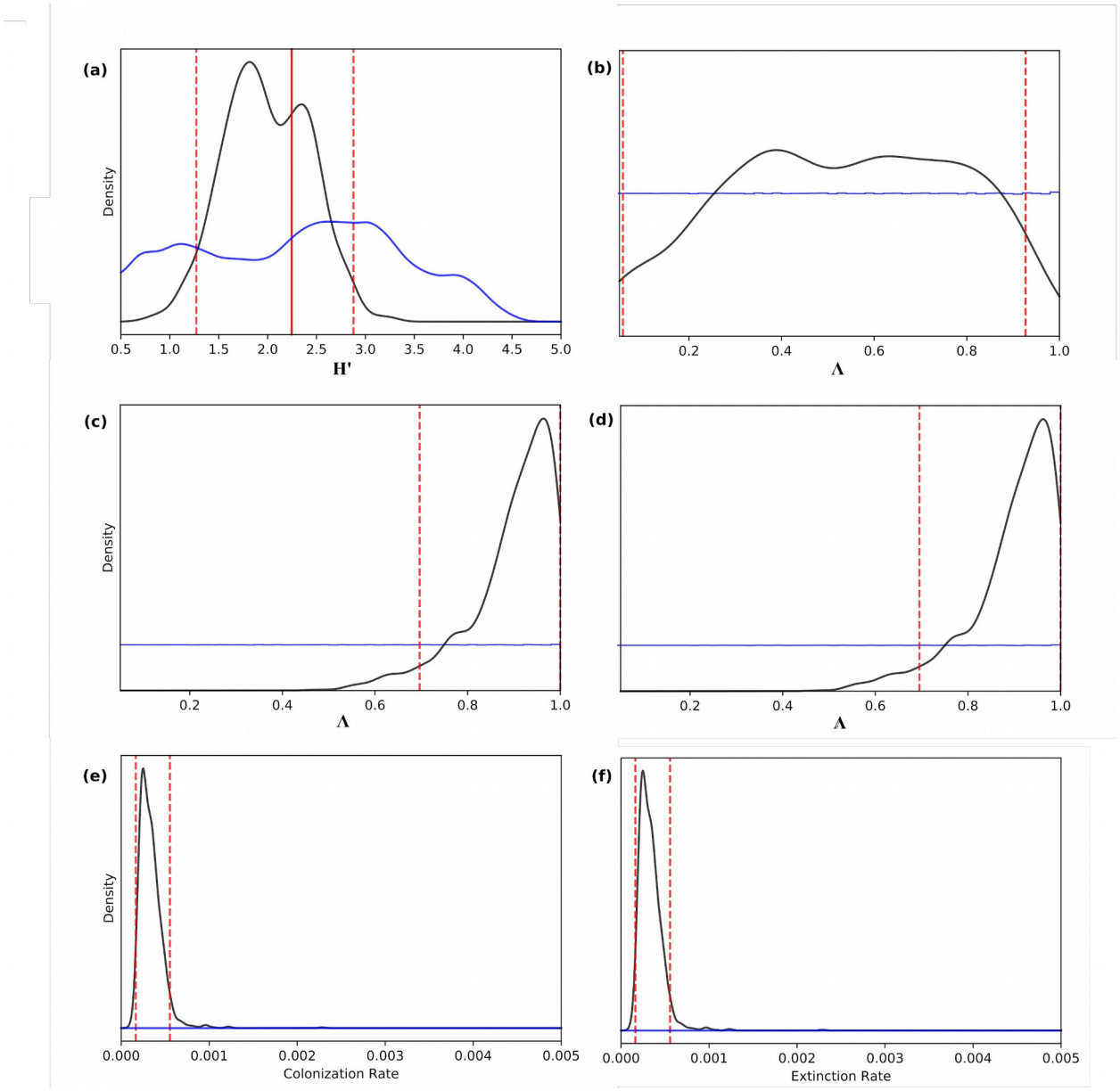

Figure 1.4 ABC posterior estimates of colonization/extinction rates, H' and Λ
ABC posterior estimates of colonization and extinction rates, and H' and Λ for the spider community from the island of Réunion. **(a)** Using the 1D-SGD as the summary statistic vector (M_I), the mode estimate of H' was 1.816 (95% HPD: 1.171-2.822; red dashed lines). The true value of H' from the observed abundance data was 2.246 (red solid line). Posterior estimates of Λ using three different model configurations: **(b)** only H' as data (M_A); **(c)** only the 1D-SGD as data (M_I); and **(d)** both H' and the 1D-SGD as data (M_{AI}). Posterior estimates of colonization rate and extinction rate using model M_{AI} are depicted in panels **(e)** and **(f)**, respectively. In all panels the red dashed lines indicate the 95% HPD, and the blue line illustrates the prior distribution.

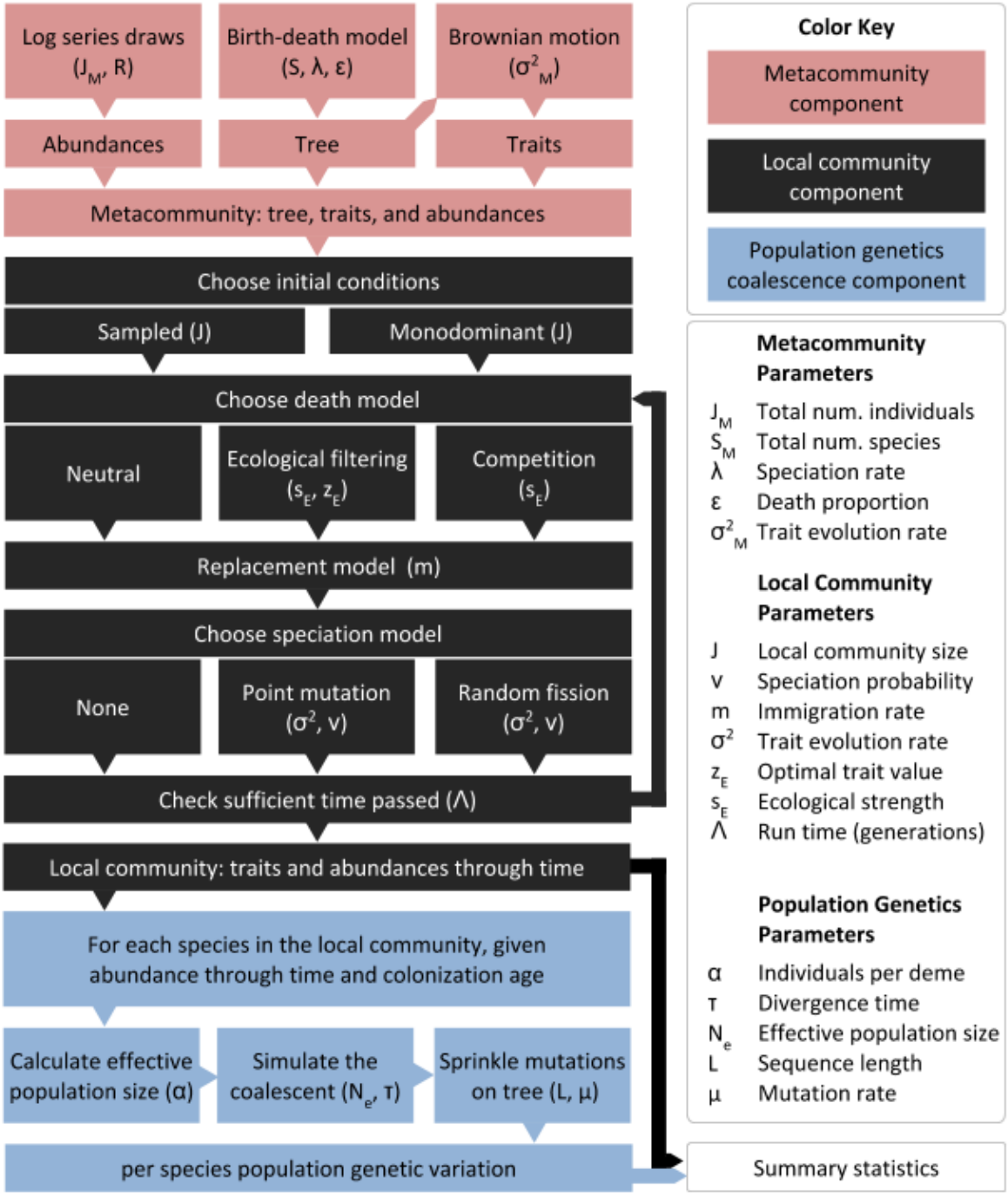

Figure 2.1 Conceptual diagram illustrating the three primary components of MESS
The metacommunity component (red) encompasses of a global phylogeny relating all species, along with species abundances and trait values evolved along the phylogeny. The local community component (black) involves a forward-time process during which a local community

assembles by birth, death, immigration, and local speciation. The population genetic component (blue) generates backward-time coalescent simulations per species which are parameterized contingent on the abundance history and colonization time generated by the forward-time component to approximate the accumulation of genetic diversity. Each box illustrates a sub-component of the model, and indicates the parameter(s) which determine the behavior of each sub-component. Arrows between sub-components indicate information flow through the process.

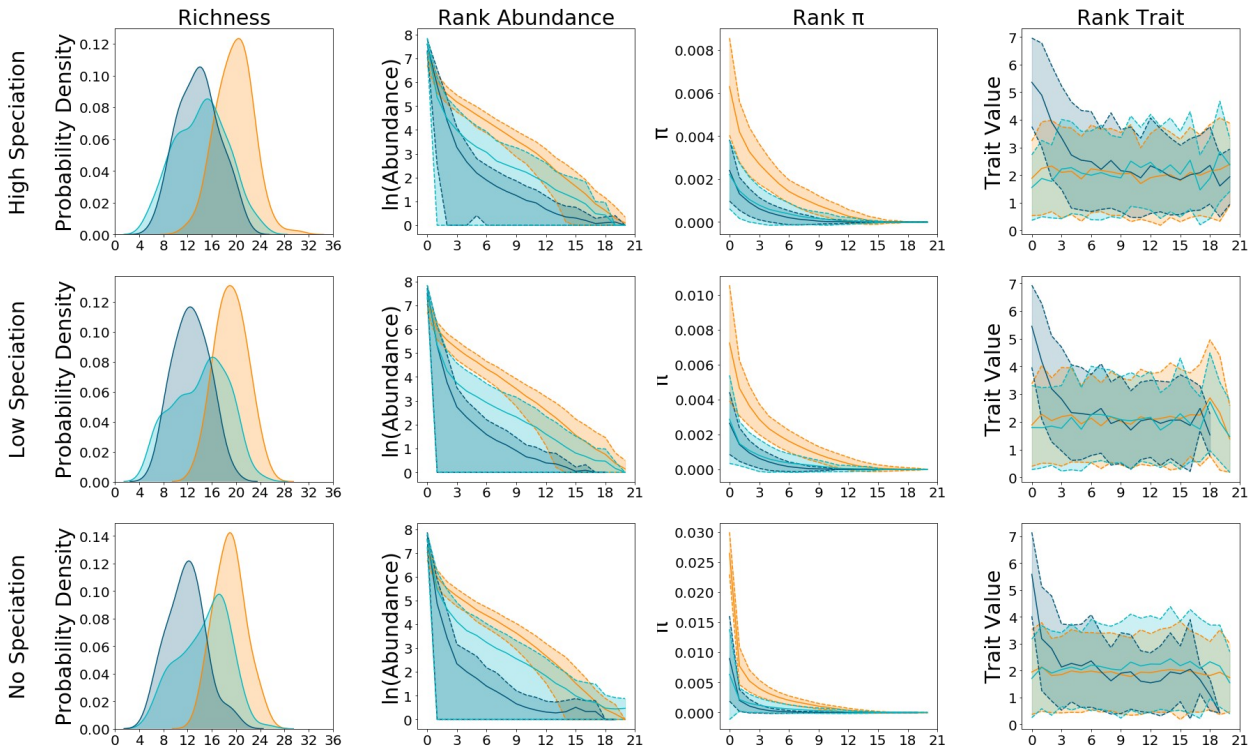

Figure 2.2 Effect of varying speciation rate and community assembly model on summary statistics

Species richness, rank abundance, rank genetic diversity, and rank trait values for 1000 simulations generated under neutral (orange), competition (dark blue) and filtering (aqua) scenarios with time fixed at 500 generations. From top to bottom, rows of panels correspond to simulations with high (ν = 0.0001), low (ν = 0.00005) and no (ν = 0) speciation. In the left column of panels, kernel density plots indicate the distribution of richness across simulations. In the rank plots (right panels), thick lines indicate average rank values and shaded areas show plus and minus one standard deviation.

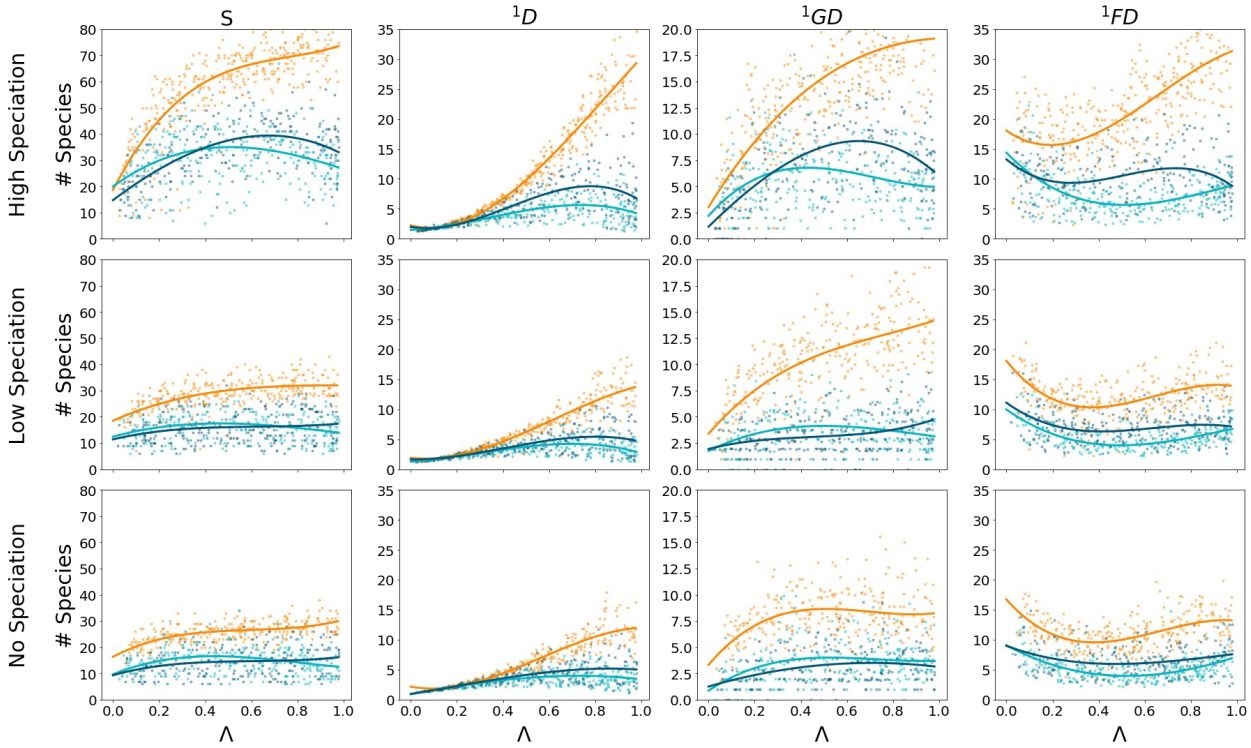

Figure 2.3 Community summary statistics through time for neutral and non-neutral models
This plot depicts the temporal change in select summary statistics for the three focal community assembly models at three different speciation rates: No, Low, and High corresponding to ν = 0, 0.0005, 0.005, respectively. Community assembly models depicted are neutral (orange), filtering (aqua), and competition (dark blue). Each subpanel shows the resultant summary statistic for 1000 simulations equally spaced through time for each model class. Simulated values are depicted as points, and a least squares polynomial is fit to better illustrate the trajectory. The far left column of panels illustrate species richness on the y-axes (S). The y-axes of the remaining columns illustrate the Hill number of order 1 for abundance, genetic diversity, and trait values, respectively.

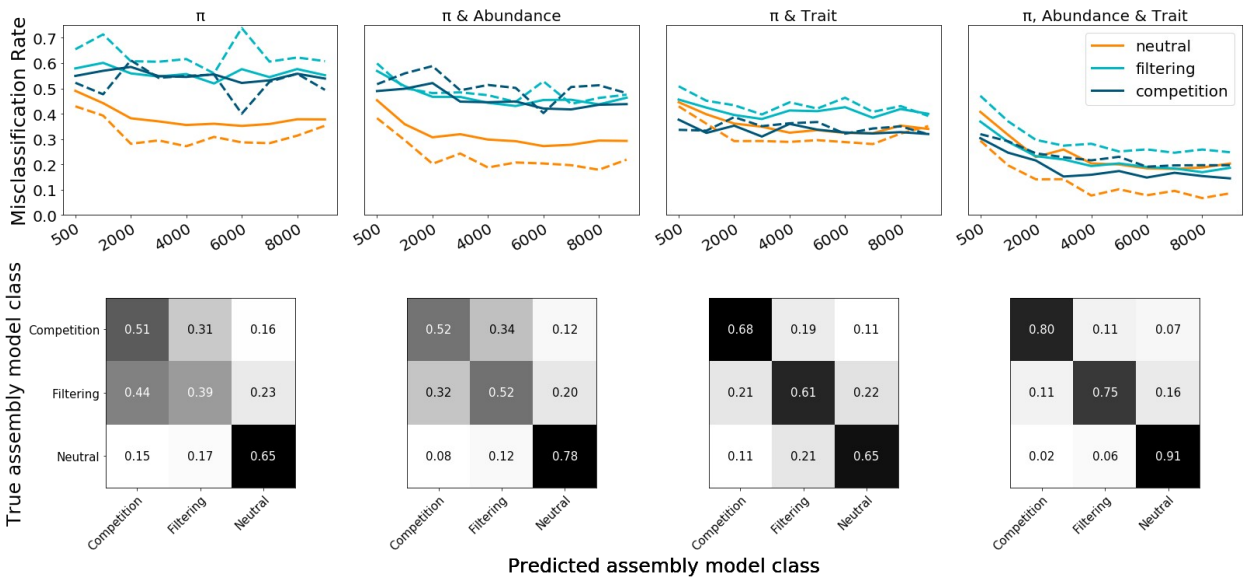

Figure 2.4 Machine learning classification error rates and confusion matrices
Top row) Random-forest misclassification error rates given different combinations of available data axes for varying sizes of local communities (J). Data axes used for each suite of simulations are indicated along the top of the figure. The x-axis indicates increasing sizes of J, from 500-10,000 in regular intervals. The y-axis indicates probability of assembly model misclassification, averaged over 1000 simulations per model class for each J (i.e. lower values indicate more accurate classification). In the figure, orange shows neutral simulations, aqua shows filtering, and dark blue shows competition. Solid lines indicate precision and dashed lines indicate recall. Bottom row) Confusion matrices depicting detailed model misclassification rates for data availability scenarios given J values between 9000 and 10,000. In these figures, values on the diagonals indicate the proportion of accurately classified simulations for each model class. Off-diagonal values indicate misclassified simulations.

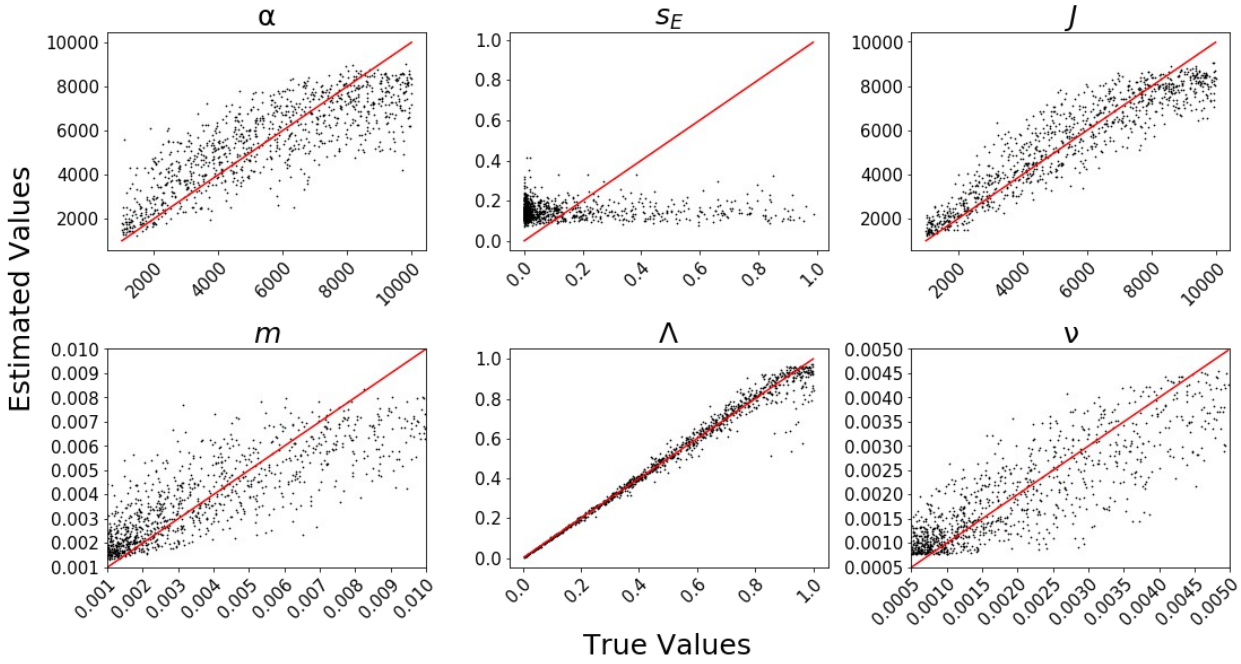

Figure 2.5 Machine learning cross-validation parameter estimation
1000 parameter estimation cross-validation (CV) replicates using neutral community assembly model simulations and summary statistics from all data axes. True parameter values are on the x-axes and the corresponding point estimates are on the y-axes. A parameter that is well estimated will have CV results that fall on or around the identity line (depicted in red). Note that ecological strength has no impact on neutral simulations, which produces the poor CV performance in estimating this parameter.

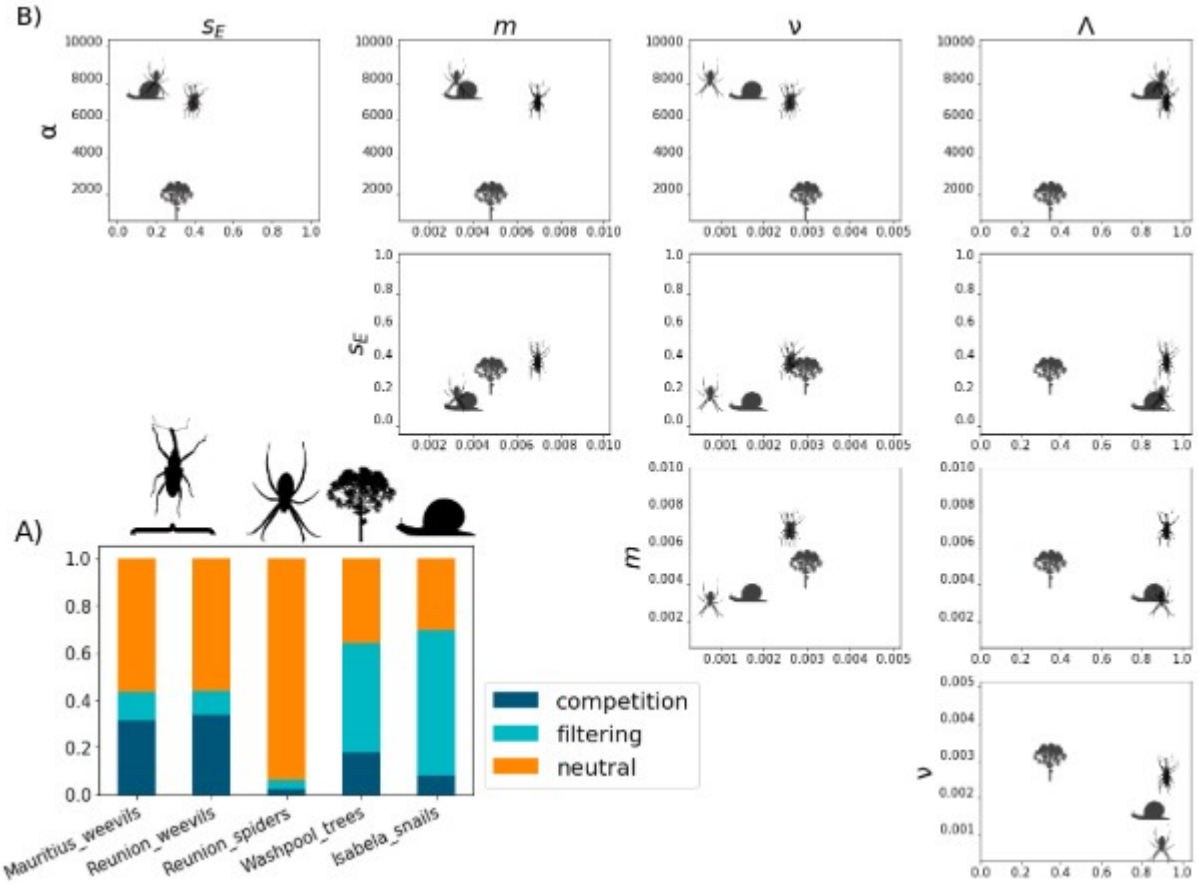

Figure 2.6 MESS empirical analysis

Empirical classification and parameter estimation of five local communities including snails, tropical trees, and island arthropods. Panel A) depicts machine learning classification probabilities for each empirical community for three focal community assembly models. The proportion of color within each bar represents the proportional predicted model class for neutrality (orange), environmental filtering (aqua), and competition (dark blue). Panel B) depicts pairwise estimates of five different model parameters under the best classified model for each local community dataset. The value along each parameter axis is indicated by the position of the representative icon.

Figure 3.1 Coral Triangle decapod community sampling localities

A digital elevation map of the Coral Triangle indicating the approximate location of the 10 sampling sites. A gross indication of one of the primary hydrological regimes of the region (the Indonesian Throughflow) is highlighted with the yellow arrow.

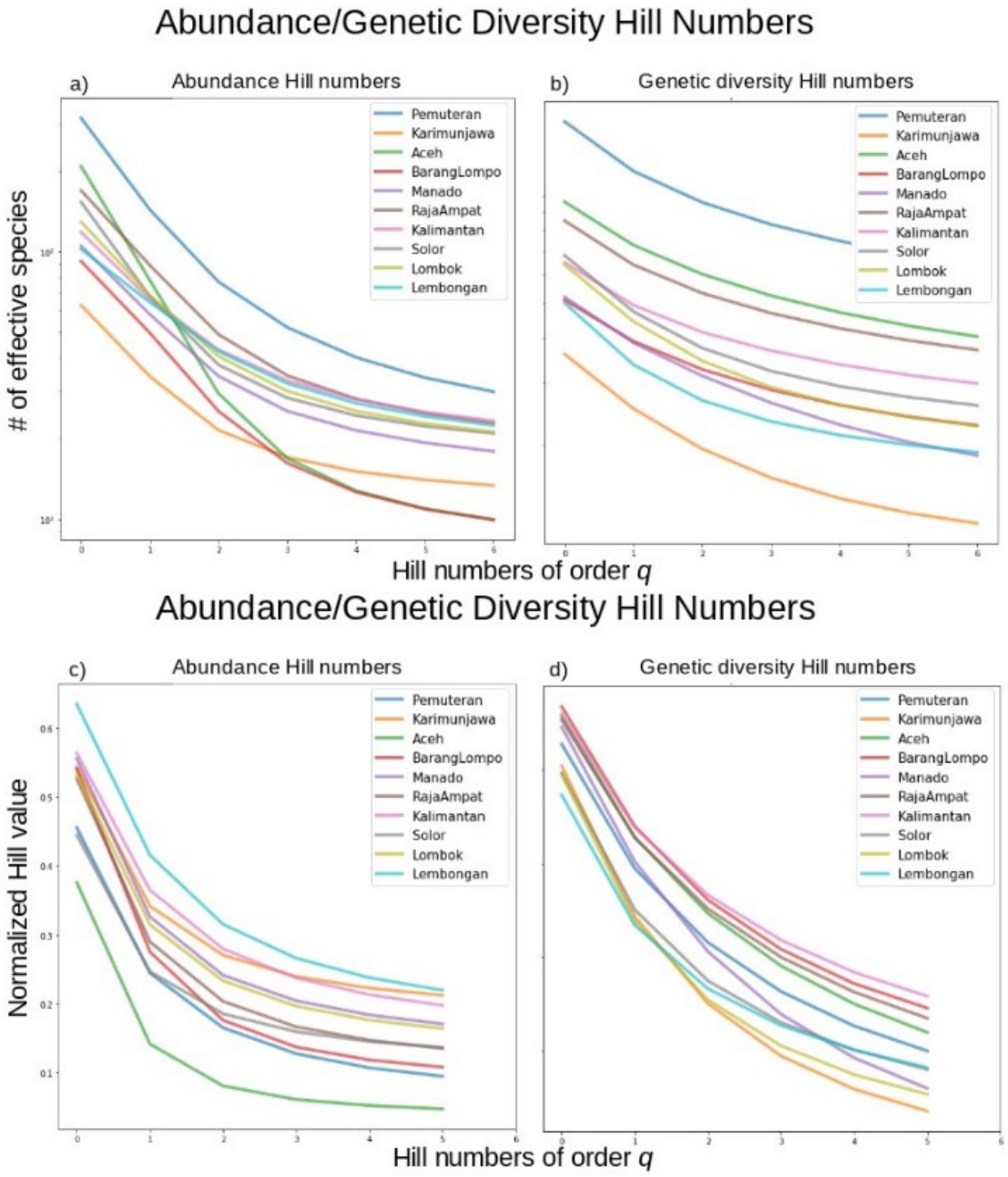

Figure 3.2 Hill numbers for abundance and genetic diversity distributions
The first 6 positive Hill numbers as well as ${}^{0}D$ (equal to species richness) for combined decapod crustacean community distributions of abundance (panel a) and genetic diversity (panel b) for each of the 10 Coral Triangle sampling sites. Hill numbers are expressed in terms of numbers of "effective species", which provide an indication of how evenly abundance and genetic diversity are distributed within communities at each sampling site. Panels c) and d) show the same data, but normalize the Hill numbers by the value of ${}^{0}D$ to allow better comparison across sites.

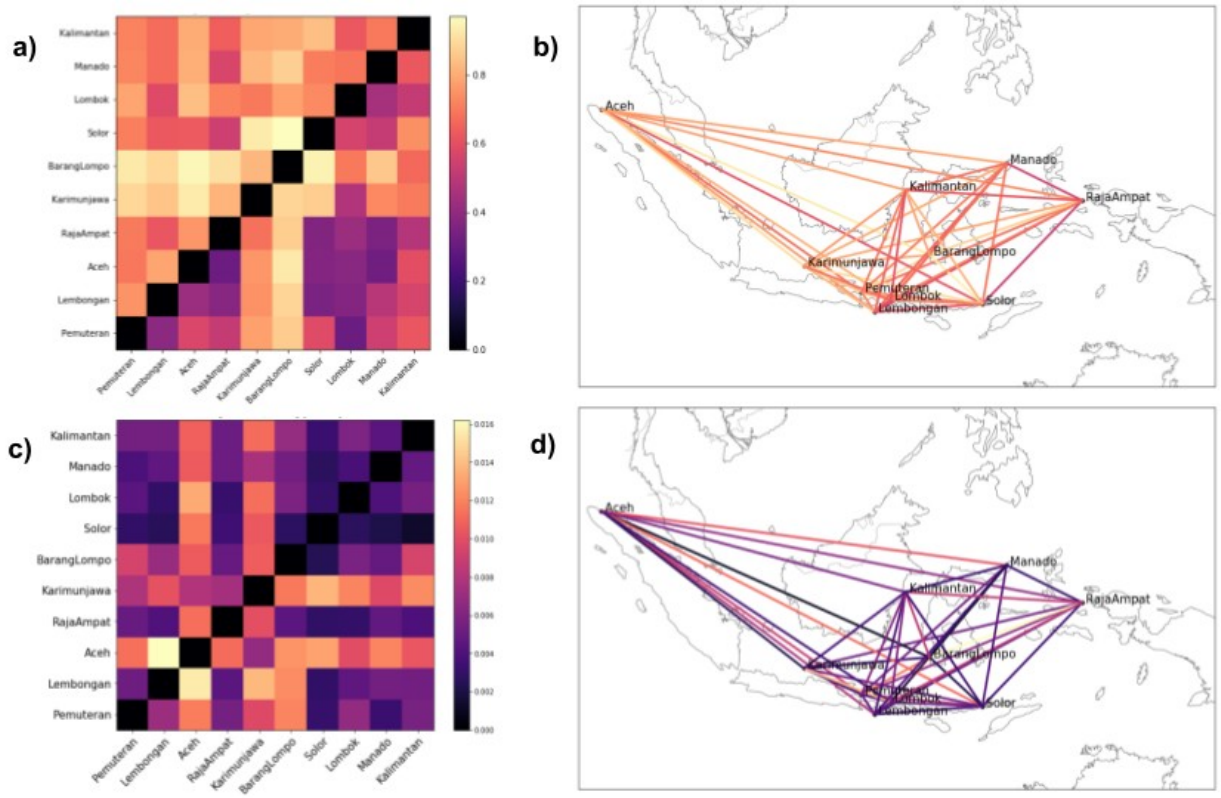

Figure 3.3 Pairwise abundance and genetic diversity turnover among sites
This figure depicts beta diversity as measured by Bray-Curtis dissimilarity (above the diagonal) and cosine distance (below the diagonal) for abundance distributions (panel a), and mean (above) and standard deviation (below) of absolute genetic divergence (D_{xy}; panel c). Pairwise Bray-Curtis dissimilarities (identical to values from panels a & c) are plotted on the map of the region for abundance (panel b) and absolute genetic divergence (panel d). In all panels darker colors indicate higher compositional similarity and lighter colors indicate lower similarity. Note that dissimilarity scales for abundance (panels a & b) and absolute genetic divergence (panels c & d) are on different scales and therefore are not directly comparable.

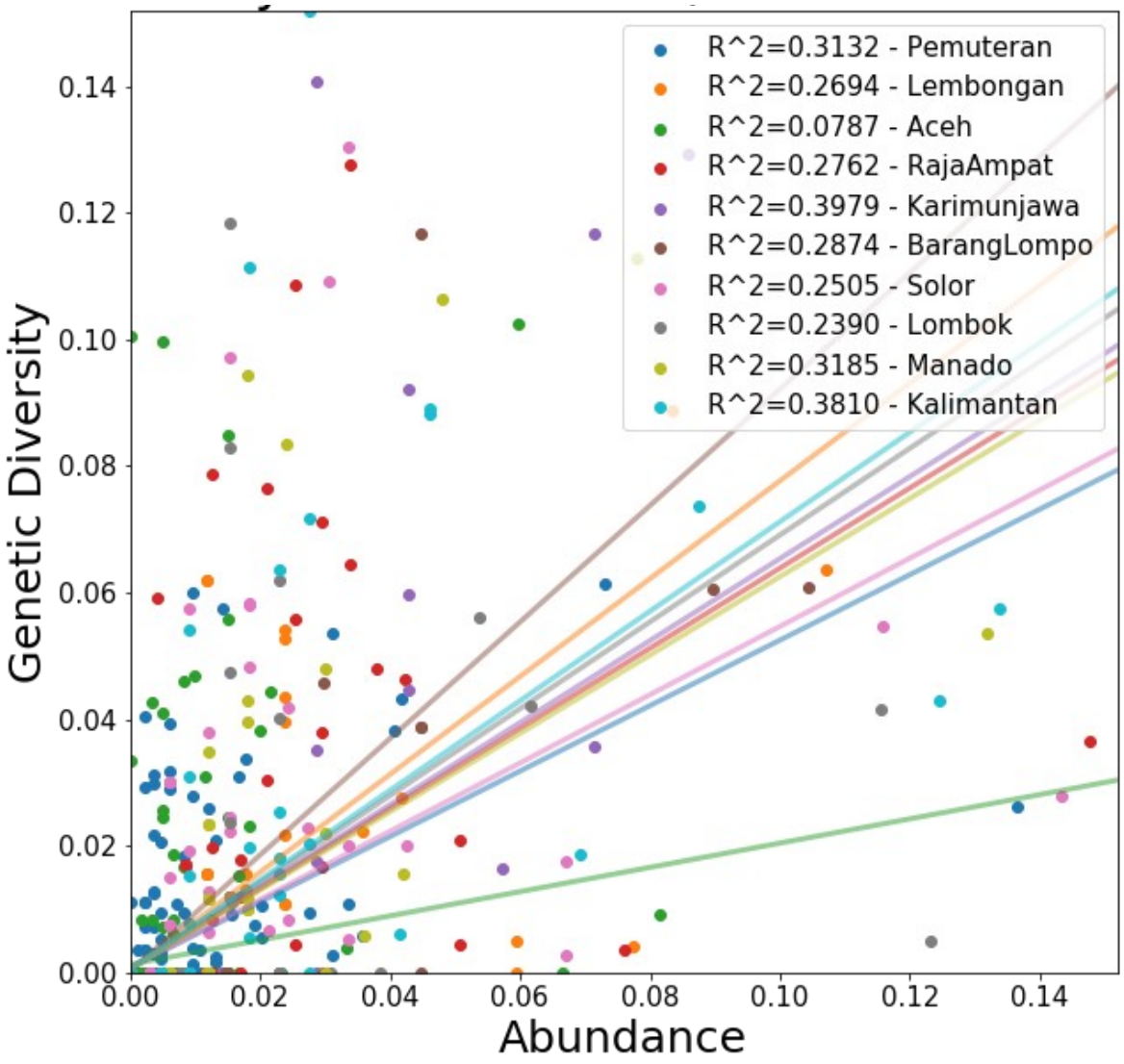

Figure 3.4 Abundance/genetic diversity correlations within sampling sites
This figure depicts the per species correlation between abundance and genetic diversity. Each point in the figure corresponds to one species, with points colored by sampling location. The values in the figure for both axes are scaled to proportional abundance and proportional genetic diversity. Least-squares regression lines are plotted for each site, and R^2 values indicate the strength of the correlations.

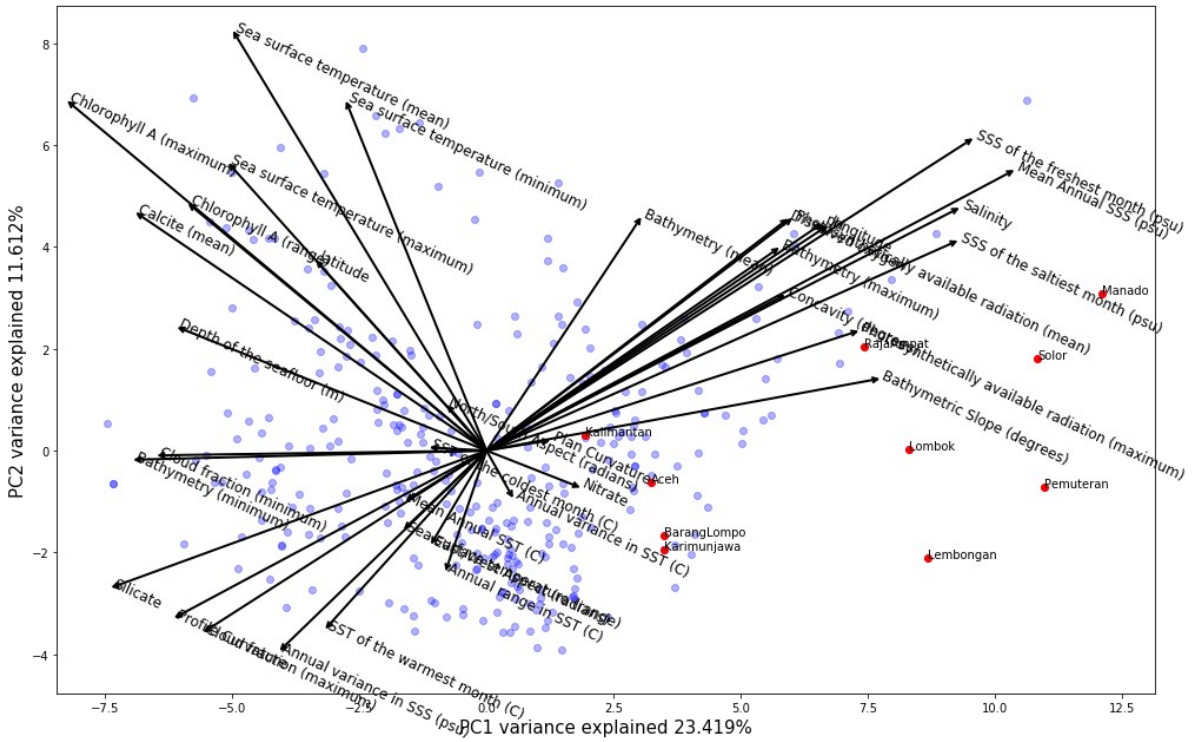

Figure 3.5 Principal component analysis of environmental space
MARSPEC and Bio-ORACLE variables projected into principal component (PC) space for the 10 sampling sites (red points) and 500 random sites (blue points) sampled from within a bounding box described by the sampling locations. The first two PCs are depicted, with PC1 (explained variance 23.42%) on the x-axis and PC2 (explained variance 11.61%) on the y-axis. Arrows indicate loadings on each MARSPEC variable, with the length of the arrow corresponding to increasing proportions of loading weight.

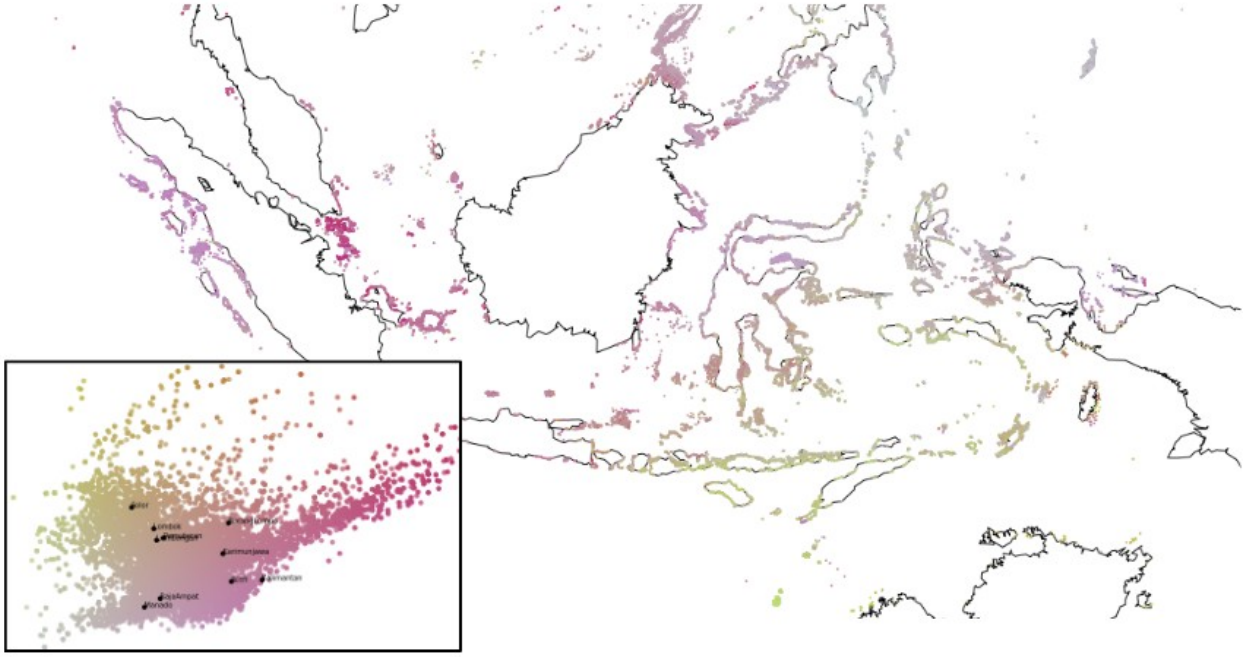

Figure 3.6 Principal component analysis of environmental data projected across the region
All MARSPEC and Bio-ORACLE variables are projected into environmental principal component (ePC) space, and then plotted across the landscape, which is masked to the known distribution of corals in the region. The color of each point on the map corresponds to the location of the environment at that point within ePC space. Inset shows ~80,000 points sampled from the background of known coral occurrences, and colored according to their location in ePC space. Loadings are as in figure 5, but are removed here for clarity. Locations of each of the 10 empirical sampling sites are also plotted in ePC space, indicated by the labeled black points.

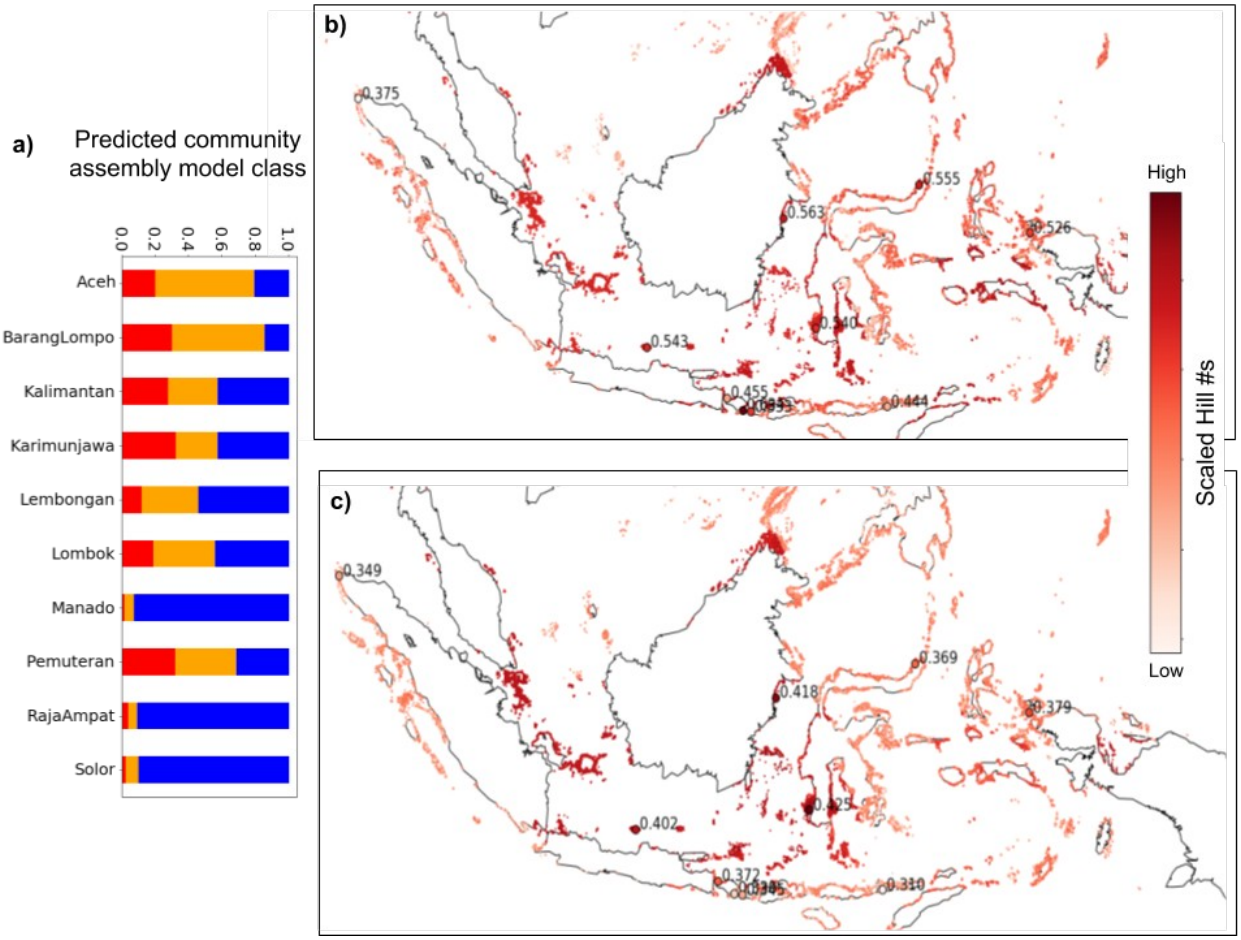

Figure 3.7 Predicted community assembly model class and predicted abundance and genetic diversity distributions projected across the landscape

The machine learning assembly model class prediction probabilities for each sampling site (panel a) with support for neutral (blue), filtering (orange), and competition (red) models. The fraction of the bar indicates the proportion of support for each model class. Panels b) and c) show the projected abundance (^1D) and genetic diversity (^1GD) structure summarized as the first Hill number normalized by species richness. A machine learning algorithm was trained on environmental correlates with ^1D and ^1GD as the target variables. Each point in the figures show the predicted ^1D and ^1GD at 80,000 points sampled across the landscape and masked to the known distribution of corals in the region, using the machine learning algorithm trained from the 10 sampled localities. Darker colors indicate higher values, and lighter colors indicate lower. The location and known ^1D and ^1GD of the 10 sampling locations are also indicated on the figure.

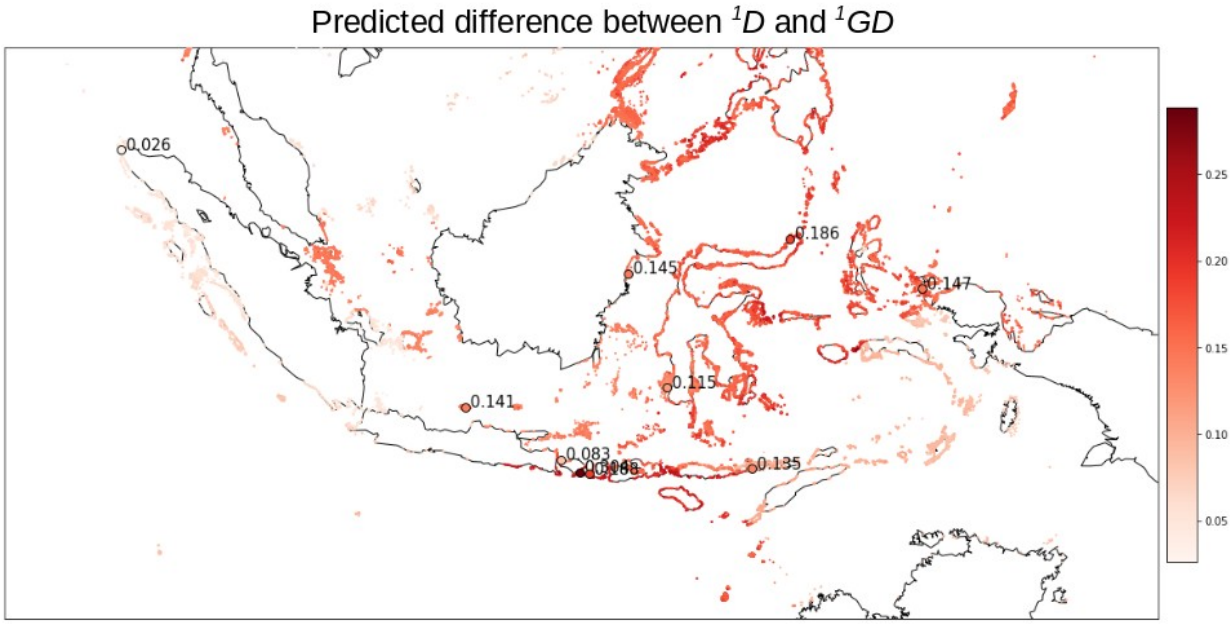

Figure 3.8 Difference between predicted ^1D and ^1GD projected across the landscape
This figure illustrates the difference between predicted ^1D and ^1GD values for all reef sites throughout the Coral Triangle. The difference between ^1D and ^1GD for the sampled communities are also indicated. Here darker values indicate greater difference between ^1D and ^1GD, independent of magnitude, and lighter colors indicate more similarity. Because some species may be present in a community and yet have no genetic diversity, the scaled values of ^1GD will always be smaller than those of ^1D.

Literature Cited

Adams, C. I. M., Knapp, M., Gemmell, N. J., Jeunen, G.-J., Bunce, M., Lamare, M. D., & Taylor, H. R. (2019). Beyond Biodiversity: Can Environmental DNA (eDNA) Cut It as a Population Genetics Tool? *Genes*, *10*(3).

Al Malik, M. D., Kholilah, N., Kurniasih, E. M., Sembiring, A., Pertiwi, N. P. D., Ambariyanto, A., … Meyer, C. (2018). Biodiversity of Cryptofauna (Decapods) and Their Correlation with Dead Coral Pocillopora sp. Volume at Bunaken Island, North Sulawesi. *IOP Conference Series: Earth and Environmental Science*, *116*(1), 012053.

Anderson, M. J., & Santana-Garcon, J. (2015). Measures of precision for dissimilarity-based multivariate analysis of ecological communities. *Ecology Letters*, *18*(1), 66–73.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews. Genetics*, *17*(2), 81–92.

Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P., & Emerson, B. C. (2018). Why the COI barcode should be the community DNA metabarcode for the metazoa. *Molecular Ecology*, *27*(20), 3968–3975.

Assis, J., Tyberghein, L., & Bosch, S. (2018). Bio ORACLE v2. 0: Extending marine data layers for bioclimatic modelling. *Global Ecology and Biogeography*

Avise, J. C., Bowen, B. W., & Ayala, F. J. (2016). In the light of evolution X: Comparative phylogeography. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(29), 7957–7961.

Azaele, S., Maritan, A., Cornell, S. J., Suweis, S., Banavar, J. R., Gabriel, D., & Kunin, W. E. (2015). Towards a unified descriptive theory for spatial ecology: predicting biodiversity patterns across spatial scales. *Methods in Ecology and Evolution*, *6*(3), 324–332.

Azaele, S., Pigolotti, S., Banavar, J. R., & Maritan, A. (2006). Dynamical evolution of ecosystems. *Nature*, *444*(7121), 926–928.

Bálint, M., Pfenninger, M., Grossart, H.-P., Taberlet, P., Vellend, M., Leibold, M. A., … Bowler, D. (2018). Environmental DNA Time Series in Ecology. *Trends in Ecology & Evolution*, *33*(12), 945–957.

Barabás, G., D'Andrea, R., Rael, R., Meszéna, G., & Ostling, A. (2013). Emergent neutrality or hidden niches? *Oikos* , *122*(11), 1565–1572.

Baselga, A., Fujisawa, T., Crampton-Platt, A., Bergsten, J., Foster, P. G., Monaghan, M. T., & Vogler, A. P. (2013). Whole-community DNA barcoding reveals a spatio-temporal continuum of biodiversity at species and genetic levels. *Nature Communications*, *4*, 1892.

Baselga, A., Gómez-Rodríguez, C., & Vogler, A. P. (2015). Multi-hierarchical macroecology at species and genetic levels to discern neutral and non-neutral processes. *Global Ecology and Biogeography*, *24*(8), 873–882.

Baselga, A., & Orme, C. D. L. (2012). betapart: an R package for the study of beta diversity. *Methods in Ecology and Evolution / British Ecological Society*, *3*(5), 808–812.

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, *162*(4), 2025.

Bellwood, D. R., Renema, W., & Rosen, B. R. (2012). Biodiversity hotspots, evolution and coral reef biogeography. *Biotic Evolution and Environmental Change in Southeast Asia*, *216*.

Bohan, D. A., Vacher, C., Tamaddoni-Nezhad, A., Raybould, A., Dumbrell, A. J., & Woodward, G. (2017). Next-Generation Global Biomonitoring: Large-scale, Automated Reconstruction

of Ecological Networks. *Trends in Ecology & Evolution*, *32*(7), 477–487.
Boltzmann, L. (1872). *Sitzungsberichte Akad. Wiss., Vienna, II, 66: 275-370; English transl.: Brush, SG (1966) Kinetic Theory: Vol. 2 Irreversible Processes*. Permagon Press, Oxford.
Bosch, S., Tyberghein, L., & De Clerck, O. (2017). sdmpredictors: Species distribution modelling predictor datasets. *R Package Version 0. 2*, *6*.
Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, *27*(4), 325–349.
Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.
Brower, A. V. (1994). Rapid morphological radiation and convergence among races of the butterfly Heliconius erato inferred from patterns of mitochondrial DNA evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(14), 6491–6495.
Bucklin, A., Steinke, D., & Blanco-Bercial, L. (2011). DNA barcoding of marine metazoa. *Annual Review of Marine Science*, *3*, 471–508.
Bunnefeld, L., Hearn, J., Stone, G. N., & Lohse, K. (2018). Whole-genome data reveal the complex history of a diverse ecological community. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(28), E6507–E6515.
Burbrink, F. T., Chan, Y. L., Myers, E. A., Ruane, S., Smith, B. T., & Hickerson, M. J. (2016). Asynchronous demographic responses to Pleistocene climate change in Eastern Nearctic vertebrates. *Ecology Letters*, *19*(12), 1457–1467.
Burbrink, F. T., McKelvy, A. D., Pyron, A. R., & Myers, E. A. (2015). Predicting community structure in snakes on Eastern Nearctic islands using ecological neutral theory and phylogenetic methods. *Proceedings of the Royal Society of London B*.
Butler, M. A., & King, A. A. (2004). Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *The American Naturalist*, *164*(6), 683–695.
Bzdok, D., Krzywinski, M., & Altman, N. (2017). Points of Significance: Machine learning: a primer. *Nature Methods*, *14*(12), 1119–1120.
Cabral, J. S., Wiegand, K., & Kreft, H. (2019). Interactions between ecological, evolutionary and environmental processes unveil complex dynamics of insular plant diversity. *Journal of Biogeography*, *103*, 9130.
Capo, E., Debroas, D., Arnaud, F., & Guillemot, T. (2016). Long term dynamics in microbial eukaryotes communities: a palaeolimnological view based on sedimentary DNA. *Molecular Ecology*, *25*, 5925–5943.
Carstens, B. C., Gruenstaeudl, M., & Reid, N. M. (2016). Community trees: Identifying codiversification in the Páramo dipteran community. *Evolution*, *70*(5), 1080–1093.
Cavender-Bares, J., Gamon, J. A., Hobbie, S. E., Madritch, M. D., Meireles, J. E., Schweiger, A. K., & Townsend, P. A. (2017). Harnessing plant spectra to integrate the biodiversity sciences across biological and spatial scales. *American Journal of Botany*, *104*(7), 966–969.
Cavender-Bares, J., Kozak, K. H., Fine, P. V. A., & Kembel, S. W. (2009). The merging of community ecology and phylogenetic biology. *Ecology Letters*, *12*(7), 693–715.
Chao, A., Chiu, C.-H., & Jost, L. (2014). Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution, and Systematics*. https://doi.org/10.1146/annurev-ecolsys-120213-091540
Chase, J. M. (2010). Stochastic community assembly causes higher biodiversity in more productive environments. *Science*, *328*(5984), 1388–1391.
Chave, J., & Leigh, E. G., Jr. (2002). A spatially explicit neutral model of beta-diversity in

tropical forests. *Theoretical Population Biology*, *62*(2), 153–168.
Chave, J., Muller-Landau, H. C., & Levin, S. A. (2002). Comparing classical community models: theoretical consequences for patterns of diversity. *The American Naturalist*, *159*(1), 1–23.
Chesson, P. (2000). Mechanisms of maintenance of species diversity. *Annual Review of Ecology and Systematics*, *31*(1), 343–366.
Chisholm, R. A., & O'Dwyer, J. P. (2014). Species ages in neutral biodiversity models. *Theoretical Population Biology*, *93*, 85–94.
Chust, G., Irigoien, X., Chave, J., & Harris, R. P. (2013). Latitudinal phytoplankton distribution and the neutral theory of biodiversity. *Global Ecology and Biogeography*, *22*(5), 531–543.
Coissac, E., Hollingsworth, P. M., Lavergne, S., & Taberlet, P. (2016). From barcodes to genomes: extending the concept of DNA barcoding. *Molecular Ecology*, *25*(7), 1423–1428.
Condit, R., Pitman, N., Leigh, E. G., Jr, Chave, J., Terborgh, J., Foster, R. B., … Hubbell, S. P. (2002). Beta-diversity in tropical forest trees. *Science*, *295*(5555), 666–669.
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.
Craven, D., Knight, T. M., Barton, K. E., Bialic-Murphy, L., & Chase, J. M. (2019). Dissecting macroecological and macroevolutionary patterns of forest biodiversity across the Hawaiian archipelago. *Proceedings of the National Academy of Sciences of the United States of America*.
Csilléry, K., François, O., & Blum, M. G. B. (2012). abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, *3*(3), 475–479.
D'Amen, M., Dubuis, A., Fernandes, R. F., Pottier, J., Pellissier, L., & Guisan, A. (2015). Using species richness and functional traits predictions to constrain assemblage predictions from stacked species distribution models. *Journal of Biogeography*, *42*(7), 1255–1266.
Davies, N., Field, D., Amaral-Zettler, L., Clark, M. S., Deck, J., Drummond, A., … GOs-COS. (2014). The founding charter of the Genomic Observatories Network. *GigaScience*, *3*(1), 2.
Davies, T. J., Allen, A. P., Borda-de-Água, L., Regetz, J., & Melián, C. J. (2011). Neutral biodiversity theory can explain the imbalance of phylogenetic trees but not the tempo of their diversification. *Evolution*, *65*(7), 1841–1850.
Deck, J., Gaither, M. R., Ewing, R., Bird, C. E., Davies, N., Meyer, C., … Crandall, E. D. (2017). The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples. *PLoS Biology*, *15*(8), e2002925.
Degenhardt, F., Seifert, S., & Szymczak, S. (2019). Evaluation of variable selection methods for random forests and omics data sets. *Briefings in Bioinformatics*, *20*(2), 492–503.
Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., … Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895.
DePatta Pillar, V. (1998). Sampling sufficiency in ecological surveys. *Abstracta Botanica*, 37–48.
Derocles, S. A. P., Bohan, D. A., Dumbrell, A. J., Kitson, J. J. N., Massol, F., Pauvert, C., … Evans, D. M. (2018). Chapter One - Biomonitoring for the 21st Century: Integrating Next-Generation Sequencing Into Ecological Network Analysis. In D. A. Bohan, A. J. Dumbrell, G. Woodward, & M. Jackson (Eds.), *Advances in Ecological Research* (Vol. 58, pp. 1–62). Academic Press.
Distler, T., Schuetz, J. G., Velásquez-Tibatá, J., & Langham, G. M. (2015). Stacked species distribution models and macroecological models provide congruent projections of avian species richness under climate change. *Journal of Biogeography*, *42*(5), 976–988.

Dopheide, A., Tooman, L. K., Grosser, S., Agabiti, B., Rhode, B., Xie, D., … Newcomb, R. D. (2019). Estimating the biodiversity of terrestrial invertebrates on a forested island using DNA barcodes and metabarcoding data. *Ecological Applications*, *29*(4), e01877.
Ellis, N., Smith, S. J., & Pitcher, C. R. (2012). Gradient forests: calculating importance gradients on physical predictors. *Ecology*, *93*(1), 156–168.
Emerson, B. C., Casquet, J., López, H., Cardoso, P., Borges, P. A. V., Mollaret, N., … Thébaud, C. (2017). A combined field survey and molecular identification protocol for comparing forest arthropod biodiversity across spatial scales. *Molecular Ecology Resources*, *17*(4), 694–707.
Engen, S., Solbu, E. B., & Sæther, B. E. (2017). Neutral or non neutral communities: temporal dynamics provide the answer. *Oikos*.
Epp, L. S. (2019). A global perspective for biodiversity history with ancient environmental DNA. *Molecular Ecology*, *28*(10), 2456–2458.
Etienne, R. S., & Haegeman, B. (2011). The neutral theory of biodiversity with random fission speciation. *Theoretical Ecology*, *4*(1), 87–109.
Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, *61*(5), 717–726.
Fitzpatrick, M. C., & Keller, S. R. (2015). Ecological genomics meets community-level modelling of biodiversity: Mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters*, *18*(1), 1–16.
Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.
Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, *21*(2), 137–146.
Gaggiotti, O. E., Chao, A., Peres-Neto, P., Chiu, C.-H., Edwards, C., Fortin, M.-J., … Selkoe, K. A. (2018). Diversity from genes to ecosystems: A unifying framework to study variation across biological metrics and scales. *Evolutionary Applications*, *11*(7), 1176–1193.
Garrick, R. C., Bonatelli, I. A. S., Hyseni, C., Morales, A., Pelletier, T. A., Perez, M. F., … Carstens, B. C. (2015). The evolution of phylogeographic data sets. *Molecular Ecology*, *24*(6), 1164–1171.
Gascuel, F., Laroche, F., Bonnet-Lebrun, A.-S., & Rodrigues, A. S. L. (2016). The effects of archipelago spatial structure on island diversity and endemism: predictions from a spatially-structured neutral model. *Evolution*, *70*(11), 2657–2666.
Gavrilets, S., & Vose, A. (2005). Dynamic patterns of adaptive radiation. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(50), 18040–18045.
Gelman, A. (2003). A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-fit Testing. *International Statistical Review = Revue Internationale de Statistique*, *71*(2), 369–382.
Gorelick, R. (2006). Combining richness and abundance into a single diversity index using matrix analogues of Shannon's and Simpson's indices. *Ecography*, *29*(4), 525–530.
Gotelli, N. J., & Abele, L. G. (1983). Community patterns of coral-associated decapods. *Marine Ecology Progress Series. Oldendorf*, *13*(2), 131–139.
Graham, C. H., & Fine, P. V. A. (2008). Phylogenetic beta diversity: linking ecological and evolutionary processes across space in time. *Ecology Letters*, *11*(12), 1265–1277.
Gratton, P., Marta, S., Bocksberger, G., Winter, M., Trucchi, E., & Kühl, H. (2017). A world of sequences: can we use georeferenced nucleotide databases for a robust automated

phylogeography? *Journal of Biogeography*, *44*(2), 475–486.
Grummer, J. A., Beheregaray, L. B., Bernatchez, L., Hand, B. K., Luikart, G., Narum, S. R., & Taylor, E. B. (2019). Aquatic Landscape Genomics and Environmental Effects on Genetic Variation. *Trends in Ecology & Evolution*, *34*(7), 641–654.
Grundler, M. R., Singhal, S., Cowan, M. A., & Rabosky, D. L. (2019). Is genomic diversity a useful proxy for census population size? Evidence from a species-rich community of desert lizards. *Molecular Ecology*.
Harfoot, M. B. J., Newbold, T., Tittensor, D. P., Emmott, S., Hutton, J., Lyutsarev, V., … Purves, D. W. (2014). Emergent global patterns of ecosystem structure and function from a mechanistic general ecosystem model. *PLoS Biology*, *12*(4), e1001841.
Harmon, L. J., & Harrison, S. (2015). Species diversity is dynamic and unbounded at local and continental scales. *The American Naturalist*, *185*(5), 584–593.
Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E., & Challenger, W. (2008). GEIGER: investigating evolutionary radiations. *Bioinformatics* , *24*(1), 129–131.
Harris, D. J. (2015). Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*, *6*(4), 465–473.
Head, C. E. I., Bonsall, M. B., Jenkins, T. L., Koldewey, H., Pratchett, M. S., Taylor, M. L., & Rogers, A. D. (2018). Exceptional biodiversity of the cryptofaunal decapods in the Chagos Archipelago, central Indian Ocean. *Marine Pollution Bulletin*, *135*, 636–647.
Hersch-Green, E. I., Turley, N. E., & Johnson, M. T. J. (2011). Community genetics: what have we accomplished and where should we be going? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *366*(1569), 1453–1460.
Hickerson, M. J., Carstens, B. C., Cavender-Bares, J., Crandall, K. A., Graham, C. H., Johnson, J. B., … Yoder, A. D. (2010). Phylogeography's past, present, and future: 10 years after Avise, 2000. *Molecular Phylogenetics and Evolution*.
Hill, M. O. (1973). Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology*, *54*(2), 427–432.
Hoeksema, B. W. (2007). Delineation of the Indo-Malayan Centre of Maximum Marine Biodiversity: The Coral Triangle. In W. Renema (Ed.), *Biogeography, Time, and Place: Distributions, Barriers, and Islands* (pp. 117–178). Dordrecht: Springer Netherlands.
Huang, W., Takebayashi, N., Qi, Y., & Hickerson, M. J. (2011). MTML-msBayes: approximate Bayesian comparative phylogeographic inference from multiple taxa and multiple loci with rate heterogeneity. *BMC Bioinformatics*, *12*, 1.
Hubbell, S. P. (2001). *The unified neutral theory of biodiversity and biogeography* (Vol. 32). Princeton University Press.
Hutchinson, M. F. (1998). Interpolation of rainfall data with thin plate smoothing splines. Part I: Two dimensional smoothing of data with short range correlation. *Journal of Geographic Information and Decision Analysis*, *2*(2), 139–151.
Jabot, F., & Chave, J. (2009). Inferring the parameters of the neutral theory of biodiversity using phylogenetic information and implications for tropical forests. *Ecology Letters*, *12*(3), 239–248.
Jabot, F., Laroche, F., Massol, F., Arthaud, F., & Crabot, J. (2018). Assessing metacommunity processes through signatures in spatiotemporal turnover of community composition. *bioRxiv*.
Jasiewicz, J., & Stepinski, T. F. (2013). Geomorphons—a pattern recognition approach to classification and mapping of landforms. *Geomorphology* , *182*, 147–156.
Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M. J., … Turak, E. (2019). Essential biodiversity variables for mapping and monitoring species populations.

Nature Ecology & Evolution, *3*(4), 539–551.
Jetz, W., & Rahbek, C. (2002). Geographic range size and determinants of avian species richness. *Science*, *297*(5586), 1548–1551.
Jordan, S. M. R., Barraclough, T. G., & Rosindell, J. (2016). Quantifying the effects of the break up of Pangaea on global terrestrial diversification with neutral theory. *Philosophical Transactions of the Royal Society B*, *371*(1691).
Kalyuzhny, M., Kadmon, R., & Shnerb, N. M. (2015). A neutral theory with environmental stochasticity explains static and dynamic properties of ecological communities. *Ecology Letters*, *18*(6), 572–580.
Keil, P., & Chase, J. M. (2019). Global patterns and drivers of tree diversity integrated across a continuum of spatial grains. *Nature Ecology & Evolution*, *3*(3), 390–399.
Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Computational Biology*, *12*(5), e1004842.
Kerr, J. T., & Packer, L. (1997). Habitat heterogeneity as a determinant of mammal species richness in high-energy regions. *Nature*, *385*(6613), 252.
Kholilah, N., Al Malik, M. D., Kurniasih, E. M., Sembiring, A., Ambariyanto, A., & Mayer, C. (2018). Conditions of Decapods Infraorders in Dead Coral Pocillopora sp. at Pemuteran, Bali: Study Case 2011 and 2016. *IOP Conference Series: Earth and Environmental Science*, *116*(1), 012070.
Kirpich, A., Ainsworth, E. A., Wedow, J. M., Newman, J. R. B., Michailidis, G., & McIntyre, L. M. (2018). Variable selection in omics data: A practical evaluation of small sample sizes. *PloS One*, *13*(6), e0197910.
Kitson, J. J. N., Warren, B. H., Thébaud, C., Strasberg, D., & Emerson, B. C. (2018). Community assembly and diversification in a species-rich radiation of island weevils (Coleoptera: Cratopini). *Journal of Biogeography*, *45*(9), 2016–2026.
Knowlton, N., & Leray, M. (2015). Exploring Coral Reefs Using the Tools of Molecular Genetics. In C. Birkeland (Ed.), *Coral Reefs in the Anthropocene* (pp. 117–132). Dordrecht: Springer Netherlands.
Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, *14*, 1137–1145. Montreal, Canada.
Kraemer, A. C., Philip, C. W., Rankin, A. M., & Parent, C. E. (2019). Trade-offs direct the evolution of coloration in Galápagos land snails. *Proceedings. Biological Sciences / The Royal Society*, *286*(1894), 20182278.
Krehenwinkel, H., Kennedy, S. R., Rueda, A., Lam, A., & Gillespie, R. G. (2018). Scaling up DNA barcoding–Primer sets for simple and cost efficient arthropod systematics by multiplex PCR and Illumina amplicon sequencing. *Methods in Ecology and Evolution*, *9*(11), 2181-2193.
Krehenwinkel, H., Kennedy, S. R., Adams, S. A., Stephenson, G. T., Roy, K., & Gillespie, R. G. (2019). Multiplex PCR targeting lineage specific SNPs: A highly efficient and simple approach to block out predator sequences in molecular gut content analysis. *Methods in Ecology and Evolution*.
Krehenwinkel, H., Pomerantz, A., Henderson, J. B., Kennedy, S. R., Lim, J. Y., Swamy, V., ... & Prost, S. (2019). Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. *GigaScience*, *8*(5), giz006.
Kunstler, G., Lavergne, S., Courbaud, B., Thuiller, W., Vieilledent, G., Zimmermann, N. E., …

Coomes, D. A. (2012). Competitive interactions between forest trees are driven by species' trait hierarchy, not phylogenetic or functional similarity: implications for forest community assembly. *Ecology Letters*, *15*(8), 831–840.
Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, *36*(11), 1–13.
Lamy, T., Laroche, F., David, P., Massol, F., & Jarne, P. (2017). The contribution of species-genetic diversity correlations to the understanding of community assembly rules. *Oikos* , *126*(6), 759–771.
Laroche, F., Jarne, P., Lamy, T., David, P., & Massol, F. (2015). A neutral theory for interpreting correlations between species and genetic diversity in communities. *The American Naturalist*, *185*(1), 59–69.
Leibold, M. A., & Chase, J. M. (2017). *Metacommunity Ecology*. Princeton University Press.
Leidinger, L., & Cabral, J. S. (2017). Biodiversity Dynamics on Islands: Explicitly Accounting for Causality in Mechanistic Models. *Diversity*, *9*(3), 30.
Lemaire, L., Jay, F., Lee, I.-H., Csilléry, K., & Blum, M. G. B. (2016). Goodness-of-fit statistics for approximate Bayesian computation.
Lever, J., Krzywinski, M., & Altman, N. (2016). Points of significance: model selection and overfitting. *Nature Methods*, *13*(9), 703–704.
Likens, G., & Lindenmayer, D. (2018). *Effective ecological monitoring*. CSIRO publishing.
Lindenmayer, D. B., Likens, G. E., Andersen, A., Bowman, D., Bull, C. M., Burns, E., … Others. (2012). Value of long-term ecological studies. *Austral Ecology*, *37*(7), 745–757.
Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., & Corander, J. (2017). Fundamentals and Recent Developments in Approximate Bayesian Computation. *Systematic Biology*, *66*(1), e66–e82.
Lomolino, M. V. (2004). Conservation biogeography. *Frontiers of biogeography*, 293
Lozier, J. D., Aniello, P., & Hickerson, M. J. (2009). Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modeling. Journal of Biogeography. *Journal of Biogeography*, *10.1111/j*.
MacArthur, R. H., & Wilson, E. O. (1967). *Theory of Island Biogeography*. Princeton University Press.
Manceau, M., Lambert, A., & Morlon, H. (2015). Phylogenies support out-of-equilibrium models of biodiversity. *Ecology Letters*, *18*(4), 347–356.
Marchesini, A., Vernesi, C., Battisti, A., & Ficetola, G. F. (2018). Deciphering the drivers of negative species-genetic diversity correlation in Alpine amphibians. *Molecular Ecology*, *27*(23), 4916–4930.
Marquet, P. A., Allen, A. P., Brown, J. H., Dunne, J. A., Enquist, B. J., Gillooly, J. F., … West, G. B. (2014). On Theory in Ecology. *Bioscience*, *64*(8), 701–710.
McCarthy, J. K., Mokany, K., Ferrier, S., & Dwyer, J. M. (2018). Predicting community rank abundance distributions under current and future climates. *Ecography*.
McGill, B. J. (2003). A test of the unified neutral theory of biodiversity. *Nature*, *422*(6934), 881–885.
McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Benecha, H. K., … White, E. P. (2007). Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, *10*(10), 995–1015.
Meinshausen, N. (2006). Quantile Regression Forests. *Journal of Machine Learning Research: JMLR*, *7*(Jun), 983–999.
Miller, R. I. (1994). *Mapping the diversity of nature*. Springer Science & Business Media.

Miraldo, A., Li, S., Borregaard, M. K., Flórez-Rodríguez, A., Gopalakrishnan, S., Rizvanovic, M., … Nogués-Bravo, D. (2016). An Anthropocene map of genetic diversity. *Science*, *353*(6307), 1532–1535.
Missa, O., Dytham, C., & Morlon, H. (2016). Understanding how biodiversity unfolds through time under neutral theory. *Philosophical Transactions of the Royal Society B*, *371*(1691).
Monismith, S. G. (2006). Hydrodynamics of Coral Reefs. *Annual Review of Fluid Mechanics*.
Morlon, H. (2014). Phylogenetic approaches for studying diversification. *Ecology Letters*, *17*(4), 508–525.
Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, *403*(6772), 853–858.
Nee, S. (2005). The neutral theory of biodiversity: do the numbers add up? *Functional Ecology*, *19*(1), 173–176.
Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press.
Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, *76*(10), 5269–5273.
O'Dwyer, J. P., & Cornell, S. J. (2017). Cross-scale neutral ecology and the maintenance of biodiversity. *Scientific reports*, 8(1), 10200.
O'Dwyer, J. P., & Green, J. L. (2010). Field theory for biogeography: a spatially explicit model for predicting patterns of biodiversity. *Ecology Letters*, *13*(1), 87–95.
Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., O'hara, R. B., Simpson, G. L., … Wagner, H. (2010). Vegan: community ecology package. R package version 1.17-4.
Ovaskainen, O., Roy, D. B., Fox, R., & Anderson, B. J. (2016). Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution / British Ecological Society*, *7*(4), 428–436.
Overcast, I., Emerson, B. C., & Hickerson, M. J. (2019a). An integrated model of population genetics and community ecology. *Journal of Biogeography*, *46*(4), 816–829.
Overcast, I., Ruffley, M., Rosindell, J., Harmon, L., Borges, P., Chase, J., … Rominger, A. J. (2019b). What a MESS!: On the distribution of abundance, genetic, and functional diversity in ecological communities. In Prep.
Papadopoulou, A., Anastasiou, I., Spagopoulou, F., Stalimerou, M., Terzopoulou, S., Legakis, A., …. Vogler, A. P. (2011). Testing the Species–Genetic Diversity Correlation in the Aegean Archipelago: Toward a Haplotype-Based Macroecology? *The American Naturalist*, *178*(2), 241–255.
Papadopoulou, A., & Knowles, L. L. (2016). Toward a paradigm shift in comparative phylogeography driven by trait-based hypotheses. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(29), 8018–8024.
Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* , *20*(2), 289–290.
Parent, C. E., & Crespi, B. J. (2006). Sequential colonization and diversification of Galapagos endemic land snail genus Bulimulus (Gastropoda, Stylommatophora). *Evolution*, *60*(11), 2311–2328.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research: JMLR*, *12*(Oct), 2825–2830.
Pelletier, T. A., & Carstens, B. C. (2018). Geographical range size and latitude predict population genetic structure in a global survey. *Biology Letters*, *14*(1).

Pennell, M. W., & Harmon, L. J. (2013). An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Annals of the New York Academy of Sciences*, *1289*, 90–105.

Pertiwi, N. P. D., Malik, M. D. A., & Kholilah, N. (2018). Community Structure of Decapod Inhabit Dead Coral Pocillopora sp. in Pemuteran, Bali. In *IOP Conference Series: Earth and Environmental Science* (Vol. 116, No. 1, p. 012055). IOP Publishing.

Pettorelli, N., Wegmann, M., Skidmore, A., Mücher, S., Dawson, T. P., Fernandez, M., … Geller, G. N. (2016). Framing the concept of satellite remote sensing essential biodiversity variables: challenges and future directions. *Remote Sensing in Ecology and Conservation*, *2*(3), 122–131.

Phillips, S. J., Dudík, M., & Schapire, R. E. (2004). A Maximum Entropy Approach to Species Distribution Modeling. *Proceedings of the Twenty-First International Conference on Machine Learning*, 83.

Pigot, A. L., & Etienne, R. S. (2015). A new dynamic null model for phylogenetic community structure. *Ecology Letters*, *18*(2), 153–163.

Plaisance, L., Caley, M. J., Brainard, R. E., & Knowlton, N. (2011). The diversity of coral reefs: what are we missing? *PloS One*, *6*(10), e25026.

Plaisance, L., Knowlton, N., Paulay, G., & Meyer, C. (2009). Reef-associated crustacean fauna: biodiversity estimates using semi-quantitative sampling and DNA barcoding. *Coral Reefs* , *28*(4), 977–986.

Pontarp, M., Brännström, Å., & Petchey, O. L. (2019). Inferring community assembly processes from macroscopic patterns using dynamic eco evolutionary models and Approximate Bayesian Computation (ABC). *Methods in Ecology and Evolution*, *10*(4), 450–460.

Porter, T. M., & Hajibabaei, M. (2018). Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology*, *27*(2), 313–338.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems* , *9*(2), 181–199.

Rabosky, D. L., & Hurlbert, A. H. (2015). Species richness at continental scales is dominated by ecological limits. *The American Naturalist*, *185*(5), 572–583.

Reinke, B. A., Miller, D. A. W., & Janzen, F. J. (2019). What Have Long-Term Field Studies Taught Us About Population Dynamics? *Annual Review of Ecology*.

Ricklefs, R. E. (2006). The unified neutral theory of biodiversity: do the numbers add up? *Ecology*, *87*(6), 1424–1431.

Ricklefs, R. E., & Bermingham, E. (2001). Nonequilibrium diversity dynamics of the Lesser Antillean avifauna. *Science*, *294*(5546), 1522–1524.

Ridenhour, B. J., Brooker, S. L., Williams, J. E., Van Leuven, J. T., Miller, A. W., Dearing, M. D., & Remien, C. H. (2017). Modeling time-series data from microbial communities. *The ISME Journal*.

Rominger, A. J., Goodman, K. R., Lim, J. Y., Armstrong, E. E., Becking, L. E., Bennett, G. M., … Others. (2016). Community assembly on isolated islands: macroecology meets evolution. *Global Ecology and Biogeography: A Journal of Macroecology*, *25*(7), 769–780.

Rosenberg, N. A., & Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, *3*(5), 380–390.

Rosindell, J., & Cornell, S. J. (2007). Species–area relationships from a spatially explicit neutral model in an infinite landscape. *Ecology Letters*, *10*(7), 586–595.

Rosindell, J., Cornell, S. J., Hubbell, S. P., & Etienne, R. S. (2010). Protracted speciation

revitalizes the neutral theory of biodiversity. *Ecology Letters*, *13*(6), 716–727.
Rosindell, J., & Harmon, L. J. (2013). A unified model of species immigration, extinction and abundance on islands. *Journal of Biogeography*, *40*(6), 1107–1118.
Rosindell, J., Harmon, L. J., & Etienne, R. S. (2015). Unifying ecology and macroevolution with individual-based theory. *Ecology Letters*, *18*(5), 472–482.
Rosindell, J., Hubbell, S. P., He, F., Harmon, L. J., & Etienne, R. S. (2012). The case for ecological neutral theory. *Trends in Ecology & Evolution*, *27*(4), 203–208.
Rossetto, M., McPherson, H., Siow, J., Kooyman, R., van der Merwe, M., & Wilson, P. D. (2015). Where did all the trees come from? A novel multispecies approach reveals the impacts of biogeographical history and functional diversity on rain forest assembly. *Journal of Biogeography*, *42*(11), 2172–2186.
Satler, J. D., & Carstens, B. C. (2016). Phylogeographic concordance factors quantify phylogeographic congruence among co-distributed species in the Sarracenia alata pitcher plant system. *Evolution*.
Satler, J. D., & Carstens, B. C. (2017). Do ecological communities disperse across biogeographic barriers as a unit? *Molecular Ecology*, *26*(13), 3533–3545.
Satler, J. D., & Carstens, B. C. (2017). Do ecological communities disperse across biogeographic barriers as a unit? *Molecular Ecology*.
Sbrocco, E. J. (2014). Paleo-MARSPEC: gridded ocean climate layers for the mid-Holocene and Last Glacial Maximum: Ecological Archives E095-149. *Ecology*, *95*(6), 1710–1710.
Sbrocco, E. J., & Barber, P. H. (2013). MARSPEC: ocean climate layers for marine spatial ecology: Ecological Archives E094-086. *Ecology*, *94*(4), 979–979.
Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541.
Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., … Others. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences*, *109*(16), 6241–6246.
Schraiber, J. G., & Akey, J. M. (2015). Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, *16*(12), 727–740.
Schrider, D. R., & Kern, A. D. (2018). Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends in Genetics: TIG*, *34*(4), 301–312.
Shannon, C. (1948). A mathematical theory of communication, Part I, Part II. *Bell Systems Technical Journal*, *27*, 623–656.
Sheehan, S., & Song, Y. S. (2016). Deep Learning for Population Genetic Inference. *PLoS Computational Biology*, *12*(3), e1004845.
Smith, B. T., Seeholzer, G. F., Harvey, M. G., Cuervo, A. M., & Brumfield, R. T. (2017). A latitudinal phylogeographic diversity gradient in birds. *PLoS Biology*, *15*(4), e2001073.
Smith, E. P., Pontasch, K. W., & Cairns, J. (1990). Community similarity and the analysis of multispecies environmental data: A unified statistical approach. *Water Research*, *24*(4), 507–514.
Soininen, J., McDonald, R., & Hillebrand, H. (2007). The distance decay of similarity in ecological communities. *Ecography*, *30*(1), 3–12.
Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*,

134, 93–101.
Stadler, T. (2019). TreeSim: Simulating Phylogenetic Trees. *R Package*, *2.4*.
Stone, G. N., Lohse, K., Nicholls, J. A., Fuentes-Utrilla, P., Sinclair, F., Schönrogge, K., … Hickerson, M. J. (2012). Reconstructing community assembly in time and space reveals enemy escape in a Western Palearctic insect community. *Current Biology: CB*, *22*(6), 532–537.
Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*(8), 2045–2050.
Takahata, N., & Nei, M. (1985). Gene genealogy and variance of interpopulational nucleotide differences. *Genetics*, *110*, 325–344.
Tilman, D. (2004). Niche tradeoffs, neutrality, and community structure: a stochastic theory of resource competition, invasion, and community assembly. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(30), 10854–10861.
Title, P. O., & Bemmels, J. B. (2018). ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography*, *41*(2), 291–307.
Triantis, K. A., Rigal, F., Parent, C. E., Cameron, R. A. D., Lenzner, B., Parmakelis, A., … Cowie, R. H. (2016). Discordance between morphological and taxonomic diversity: land snails of oceanic archipelagos. *Journal of Biogeography*, *43*(10), 2050–2061.
Turner, W., Spector, S., Gardiner, N., Fladeland, M., Sterling, E., & Steininger, M. (2003). Remote sensing for biodiversity science and conservation. *Trends in Ecology & Evolution*, *18*(6), 306–314.
Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F., & De Clerck, O. (2012). Bio-ORACLE: a global environmental dataset for marine species distribution modelling. *Global Ecology and Biogeography*, *21*(2), 272–281.
UNEP-WCMC, WorldFish Centre, WRI, TNC. (2010). *Global distribution of warm-water coral reefs, compiled from multiple sources including the Millennium Coral Reef Mapping Project*. UNEP World Conservation Monitoring Centre Cambridge, UK.
Uyeda, J. C., Hansen, T. F., Arnold, S. J., & Pienaar, J. (2011). The million-year wait for macroevolutionary bursts. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(38), 15908–15913.
Valente, L., Etienne, R. S., & Dávalos, L. M. (2017). Recent extinctions disturb path to equilibrium diversity in Caribbean bats. *Nature Ecology & Evolution*, *1*(2), s41559–016 – 0026.
Valente, L. M., Phillimore, A. B., & Etienne, R. S. (2015). Equilibrium and non equilibrium dynamics simultaneously operate in the Galápagos islands. *Ecology Letters*.
Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., … Others. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, *25*(4), 929–942.
Vellend, M. (2005). Species diversity and genetic diversity: parallel processes and correlated patterns. *The American Naturalist*, *166*(2), 199–215.
Vellend, M. (2010). Conceptual synthesis in community ecology. *The Quarterly Review of Biology*, *85*(2), 183–206.
Vellend, M. (2016). *The Theory of Ecological Communities (MPB-57)*. Princeton University Press.
Vellend, M., Lajoie, G., Bourret, A., Múrria, C., Kembel, S. W., & Garant, D. (2014). Drawing

ecological inferences from coincident patterns of population- and community-level biodiversity. *Molecular Ecology*, *23*(12), 2890–2901.
Venkataraman, A., Bassis, C. M., Beck, J. M., Young, V. B., Curtis, J. L., Huffnagle, G. B., & Schmidt, T. M. (2015). Application of a neutral community model to assess structuring of the human lung microbiome. *mBio*, *6*(1).
Vergnon, R., van Nes, E. H., & Scheffer, M. (2012). Emergent neutrality leads to multimodal species abundance distributions. *Nature Communications*, *3*, 663.
Wagner, C. E., Harmon, L. J., & Seehausen, O. (2014). Cichlid species-area relationships are shaped by adaptive radiations that scale with area. *Ecology Letters*, *17*(5), 583–592.
Webb, C. O., Ackerly, D. D., McPeek, M. A., & Donoghue, M. J. (2002). Phylogenies and Community Ecology. *Annual Review of Ecology and Systematics*, *33*(1), 475–505.
Werner, F. E., Cowen, R. K., & Paris, C. B. (2007). Coupled biological and physical models: present capabilities and necessary developments for future studies of population connectivity. *Oceanography* , *20*(3), 54–69.
Xue, A. T., & Hickerson, M. J. (2017). Multi DICE: R package for comparative population genomic inference under hierarchical co demographic models of independent single population size changes. *Molecular Ecology Resources*.
Zellweger, F., Baltensweiler, A., Ginzler, C., Roth, T., Braunisch, V., Bugmann, H., & Bollmann, K. (2016). Environmental predictors of species richness in forest landscapes: abiotic factors versus vegetation structure. *Journal of Biogeography*, *43*(6), 1080–1090.