Towards a genetic theory of island biogeography: Inferring processes from multidimensional community-scale data

Isaac Overcast1,2 | Guillaume Achaz3,4,5 | Robin Aguilée6 | Carmelo Andújar7 | Paula Arribas7 | Thomas J. Creedy8,9 | Evan P. Economo10,11 | Rampal S. Etienne12 | Rosemary Gillespie13 | Claire Jacquet14,15 | Flora Jay16 | Susan Kennedy10 | Henrik Krehenwinkel17 | Amaury Lambert1,18 | Emmanouil Meramveliotakis19 | Víctor Noguerales7,19 | Benoit Perez-Lamarque1 | George Roderick13 | Haldre Rogers20 | Megan Ruffley21 | Isabel Sanmartin22 | Alfred P. Vogler8,9 | Anna Papadopoulou19 | Brent C. Emerson7 | Hélène Morlon1

1 Institut de Biologie de l’ENS (IBENS), Ecole Normale Supérieure, CNRS, INSERM, Université PSL, Paris, France
2 Department of Vertebrate Zoology, American Museum of Natural History, New York, New York, USA
3 Institut Systématique Evolution Biodiversité (ISYEB), Muséum National d’Histoire Naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, Paris, France
4 SMILE Group, CIRB, UMR 7241, Collège de France, CNRS, INSERM, Paris, France
5 Éco-anthropologie, Muséum National d’Histoire Naturelle, CNRS UMR 7206, Université de Paris, Paris, France
6 UMR5174 (Laboratoire Evolution et Diversité Biologique), CNRS, Université Toulouse III Paul Sabatier, Toulouse, France
7 Island Ecology and Evolution Research Group, Instituto de Productos Naturales y Agrobiología, CSIC (IPNA-CSIC), Santa Cruz de Tenerife, Spain
8 Department of Life Sciences, Natural History Museum, London, UK
9 Department of Life Sciences, Imperial College London, Ascot, UK
10 Biodiversity and Biocomplexity Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan
11 Radcliffe Institute for Advanced Study, Harvard University, Cambridge, Massachusetts, USA
12 Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands
13 Department of Environmental Science, Policy, and Management, University of California, Berkeley, California, USA
14 Department of Aquatic Ecology, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
15 Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zürich, Switzerland
16 Université Paris-Saclay, CNRS UMR9015, Inria, Laboratoire Interdisciplinaire des Sciences du Numérique, Orsay, France
17 Department of Biogeography, Trier University, Trier, Germany
18 Center for Interdisciplinary Research in Biology (CIRB), Collège de France, CNRS UMR 7241, INSERM U 1050, Université PSL, Paris, France
19 Department of Biological Sciences, University of Cyprus, Nicosia, Cyprus
20 Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, USA
21 Carnegie Institution for Science, Stanford, California, USA
22 Real Jardín Botánico (RJB), CSIC, Madrid, Spain

Correspondence
Isaac Overcast, Institut de Biologie de l’ENS (IBENS), Ecole Normale Supérieure, CNRS, INSERM, Université PSL, Paris 75005, France.
Email: overcast@biologie.ens.fr

Abstract

Background: MacArthur and Wilson’s theory of island biogeography has been a foundation for obtaining testable predictions from models of community assembly and for developing models that integrate across scales and disciplines. Historically, however,
these developments have focused on integration across ecological and macroevolutionary scales and on predicting patterns of species richness, abundance distributions, trait data and/or phylogenies. The distribution of genetic variation across species within a community is an emerging pattern that contains signatures of past population histories, which might provide an historical lens for the study of contemporary communities. As intraspecific genetic diversity data become increasingly available at the scale of entire communities, there is an opportunity to integrate microevolutionary processes into our models, moving towards development of a genetic theory of island biogeography.

**Motivation/goal:** We aim to promote the development of process-based biodiversity models that predict community genetic diversity patterns together with other community-scale patterns. To this end, we review models of ecological, microevolutionary and macroevolutionary processes that are best suited to the creation of unified models, and the patterns that these predict. We then discuss ongoing and potential future efforts to unify models operating at different organizational levels, with the goal of predicting multidimensional community-scale data including a genetic component.

**Main conclusions:** Our review of the literature shows that despite recent efforts, further methodological developments are needed, not only to incorporate the genetic component into existing island biogeography models, but also to unify processes across scales of biological organization. To catalyse these developments, we outline two potential ways forward, adopting either a top-down or a bottom-up approach. Finally, we highlight key ecological and evolutionary questions that might be addressed by unified models including a genetic component and establish hypotheses about how processes across scales might impact patterns of community genetic diversity.

**Keywords**
- biodiversity dynamics, community ecology, community genetic diversity, metabarcoding, metagenomics, population genetics, stochastic biodiversity models

## 1 | INTRODUCTION

General interest in understanding the processes that have given rise to current patterns of biodiversity has spurred the development of diverse models to explain these patterns. One of the more influential is MacArthur and Wilson’s equilibrium theory of island biogeography (ETIB; MacArthur & Wilson, 1963, 1967). Although originally developed to understand biodiversity patterns on islands, the ETIB has been used more generally to predict community patterns, thanks to the analogy between island and community assemblages, both of which result from limited dispersal from a larger species pool. The ETIB has catalysed >50 years of research on models that represent how species from regional or global pools (i.e., the metacommunity) assemble into local communities (Chesson, 2000; Hubbell, 2001; Leibold & Chase, 2017; Tilman, 2004; Vellend, 2016). These community assembly models provide explicit predictions about patterns that can be measured in the field, such as species abundance distributions (SADs; McGill et al., 2007), spatial turnover (Nekola & White, 1999), temporal turnover (Magurran et al., 2010) and species–area relationships (Lomolino, 2000). The availability of other types of data has spurred efforts to develop models that also generate community-scale predictions for such data, including species traits (Jacquet et al., 2017; Xu et al., 2020), interaction networks (Gravel et al., 2011; Thébault & Fontaine, 2010) and phylogenetic trees (Jabot & Chave, 2009; Morlon et al., 2011; Sanmartín et al., 2008; Valente et al., 2015). These models often require integration across disciplines to represent processes operating over different organizational levels (individuals/genotypes/species) and spatial scales, within the community or within the metacommunity, such as environmental filtering, interspecific interactions, trait evolution, speciation and extinction. Such unified models are challenging to build, but they allow a powerful inference of multiple processes potentially operating at and interacting across various scales.

One dimension of biodiversity that remains largely absent from predictions of such integrative models is the distribution of population genetic variation across species within communities (Ellegren & Galger, 2016). This shortcoming means that community-level data for genetic diversity cannot be used to reinforce inferences of ecological and evolutionary processes using island biodiversity models. This is unfortunate, because community-scale genetic
data provide a record of the formation of ecological communities on a population-genetic (intermediate) time-scale (on the order of tens to tens of thousands of years), which can complement information obtained from other community-scale patterns (Overcast et al., 2020). For example, the strength of the correlation between species abundance and genetic diversity might provide an indication of community age, because all species in young communities (e.g., those on recently formed volcanic islands) should have low genetic diversity (but see Caujapé-Castells et al., 2017), whereas abundant species from old communities (e.g., those on relatively ancient volcanic islands) might have accumulated more genetic diversity than rare species (Overcast et al., 2020). Genetic diversity might also capture signatures of historical changes in population size and cycles of genetic connectivity that drive incipient species into and through the “grey zone” of speciation (Roux et al., 2016; Salces-Castellano et al., 2020). Importantly, in the context of considering processes across organizational levels and spatial scales, genetic diversity provides a snapshot of population history at a time-scale that can shed light on the link between microevolutionary, ecological and macroevolutionary processes, which encompass patterns of intraspecific allele frequencies, species abundances and phylogenetic relationships, respectively. Therefore, the development of unified models of biodiversity that provide predictions for community-scale genetic data, along with other community-level predictions, has the potential to provide insight into numerous key ecological and evolutionary questions (Box 1).

Data availability is no longer a major obstacle to the integration of microevolutionary processes in inferences from unified models of biodiversity. Obtaining genetic data at the community scale has long been hindered by logistical constraints, but metagenetic methods, including whole-organism community DNA metabarcoding (Creedy et al., 2021; Deiner et al., 2017), environmental DNA (Cordier et al., 2020; Taberlet et al., 2012) or mitochondrial metagenomics (Crampton-Platt et al., 2016; Gómez-Rodríguez et al., 2015, 2017), are increasingly removing this constraint. Conventionally, these approaches have been implemented to approximate community profiles at the species level, but their potential also to generate population-level genetic data for entire communities has recently been revealed (Andújar et al., 2021; Arribas et al., 2021; Elbrecht et al., 2018; Schloissnig et al., 2013; Tsuji et al., 2020; Turon et al., 2020). As advances in sequencing technology continue to produce higher yields, metagenetics is increasingly being used for the characterization and monitoring of biodiversity in marine, terrestrial and freshwater environments. As a consequence, community-level genetic data are accumulating in massive online repositories (e.g., GenBank; https://www.ncbi.nlm.nih.gov/genbank), such that it is becoming increasingly possible to study intraspecific genetic variation across multiple communities from regional to global scales (i.e., “macrogenetics”; Leigh et al., 2021; Miraldo et al., 2016). One focus, therefore, is to encourage development of models that will provide an explicit biogeographical context (Edwards et al., 2022) for macrogenetics by integrating community-scale intraspecific genetic variation into mechanistic models to understand and predict how biodiversity accumulates across spatial scales and levels of organization.

The field of population genetics has produced countless models to infer historical processes of individual (or closely related) species (e.g., Kingman, 1982; Wright, 1931). However, these models are not well adapted to analyse community-scale data, because their focal unit of study is individual species. Comparative phylogeographical models aggregate multiple individual-population genetic models for co-distributed species (Arbogast & Kenagy, 2001; Edwards et al., 2022), but they lack an ecological (community) context. There are a handful of precedent efforts to develop community-level models that include predictions for the genetic diversity of all constituent species. Johnson et al. (2000) derived a population genetic model of island biogeography that allowed estimation of rates of migration, extinction and speciation and that made explicit predictions about how genetic divergence relates to island area and distance from the mainland (Ricklefs & Bermingham, 2004). Inspired by the unified neutral theory of biodiversity (UNTB; Hubbell, 2001), which is an individual-based neutral model of community assembly following ETIB, Laroche et al. (2015) developed a model that predicts the species genetic diversity correlation (SGDC), which measures the correlation between the genetic diversity of a focal taxon and species richness in the local community (Vellend, 2005). Empirical studies across a broad array of taxa have found either positive (Lamy et al., 2013; Papadopoulou et al., 2011) or negative genetic diversity–species richness correlations (Marchesini et al., 2018; Xu et al., 2016), consistent with predictions by Laroche et al. (2015). Finally, Overcast et al. (2020) developed the massive eco-evolutionary synthesis simulations (MESS) model, which combines Hubbell’s UNTB with the neutral model of population genetics (Kimura, 1983). MESS provides predictions for the correlation between genetic diversity and abundance and for the species genetic distribution (SGD; Overcast et al., 2019), an analogue to the SAD, but where species abundance is replaced by a measure of species genetic diversity.

Given these few existing developments, the power of unified models and trends towards increased access to community-level population genetic data, we aim to promote the development of process-based biodiversity models that predict community genetic diversity patterns together with other community-scale patterns. For simplicity, we focus on genetic diversity patterns obtained from single-locus data, consistent with data that are typically generated within metabarcoding studies, but note that modelling of multilocus or metagenomic data is also possible and would only increase the resolution of model predictions. We review current models for studying biodiversity dynamics from the perspective of community ecology, microevolution and macroevolution, with a focus on their applicability to island-like systems. We highlight the strengths and weaknesses of these existing models, in terms of integrating across organizational levels and predicting realistic multidimensional community data including a genetic component. Finally, we discuss ongoing and potential future efforts to unify models operating at different organizational levels, in order to predict multidimensional community-scale data better.
BOX 1  Why are unified models including community-scale genetic variation important for understanding the processes that generate and maintain biodiversity?

Many open questions remain concerning how the biodiversity we observe in nature has accumulated. For example, how do the dynamics of adaptation, competition, speciation and colonization result in the communities that we observe today? And how have local environmental conditions and human-mediated disturbance modulated these processes? These questions can be addressed, in part, without unified models; for example, community ecology studies often suppose that the dynamics of community assembly play out on time-scales so short that speciation can be ignored, and macroevolutionary investigations rarely consider processes such as mutation and the accumulation of genetic diversity. The main value of unified models is precisely to facilitate investigation of feedback between such processes (i.e., addressing questions such as: how does in situ speciation reshape communities locally, and how does it impact the genetic diversity of the two descendant species and of all other species in the community?). A growing awareness of the importance of such feedback has bolstered interest in unified models (Cavender-Bares & Wilczek, 2003; McGill, 2010; Pennell & Harmon, 2013). Unified models might also increase our ability to infer which, among a series of alternative eco-evolutionary scenarios, is most likely to have produced the data we observe, by simultaneously leveraging the complementary information contained in multiple axes of biodiversity data. For example, scenarios that are not distinguishable from species abundance distributions alone might be distinguishable when also including community genetic data. We identify here a non-exhaustive set of questions for which models predicting individual (or few) types of data have had some, albeit limited, success in resolving, but for which the development of unified models would provide further advances.

- **Factors modulating commonness and rarity within communities.** How much do colonization times and early establishment demographics determine which species are common and which are rare in contemporary communities (e.g., tropical forests in Hawaii; Craven et al., 2019)? Have abundance relationships been maintained over time, and have historical changes in abundance followed deterministic (niche partitioning) or stochastic (neutral assembly) processes (e.g., tropical forests in Panama; Wang et al., 2013)? Ecological models of community assembly predicting only contemporary species abundances are unable to answer such questions, but unified models including a population genetic component for predicting the past and present genetic diversity of contemporary species could do so.

- **Factors promoting species richness within communities.** How much does in situ speciation contribute to the diversity of local communities (e.g., Caribbean Anolis lizards; Losos & Schluter, 2000)? How much do the dynamics of speciation and extinction at the metacommunity and community scales versus ecological limits imposed by local environmental conditions (e.g., mediated by total abundance in the community) explain the diversity of local communities (e.g., Rabosky & Glor, 2010)? Standard community assembly models do not account for diversification dynamics, whereas standard macroevolutionary models consider ecological limits at the clade rather than community level. These processes could be untangled with unified models of lineage diversification with a spatial local community component to generate nested local/global phylogenetic, richness and abundance patterns.

- **Impact of community context on genetic diversity.** Genetic diversity of individual species is the outcome of processes that take place within a community context, yet this context is rarely considered in population genetic models. For example, dispersal limitation between, competition (or other types of interactions) within, and adaptation to ecological conditions of islands or island-like habitat patches will all contribute to shape individual (and collective) species genetic diversity (e.g., Malagasy herpetofauna; Kuhn et al., 2022). Unified models of colonization, abundance and competition/interaction that predict abundance and genetic diversity for all species can provide such a community context for population genetic models.

- **Diversification and genetic diversity.** How much do metacommunity diversification dynamics explain community-scale genetic diversity? How does in situ speciation impact the genetic diversity of descendant species, and how does this contribute to shaping genetic diversity across the entire community (e.g., Neotropical cichlids; Kautt et al., 2020)? Answering these questions requires a unified model of macroevolution, community assembly and population genetics.

2  EXISTING MODELS OF COMMUNITY, GENETIC AND COMPARATIVE BIODIVERSITY DATA

A model that aims to incorporate predictions of genetic diversity, as an additional dimension of community-scale patterns, together with abundances, traits and/or network structure, species richness and phylogenetic patterns (Figure 1) needs to incorporate processes ranging from ecological (e.g., dispersal, environmental filtering, demographic processes modulated by environmental factors and intra- and interspecific interactions) to microevolutionary (e.g., mutation, genetic drift and selection) and macroevolutionary (e.g., speciation, extinction and phenotypic evolution) (Figure 2). We expect that these processes will modulate community-scale genetic diversity
patterns (Box 2). If one is additionally interested in predicting spatially explicit patterns, similar to the species–area relationship (SAR; Durrett & Levin, 1996), the distance decay of similarity relationship (Morlon et al., 2008) or equivalent patterns for networks (Brose et al., 2004; Galiana et al., 2018), traits (Penone et al., 2016; Weinstein et al., 2014), genetic diversity (Basilga et al., 2015; Papadopoulou et al., 2011) and phylogenetic diversity (Morlon et al., 2011), then the model needs to incorporate spatially explicit dispersal and, potentially, environmental variation. Finally, prediction of temporal patterns requires following how patterns unfold through time under the chosen model (e.g., Missa et al., 2016). A rich suite of ecological, microevolutionary and macroevolutionary models that predict some of these patterns already exists. We begin with a short overview of some of these models, evaluating which patterns they are able to predict and whether the models provide realistic predictions, before moving into integration of genetic data into community-level models. We consider models that can be fitted to data using a variety of statistical techniques, ranging from likelihood-based to simulation-based inference (Box 3).

2.1 | Ecological models derived from the theory of island biogeography

The ETIB (MacArthur & Wilson, 1963, 1967) has been instrumental in ecology, inspiring many of the models that have subsequently been used in the field to predict community-scale data. Under the assumption that species are functionally equivalent (i.e., there are no interspecific differences in colonization or extinction rates), the theory predicts equilibrium richness as a product of immigration, which is dictated by distance from the mainland source, and extinction, which is affected by island size. Beyond predicting biodiversity patterns on islands, the ETIB has been used to predict community patterns, thanks to the analogy between continent–island and metacommunity–community systems (Hanski, 1998). For example, the ETIB naturally predicts SARs (Lomolino, 2000), and in general, these provide a reasonably good fit to empirical systems (Warren et al., 2015). Body-size distributions can be predicted by modelling body-size variation (e.g., the allometric theory of island biogeography; Jacquet et al., 2017),
under the assumption that island area and isolation will have differential impacts on colonization and extinction probability for large- versus small-bodied organisms. Food web properties, such as connectance (the fraction of realized interactions), can also be predicted (Gravel et al., 2011; Harmon et al., 2019). Several ETIB-inspired models can predict community-scale phylogenetic patterns, including DAISIE (Valente et al., 2015, 2020), which models immigration, extinction and speciation in an island/archipelago context, and DAMOCLES (Pigot & Etienne, 2015), which incorporates standard birth–death diversification processes at the regional scale with colonization and local extinction (but not speciation) processes at the local scale. The ETIB and its descendants are lineage based and, as such, they do not make predictions of abundance or genetic diversity. These models are also spatially implicit, in the sense that, although the rate of colonization can be a proxy for geographical distance, physical geography and landscape features are not modelled, hence they do not provide predictions of continuous spatial biodiversity patterns, such as the distance decay of similarity relationship.

The UNTB (Hubbell, 2001; Rosindell et al., 2011) is an individual-based model inspired by the ETIB, which models stochastic demographic processes in a small local community, within which all individuals of all species are functionally equivalent and which is at an equilibrium between local extinction and colonization from a large source pool of potential immigrants (metacommunity). The UNTB provides equilibrium predictions of species abundance distributions at local scales (Etienne, 2005, 2007) that are a good fit to temporally static SADs (Hubbell, 2001; Matthews & Whittaker, 2014) and can be relaxed to allow modelling of non-equilibrium abundance dynamics (Manceau et al., 2015; Missa et al., 2016; Overcast et al., 2019, 2020; Rosindell & Harmon, 2013), facilitating predictions of temporal patterns, such as changes in community composition through time (e.g., Kalyuzhny et al., 2015). Body-size distributions can be predicted within a UNTB-derived model by allowing variation in survival probability given body-size differences, while otherwise retaining ecological equivalence (e.g., O’Dwyer et al., 2009). Further extensions of the UNTB provide predictions for SADs on both sides of an interaction network, in addition to the structure of the network (Canard et al., 2012; Mallet et al., 2020). Spatially explicit temporal patterns of α- and β-diversity have also been modelled under static (Durrett & Levin, 1996; O’Dwyer & Green, 2010) or fluctuating (Gotelli et al., 2009; Jabot et al., 2020; Pontarp & Wiens, 2017; Rangel et al., 2007) environmental conditions, although how environmental stochasticity impacts spatial and temporal abundance patterns remains to be investigated fully. The UNTB also makes predictions of phylogenetic patterns at local
and metacommunity scales that fit the tree imbalance observed within empirical phylogenies reasonably well (Davies et al., 2011; Jabot & Chave, 2009). These have been improved by incorporation of protracted (Rosindell et al., 2010) or genealogical (Manceau et al., 2015; Rosindell et al., 2015) models of speciation. The UNTB has also served as a starting point for more general (including non-neutral) models of community structure (e.g., Haegeman & Etienne, 2017; Rosindell & Phillimore, 2011).

2.2 Microevolutionary models for genetic diversity data

Population genetic models seek to represent the fundamental forces of evolution acting within species, including mutation, genetic drift, natural selection and migration, in the context of demographic history over time-scales of tens or hundreds to thousands of generations. They are typically applied to DNA sequence data from
Numerous statistical approaches exist for performing inference from real data, making trade-offs associated with model complexity, analytical and computational tractability, statistical power and accuracy, quantification of uncertainty, and interpretability. When analytical likelihood formulas are available, these methods are a classical and powerful approach for fitting models to data. Likelihood-based approaches involve computing the probability of the data (e.g., a phylogeny or an abundance distribution) under a given model characterized by a set of parameters. In practice, likelihood methods do not always scale well to large datasets, in which case a composite likelihood approach approximating the full likelihood as a product of independent sub-likelihoods can be adopted (Varin et al., 2011). Alternatively, a data-augmentation approach (which consists of enhancing the data with unobserved events, such as extinctions) can simplify the computation of the likelihood and increase efficiency (Mallet & Morlon, 2020; Quintero & Landis, 2020). When used in combination with Bayesian statistics, such as Markov chain Monte Carlo (MCMC) sampling, likelihood-based methods naturally quantify uncertainty. Likelihood-based methods are the gold standard for most comparative phylogenetic methods (Morlon, 2014) and have also been used extensively to fit data to ecological (Rosindell et al., 2011) and microevolutionary (Kuhner et al., 1998; Pritchard et al., 2000) models.

For models of increasing complexity where likelihoods cannot be computed, a simulation-based approach can be adopted in an approximate Bayesian computation (ABC) framework (Beaumont, 2010; Csilléry et al., 2010). ABC approximates posterior model probabilities and parameter estimate distributions by comparing summary statistics calculated on the real data with identical summary statistics calculated on simulated data from tens of thousands or millions of simulations. The posteriors are constructed based on the proportion of simulated models and parameter values of a small fraction of retained simulations that are "closest" to the observed data. Classical ABC can be computationally demanding, because the number of simulations needed to sample parameter space grows exponentially with the number of parameters, although recent advances in combining machine learning with ABC have reduced this burden (Pudlo et al., 2015; Raynal et al., 2019). Sequential sampling methods (Beaumont et al., 2009; Lenormand et al., 2013) increase efficiency by iteratively sampling from regions of parameter space with the highest likelihood. ABC methods have seen broad adoption for numerous applications in population genetics (Beaumont, 2010). They have been used less frequently in ecological and macroevolutionary studies, particularly to infer model parameters (Jabot & Chave, 2009; Jabot & Lohier, 2016; Janzen et al., 2015; Overcast et al., 2019; Pontarp et al., 2019; Slater et al., 2012).

Machine learning (ML) methods offer a likelihood-free approach and are gaining significant attention for application to ecological and evolutionary questions (Schrider & Kern, 2018; Sheehan & Song, 2016). In the context of inference from multidimensional biodiversity data, we focus on supervised methods that allow both model selection and parameter estimation, unlike unsupervised "clustering" methods, such as principal component analysis. All supervised methods share a common process: (1) generate simulations under different models and parameter values; (2) train the ML on a subset of simulations; (3) test the ML model on held-out simulations to evaluate accuracy and recall; and (4) confront the ML with observed data to perform model classification and parameter estimation.

Example applications of supervised ML include demographic model selection using population genetic summary statistics (e.g., Smith et al., 2017) and inference of historical population sizes using whole-genome data (e.g., Sanchez et al., 2020). ML has seen increasing use in macroevolutionary studies, for example to estimate phylogenies (Bhattacharjee & Bayzid, 2020) and for correlation of evolutionary rates (Tao et al., 2019). Early adoption of ML for ecological questions has been concentrated within the species distribution modelling literature (Gobeyn et al., 2019), with comparatively few applications in community ecology and with those being primarily focused on predicting microbial community structure (Thomas et al., 2018) and interaction patterns in ecological networks (Pichler et al., 2020). ML methods can be accurate and computationally efficient for inference under highly complex models. They avoid the curse of dimensionality (high-dimensional data are handled well), but suffer from reduced interpretability (but see Azodi et al., 2020) and difficulty in quantifying uncertainty around predictions, although this is an active area of research (Alaa & Van Der Schaar, 2020; Coulston et al., 2016).
The mean difference in \( \pi \) across the landscape (e.g., Haller & Messer, 2018), modeling patterns of genetic variation, in addition to trait values, fitness differences and selection on quantitative traits, allowing for flexibility in modeling complex evolutionary scenarios incorporating intense than coalescent approaches, yet allow for much greater genetic offset based on ecological and genetic data (Fitzpatrick & Keller, 2015; Jay et al., 2012), akin to species distribution modelling, but at the intraspecific level. Nonetheless, population genetic models are predominantly designed to investigate processes within or among closely related species and, as such, they most often do not provide predictions for patterns above the species level (i.e., phylogenetic relationships or community-scale patterns of abundance).

Efforts to scale up microevolutionary models from the species level to multiple co-distributed species have occurred in comparative phylogeography (Arbogast & Kenagy, 2001; Edwards et al., 2022). Classic comparative phylogeographical studies attempted to infer histories of community assembly (e.g., Hewitt, 2000) from shared patterns of genetic variation across the landscape. The emergence of model-based comparative phylogeography (Hickerson et al., 2007; Huang et al., 2011; Xue & Hickerson, 2020) provided for explicit modelling of joint patterns of community-scale genetic variation under co-demographic processes, such as changes in population size (Gehara et al., 2017), while also accounting for vicariance and dispersal (Smith et al., 2014; Thom et al., 2020). These models implement a hierarchical, simulation-based approach, whereby coalescent simulations per focal taxon are parameterized conditional on hyperparameters that describe the assemble-wide process (e.g., the proportion of co-expanding or co-diverging taxa). Although they describe patterns above the species level, comparative phylogeographical models lack an explicit ecological context and also do not consider or model phylogenetic relationships among co-distributed taxa.
2.3 | Macroevolutionary models for comparative data

The primary focus of macroevolutionary models is to represent the processes of diversification (speciation and extinction) and trait evolution for entire clades over time-scales of millions of years (Cornault & Sammartín, 2022; Morlon, 2014; Pennell & Harmon, 2013). They are typically applied to so-called “comparative data”; that is, the phylogenies and traits of extant species for taxonomically restricted groups of species. By focusing on taxonomically restricted groups of species, macroevolutionary models are not adapted a priori to modelling community-scale phylogenetic and trait patterns, yet they constitute an appropriate framework for communities arising from in situ radiations, as is the case for many radiations on isolated island archipelagos (Harmon et al., 2003; Mahler et al., 2010) or lakes (Janzén & Etienne, 2017).

Macroevolutionary models consider diversification and phenotypic evolution either separately (in this case, the processes are modelled hierarchically) or jointly (in this case, traits can influence diversification and vice versa; Maddison et al., 2007). The latter models, referred to as state-dependent speciation–extinction (SSE) models, have been reviewed elsewhere (Fitzjohn et al., 2012; Drury et al., 2009; Maddison & Fitzjohn, 2015). Recent advances in diversification modelling produce more realistic tree shapes by allowing variation in rate among lineages and through time (Maliet et al., 2019; Rabosky et al., 2013). Additional processes incorporated into lineage-based macroevolutionary models include prorated speciation (Etienne & Rosindell, 2012), clade-wide diversity-dependent speciation or extinction (Etienne et al., 2012; Mahler et al., 2010), inter- and/or intraspecific competition (Aristide & Morlon, 2019; Clarke et al., 2017; Drury et al., 2016; Xu et al., 2020), co-evolution across interacting lineages (Manceau et al., 2017) and the influence of environmental variation on macroevolutionary rates (Clavel & Morlon, 2017; Condamine et al., 2013). Some of these models include the modelling of dispersal events (Goldberg et al., 2011; Landis et al., 2013, 2021; Ree et al., 2005) and can be co-opted to make predictions about spatial patterns of biodiversity. They can potentially incorporate the effect of competition on trait evolution (Drury et al., 2018) or jointly on trait evolution and range occupancy (Quintero & Landis, 2020), allowing for the prediction of patterns of trait distribution and endemism (for a more complete treatment of dispersal in macroevolutionary models, see Hackel & Sammartin, 2021). These models incorporate more processes and can predict patterns that more closely match empirical ones, but they do not predict patterns other than the phylogenies and/or traits of extant taxa.

By construction, lineage-based models do not represent the individual-level processes of demography, mutation, genetic drift and both allele and trait selection that contribute to fluctuating species abundances and the accumulation of intraspecific genetic and trait diversity. The few individual-based models that also make predictions of macroevolutionary patterns include the UNTB, as already discussed above (Davies et al., 2011; Jabot & Chave, 2009; Missa et al., 2016), the model of Manceau et al. (2015), which is grounded in UNTB but relaxes both the zero-sum assumption and the point-mutation mode of speciation, and models that evolve species abundances and the traits of individuals along a phylogeny, which is either taken as data (Nuismer & Harmon, 2015) or simulated dynamically within the model (Duchen et al., 2020; McPeek, 2008). In principle, these latter models generate SADs while also modelling intraspecific genetic variation (Manceau et al., 2015) and inter- (McPeek, 2008) or intraspecific (Duchen et al., 2020; Nuismer & Harmon, 2015) trait variation; however, the predicted patterns have yet to be explored fully.

3 | TOWARDS A UNIFIED MULTIDIMENSIONAL COMMUNITY-SCALE THEORY OF ISLAND BIOGEOGRAPHY INCLUDING A GENETIC COMPONENT

To model multidimensional community data arising from processes across different organizational levels and spatial scales and including the genetic component, which was absent from many of the models described above, one must choose whether to model from the top down or bottom up. Top-down begins at the macroevolutionary scale and works down to the ecological and then microevolutionary scale (Figure 4). Bottom-up begins at the microevolutionary scale and naturally builds up to higher scales by letting patterns at all scales emerge from the underlying processes. Both strategies have advantages and disadvantages. Top-down models tend to scale well and can model clades of biologically meaningful phylogenetic diversity (hundreds of tips) but are, perhaps, less flexible in terms of modelling very complex processes. In addition, top-down models tend to be amenable to analytical likelihood expressions (Box 3), which makes them very computationally efficient. Bottom-up models tend to be based on individuals, which makes them costly to operate at biologically realistic scales, in particular up to the macroevolutionary scale, the trade-off being that they are more mechanistic (McGill, 2019). Any of these models can be fitted to multidimensional community-scale data to answer a variety of questions about the generation and maintenance of biological diversity (Box 1), using appropriate statistical inference techniques (Box 3).

3.1 | Top-down approaches

3.1.1 | From the macroevolutionary to the ecological scale

One top-down approach to combine macroevolutionary and ecological models consists of creating hybrid models that merge the metacommunity-scale dynamics of speciation and extinction with an ETIB-like community assembly process (Figure 4a). DAMOCLES (Pigot & Etienne, 2015), which incorporates standard birth–death diversification processes at the regional scale with colonization and local extinction processes at the local scale, is such a model.
The spatial macroevolutionary models of diversification discussed above could also, in principle, serve this role, if they are configured to generate metacommunity–community patterns. For example, this could be achieved within the two-area GeoSSE model (Goldberg et al., 2011) by setting the dispersal rate from one of the areas (the community) to the other (the metacommunity) to zero. This would lead to a model representing processes very similar to those included in DAMOCLES, but with the additional advantage of representing speciation both within and between areas. Taking into consideration the metacommunity birth–death processes in DAISIE (Valente et al., 2015) would be an alternative approach to developing a similar hybrid model. These models illustrate the feasibility and usefulness of macroevolutionary–ecological models and pave the way for integrating more advanced ETIB-like community assembly processes (e.g., those accounting for interspecific interactions) within more advanced macroevolutionary models (e.g., those accounting for diversification rate heterogeneities, interspecific interactions and/or palaeoenvironmental fluctuations).

Unified neutral theory of biodiversity-like models can be used instead of ETIB-like models at the community scale as an alternative approach to creating hybrid models (Figure 4b), as implemented in MESS (Overcast et al., 2020). A key consideration for hybrid models is how to account for species abundances in macroevolutionary models in a principled way. In MESS, the metacommunity-scale phylogeny is generated with a birth–death lineage-based speciation process, and contemporary abundances are sampled from a logarithmic series distribution and assigned to species randomly. A more process-based approach would be to evolve abundances directly along phylogenetic branches as the lineage-based macroevolutionary process unfolds. In such a model, extinction events would naturally arise when species abundance falls below one, or any realistic minimum viable population size estimate. Speciation events would arise as the realization of a Poisson process, with rates following any of the models considered in macroevolution. Finally, there are several ways to evolve abundances. They could, for example, evolve as a Brownian process, with splits by random fission at each
speciation event (Etienne & Haegeman, 2011; Hubbell & Lake, 2003). Alternatively, species-specific growth rates controlling changes in abundance could be evolved along branches as a Brownian process. The feedback from the ecological to the macroevolutionary scale might be realized by incorporating a protracted model of speciation (Rosindell et al., 2010), with the abundance structure of incipient lineages modulating the duration of speciation, whether complete speciation occurs and/or whether incipient species go extinct. Once a global phylogeny with abundances is generated by such a process, local community abundance distributions and phylogenetic relationships can be sampled following a classical UNTB individual-based simulation procedure.

From these neutral models, which assume that species are demographically and functionally equivalent, and following the same reasoning, several useful extensions are possible. In particular, existing macroevolutionary models of trait evolution could be used at the metacommunity scale to generate trait histories along phylogenetic branches, in addition to species-specific trait values at the tips, from which local community trait distributions can be sampled. In the non-neutral version of DAMOCLES, for example, traits influence colonization rates through habitat filtering or competition with residents. In the non-neutral version of MESS, relaxation of the ecological neutrality assumption allows the influence of biotic and abiotic interaction to modulate individual survival. Although MESS has explored simple models of competition and environmental filtering, variation within an ecologically relevant trait (e.g., body size) could influence species abundances through more biologically realistic interspecific interactions, such as mutualism or facilitation. Recent lineage-based models of phenotypic trait evolution that model inter- and intraspecific variation jointly (Gaboriau et al., 2020; Kostikova et al., 2016) could also be included within macro-ETIB and macro-UNTB approaches, allowing for more fine-grained modelling of ecological interactions.

3.1.2 From the ecological to the microevolutionary scale

After developing top-down models linking macroevolutionary and ecological processes, the next step is to extend these to the microevolutionary scale. To achieve this, one possibility is to constrain microevolutionary models per species to be contingent on divergence times and abundance histories predicted by the linked macroevolutionary–ecological model (e.g., Overcast et al., 2020). These constraints are justified by the observed correlation between census and effective size in natural populations (James & Eyre-Walker, 2020). This approach assumes that there is no post-divergence gene flow between sister species and that all lineages have sorted at the time of sampling. These simplifying assumptions are reasonable under low rates of diversification and/or small effective population sizes with respect to branching times. However, in cases where post-divergence gene flow cannot be ignored, species ages can be scaled to generation time, and migration rates among lineages can be modelled in a fully parameterized coalescent framework for the entire community.

This general approach can be applied to either macro-ETIB or macro-UNTB macroevolutionary–ecological models. As a prelude to this, Johnson et al. (2000) derived formulas based on the ETIB for average genetic divergence between island and mainland taxa as functions of time, island area and distance from the mainland. However, their model did not consider community-scale intraspecific diversity. ETIB models can naturally provide predictions for colonization times, but are not directly adapted to predicting fluctuations in species abundances required for genetic diversity predictions. The crux would be to use presupposed correlations between species abundances and characteristics of the species represented in these models. For example, the allometric theory of island biogeography (Jacquet et al., 2017) accounts for body-size differences across species and could be coupled with a documented relationship between body size and abundances (Damuth's law; Damuth, 1981) to provide community-level genetic diversity predictions at mutation–drift equilibrium. The trophic theory of island biogeography (Gravel et al., 2011) accounts for dietary breadth differences across species; by assuming that abundance scales with dietary breadth (i.e., generalists have greater abundance than specialists), community-level genetic diversity predictions at mutation–drift equilibrium could be obtained.

Macro-UNTB macroevolutionary–ecological models generate direct predictions of fluctuations in abundance within communities, which can be used to constrain microevolutionary models at the scale of individual species. Two such individual-based simulation approaches have been developed (Laroche et al., 2015; Overcast et al., 2019). Both are inspired by the UNTB, but they differ in the model used to predict genetic diversity. Laroche et al. (2015) paired a UNTB-inspired patch-based model with a forward-in-time mutational process for one focal taxon in the local community to predict the relationship between species diversity and intraspecific genetic diversity (the SGDC). This model could be generalized to predict genetic diversity for all the species in the local community, rather than for only an individual taxon. Overcast et al. (2019) developed gimmeSAD, which pairs a forward-time UNTB continent–island model tracking abundance and colonization time per species with a backward-in-time coalescent model per species to predict SADs and the community-level genetic diversity distribution jointly (SGD; Figure 2a). Another approach that suggests itself is to parameterize hierarchical comparative phylogeographical models directly using expectations from a macro-UNTB model. Indeed, lacking both a colonization and a community-assembly process, comparative phylogeographical models assume that all species have occupied the local landscape for the same amount of time and that the effective population sizes of individual species are sampled independently. Conditioning hyperparameter distributions on theoretical macro-UNTB expectations of species age and abundance would allow relaxation of these assumptions. Ultimately, these different approaches will result in a model representing the same processes.
These outlines of ETIB- and UNTB-based linked ecological and microevolutionary processes suggest numerous further avenues for development. For example, an eco-microevolutionary model of interactions among species of co-distributed guilds (e.g., plants–pollinators or hosts–parasites) could allow joint predictions of community genetic diversity and network structure, a possibility foreshadowed by recent work (Bunnefeld et al., 2018; Satler & Carstens, 2017; Stone et al., 2012). Linked spatially explicit models (Currat et al., 2019; Haller & Messer, 2019) would allow for ecologically informed predictions for novel community-scale spatial patterns of genetic diversity, such as the genetic diversity–area relationship, the correlation between genetic and species β-diversity, and the distance decay of genetic diversity (i.e., community-scale isolation by distance).

3.1.3 | From the macroevolutionary to the microevolutionary scale

There are existing conceptual frameworks that go directly from the macroevolutionary to the microevolutionary scale, without a community assembly component. Models that consider both interspecific and intraspecific divergences, such as the multispecies coalescent (MSC; Degnan & Rosenberg, 2009), the generalized mixed Yule coalescent (GMYC; Pons et al., 2006), the polymorphisms-aware phylogenetic model (POMO; De Maio et al., 2015) and automatic barcode gap discovery (ABGD; Puillandre et al., 2012), provide such a framework, although initially developed as species-delimination or phylogenetic inference methods. Simulations under these frameworks model: (1) a species tree using typical Yule or birth–death macroevolutionary processes; (2) gene trees (MSC, POMO and ABDG) or intraspecific divergences (GMYC), using typical coalescent microevolutionary processes, with or without incomplete lineage sorting; and (3) potentially, lay down mutations on the gene trees. Obtaining community-scale patterns from these existing frameworks would require sampling species from the species tree, which could be done following either the macro-ETIB or macro-UNTB approach detailed above. Although this would lead to a model that converges to the macro–eco–micro model described above, it might provide additional possibilities. For example, pairing the multi-species coalescent with a mutational model underlying quantitative trait loci can generate phenotypic trait distributions in the presence of gene tree discordance (Mendes et al., 2018).

3.2 | Bottom-up unification approaches

A key obstacle for building bottom-up biodiversity models is to model speciation by the accumulation of genetic incompatibilities between populations (Orr, 1995). Among these microevolutionary models, those incorporating trait variation governed by an explicit genetic component can link demographic and speciation processes to differences in fitness within an ecological context and predict patterns above the species level (e.g., for a radiating clade; Gavrilets & Losos, 2009). Several such models have been proposed (e.g., Aguillé et al., 2018; Garwood et al., 2019; Gavrilets & Vose, 2005), all of which are individual-based birth–death processes that model genetic evolution via mutation and recombination within an ecological context. Including biotic interactions in such models is a pathway to more biologically relevant bottom-up models (e.g., Pontarp & Wiens, 2017; Thompson et al., 2020). Given that these models are primarily concerned with speciation, patterns of intraspecific genetic diversity are generated, but have not been evaluated. Additionally, these genetically explicit models tend to represent closed systems suitable for modelling in situ radiations, but lack a colonization process adapted to modelling the community structure of spatially isolated local communities (but see Aguillé et al., 2018; Gascuel et al., 2015, which model space explicitly). Opportunities remain for the development of models underpinned by a genetically explicit speciation process to make predictions about linked patterns of SADs, the distributions of traits and genetic variation, and phylogenetic and network structure.

Alternative attempts to model speciation from the accumulation of population genetic incompatibilities have been made using models either derived from or inspired by the UNTB. Instead of the traditional phenomenological models of speciation (e.g., point mutation and various derivations; Hubbell & Lake, 2003; Rosindell et al., 2010), speciation processes governed by an underlying process of genomic evolution and accumulation of incompatibility have been incorporated (e.g., de Aguiar et al., 2009; Hagen et al., 2021; Manceau et al., 2015; Marin et al., 2020; Melián et al., 2010). In these neutral models, speciation proceeds by the accumulation of genetic differentiation and is governed by a threshold effect (Gavrilets, 2004), whereby subpopulations that are differentiated by a given amount of genetic divergence are considered to be species. Although these models typically implement a fixed speciation threshold, sampling from a distribution of speciation thresholds per species would better align with observations of the “grey zone” of speciation in empirical systems (Roux et al., 2016). These models make joint predictions about community genetic diversity, abundance and phylogenetic structure (Costa et al., 2019) and, in some cases, spatial genetic patterns (Baptestini et al., 2013; de Aguiar et al., 2009), although these patterns have yet to be explored fully. These models are neutral with respect to genetics, hence there are neither fitness differences nor selection, and therefore, no direct effect of genetics on species abundance. Models incorporating both neutral and adaptive genetic components have been proposed (e.g., Aguillé et al., 2018; Gascuel et al., 2015), and investigating their predictions of community-scale genetic diversity and fitting these to empirical data will be fruitful avenues of future research.

4 | CONCLUSIONS

The distribution of genetic diversity, within and among species within ecological communities, is an emerging area of interest, but
methodological development is needed to shed light on how ecological and evolutionary processes structure such diversity. Models of community-scale genetic diversity are in their infancy, with some limited but promising progress. Our focus in this perspective has been the modelling of single-locus data derived from metabarcoding studies, which are the predominant community-scale genetic datasets that are becoming widely obtainable. As sequencing throughput continues to increase, it should become possible to move towards the generation of single-locus community genetic data to whole-community reduced representation sequencing datasets, consisting of thousands or tens of thousands of independent anonymous loci (Andrews et al., 2016), or even beyond, to community whole-genome data (e.g., Forster et al., 2016). For some model systems, such community-scale genomic data exist already (e.g., Neotropical cichlids; Kautt et al., 2020), providing opportunities to elucidate the role of genomic architecture in driving eco-evolutionary dynamics (Rudman et al., 2018). Population genetic models capturing complex genomic processes, including multilocus selection and fine-scale genome-wide mutation and recombination rate variation, allow for modelling these new types of genetic data (Haller & Messer, 2019; Kelleher et al., 2016). Integration of these models within the unifying frameworks highlighted here has the potential to provide even greater insights into the processes shaping biodiversity across scales. Our assessment of different potential approaches for building models that integrate processes across scales highlights that, despite the diversity of models developed in community ecology, population genetics and macroevolution, distinct envisioned efforts to expand the generation of single-locus community reduced representation sequencing datasets, consisting of thousands or tens of thousands of independent anonymous loci (Andrews et al., 2016), or even beyond, to community whole-genome data (e.g., Forster et al., 2016). For some model systems, such community-scale genomic data exist already (e.g., Neotropical cichlids; Kautt et al., 2020), providing opportunities to elucidate the role of genomic architecture in driving eco-evolutionary dynamics (Rudman et al., 2018). Population genetic models capturing complex genomic processes, including multilocus selection and fine-scale genome-wide mutation and recombination rate variation, allow for modelling these new types of genetic data (Haller & Messer, 2019; Kelleher et al., 2016). Integration of these models within the unifying frameworks highlighted here has the potential to provide even greater insights into the processes shaping biodiversity across scales. Our assessment of different potential approaches for building models that integrate processes across scales highlights that, despite the diversity of models developed in community ecology, population genetics and macroevolution, distinct envisioned efforts to expand the application across scales lead to convergent models. This leads us to conclude on the optimistic note that the long-sought, truly unified theory of biodiversity might be within close reach.

AUTHOR CONTRIBUTIONS
All authors contributed substantially to conceptual development. I.O. and H.M. wrote the first draft of the manuscript, and all authors contributed to manuscript revisions.

ACKNOWLEDGEMENTS
We thank the Morlon group for thoughtful and useful discussions. We acknowledge the support of the European Union’s Horizon 2020 research and innovation programme through grant agreement number 810729.

FUNDING INFORMATION
This work is a product of the iBioGen project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 810729.

CONFLICT OF INTEREST
The authors declare no competing interests.

DATA AVAILABILITY STATEMENT
Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

REFERENCES


ORCID
Isaac Overcast https://orcid.org/0000-0001-8614-6892
Robin Aguilée https://orcid.org/0000-0002-3388-9921
Claire Jacquet https://orcid.org/0000-0002-7913-0689
Benoit Perez-Lamarque https://orcid.org/0000-0001-7112-7197
Brent C. Emerson https://orcid.org/0000-0003-4067-9858
Hélène Morlon https://orcid.org/0000-0002-3195-7521

https://orcid.org/0000-0002-3388-9921


Etienne, R. S., & Rosindell, J. (2012). Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. Systematic Biology, 61(2), 204–213.


BIOSKETCH

Members of the iBioGen consortium and collaborators promote methodological unification and theoretical synthesis for studying island biodiversity using community-scale DNA sequence data, such as can be obtained from contemporary metabarcoding studies. To this end, we develop and apply novel methods that unify across scales of biological organization for inferring ecological and evolutionary processes from multidimensional biodiversity data that include a genetic component. Further information about the iBioGen project and its partners may be found on the project website: https://www.ibiogen.eu.