Integrating Across Dimensions of Biodiversity: Developing Novel Theory and Informatics

1 Intellectual Merit

1.1 Aims

As phylogenetic and functional approaches transcend species-based estimates of diversity, we seek new answers to the age-old question [1], "What limits diversity?" At the intersection of evolutionary history, environmental opportunity, functional capacity, and genetic adaptive potential, a suite of fundamental processes (drift, migration, mutation, speciation, competition, environmental filtering) and emergent properties (e.g. community structure, ecosystem productivity) lead to largely unexplored feedbacks between species, functional, phylogenetic, and population genetic diversity. Thanks to initiatives such as NSF's Dimensions of Biodiversity (DoB) Program, we have gained enough insight into feedbacks among these multiple dimensions of biodiversity to know that they complicate previously simple interpretations of single dimensions. For instance, traits such as body size may directly shape responses to climate change by impacting population demography, thus driving biogeographic patterns of lineage distribution, which themselves define the conditions to which lineages adapt [2, 3]. However, we have yet to to add empirically tractable complexity to our predictive models of biodiversity and thus not gained a mechanistic understanding of the interactions among diversity dimensions. There is now an unprecedented opportunity to advance understanding of basic and emergent interactions among dimensions, as rich, large-scale, multi-dimensional data become available through DoB and other similar initiatives.

Because measurements of diversity reflected in the scale of DoB projects range from small archipelagos to whole biomes, we propose to synthesize and integrate across projects to understand the general rules governing biodiversity. Making sense of heterogeneous and multi-dimensional data within and among such diverse biogeographical systems requires having (1) the **informatic ability** to synthesize and share data from different origins and (2) the **quantitative tools** to extract insight from synthesized data about the complex, multi-dimensional processes underlying biodiversity patterns. **We propose to fill these two gaps by (1) building an open source platform to manage and query diverse streams of biodiversity-relevant information, and (2) developing a stochastic model of biodiversity that incorporates a range of mechanisms that jointly predicts species abundance, trait distributions, phylogenetic structure, and population genetic diversity.** Through this model we will be able to quantify the contribution of three distinct classes of processes to the structure of biodiversity across disparate systems: (i) Stochastic immigration, speciation, and extinction, (ii) species interactions with the biotic and abiotic environment, and (iii) historical contingencies.

To understand the universality of these processes, we will (3) test our model across five biogeographic systems for which multi-dimensional biodiversity data are now available, due largely to DoB efforts. They include two naturally insular systems with differing degrees of isolation, age, and adaptive radiation (the Hawaiian island chronosequence and the marine lakes of Palau) and three forest systems with key gradients of diversity and evolutionary history, spanning scales of historic continental disjunctions (temperate forests of North America and Asia), entire biomes (Amazonia), and large regions of unique endemism (the coastal Atlantic Forest of Brazil; Fig. 1). The two island projects additionally afford an explicitly temporal test of our model's dynamics—via the sediment cores from Palau's marine lakes and the island chronosequence of

Hawaii. By selectively augmenting data gathered by these DoB projects with new field observations, traits and sequences we will be able to test our model and begin to verify the universality of processes regulating the diversity of life. The new data we collect will add substantially to the Tree of Life, particularly for under-sampled clades such as arthropods and marine invertebrates.

1.2 Background

Understanding the generation and maintenance of biodiversity is critical to manage and mitigate the effects of anthropogenic change on the diversity of living systems; without intervention, recovery from the current crisis [4] could take tens of millions of years [5]. Despite the lack of concrete answers to what limits diversity, decades of research have generated important insights into the processes driving diversity across multiple dimensions in specific systems. Still, mechanistic hypotheses from these studies have yet to be quantitatively and jointly tested across systems with a rigorous framework where their universality can be evaluated.

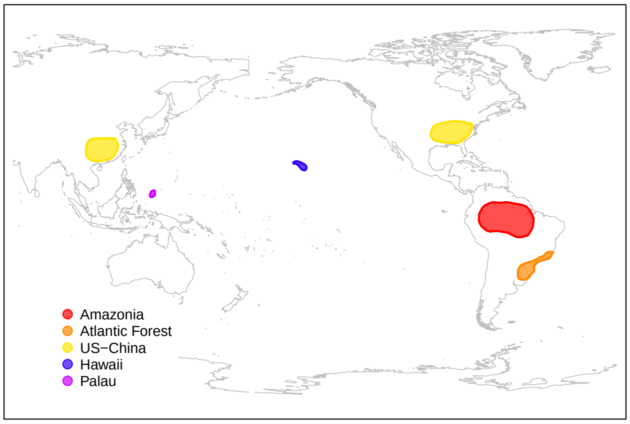

Figure 1: Map of the included DoB projects illustrating their breadth of geographic region and scale.

Current eco-evolutionary theory is not up to this task. Quantitative models inspired by the Equilibrium Theory of Island Biogeography [6] and the Neutral Theory of Biodiversity and Biogeography [7], for instance, place emphasis on the equilibrium between rates of immigration, speciation, and extinction. However, while empirical work on islands and latitudinal gradients have demonstrated the significant role of immigration, speciation, and extinction histories [8, 9], these neutral theories of eco-evolutionary drift are limited in scope and predictive power. For example, interspecific interactions and the environment have been shown empirically to be important in determining patterns such as the latitudinal diversity gradient [10–12]. Moreover, neutral drift theories do not accurately capture temporal dynamics [13, 14] and ecological interactions have been flagged as potential drivers of the fast dynamics observed in real communities that neutral theories fail to accurately predict [13, 15]. Conversely, models based on detailed ecological interactions may be so over-parameterized as to render them uninterpretable [16]. Current theory fails to predict multiple dimensions of biodiversity, which can lead to very different processes producing indistinguishable predictions for single dimensions [17–20].

The modeling perspectives from both neutral-like eco-evolutionary drift and ecological interaction-based coexistence and diversification are typically applied in equilibrium [7, 21–23]. However, empirical examples abound of non-equilibrium hysteresis. They include island biogeography [24], differences in otherwise similar biomes across Asia and North America [25, 26], and communities assembled from refugia after glacial retreat [27–29]. The potential for real systems to be far from equilibrium—increasingly relevant in the Anthropocene [30]—necessitates a modeling framework that can evaluate non-equilibrium states.

Evaluating the generality of eco-evolutionary mechanisms, from classical equilibrial models to non-equilibrium dynamics, in driving patterns of species-level, phylogenetic, population genetic, and functional trait diversity, necessitates advances in informatics and theory. We must collect, manage, and analyses multi-dimensional data on (i) the abundance of species, (ii) their geo-referenced occurrences, (iii) ecologically-relevant traits, (iv) phylogenetic and population genetic histories, and (v) the environment.

1.3 Informatic platform for dimensions of biodiversity data

Large scale, system-based collaborative projects—including many funded by the DoB program—often have difficulty making the best use of the volume and types of data they generate. Data management strategies are usually developed independently by each team or laboratory, often varying across teams. Members of collaborating laboratories frequently work on their own locally stored datasets, with local domain expertise and data administration and curation. Smaller or more narrowly focused projects can afford to absorb the administrative and cognitive overhead of fully integrating data platforms and management. However, larger projects—both in scope and scale—are significantly negatively impacted by this. The challenge is especially acute for projects with multiple PIs distributed across the globe (e.g. US-China and US-Brazil partnerships), as well as those generating vast quantities of heterogeneous data. We identified these major barriers to multi-dimensional biodiversity data management for projects that span biogeographic regions as part of a recent NSF-funded workshop on DoB data (PIs Carnaval and Soltis)

Various isolated repositories exist for warehousing some of these data independently. Initiatives such as iDigBio, GBIF, Dryad, GloBI, and NCBI provide some facility for depositing, cataloguing, and accessing occurrence records, sequence, morphological and ecological data, and have been extensively used by the many DoB projects funded to date. However, in nearly all cases the linkages that would crucially support re-integration of these data are missing, which limits the ability to derive broader understanding across projects. While all the mentioned repositories are essential services to the global scientific community, they each have internal missions that are divergent from the needs of biodiversity project teams for data integration. Further, we are limited by a lack of metadata standards that are both general and comprehensive for ecological datasets, hampering integration. We argue that simply depositing data into such repositories is not enough. A key step forward is building a framework where data resources are standardized and linked such that the right types of data can be fed into modeling frameworks. To better understand biodiversity pattern and process, DoB projects and other biodiversity research teams should be able to ask questions about the multivariate forces that influence multiple systems; thus there is a critical need to assure that datasets interoperate. **We will build a proof-of-concept platform that supports storing data and linking its shared attributes for broadest possible use and also enhances modeling frameworks, such as those discussed below**. The outcome will be a scalable, open source knowledgebase composed of relational databases in the Cloud, open repositories, and functional linkages that crosscut and connect all dimensions of spatiotemporal biodiversity data.

1.4 Theory to integrate dimensions of biodiversity

Hickerson and his team have recently developed a joint model of species abundance and community-level genetic variation through time [31]. Their work explicitly links a forward-time neutral model of community assembly with a backward-time neutral coalescent model of population genetics, fitting the model to data with approximate Bayesian computation [ABC; 32]. **We propose to extend this model to jointly predict species abundance, phylogeny, population genetic, and functional trait distributions within ecological communities (Fig. 2)**. To achieve this theoretical advance, we will incorporate speciation [33] and trait evolution to existing work in neutral theory [22, 34]. By including population genetic and phylogenetic information when estimating the model and assessing its fit to data, we will capture signatures of temporal dynamics and scrutinize our work exactly where neutral theory fails most noticeably [13]. We will also extend the neutral model to incorporate non-neutral processes, such as density dependence [35], and the ability of fitness to evolve within and between species [15], all of which will help to reconcile modeled temporal dynamics with reality. We will link these coarse-grained processes, for which mathematical details have been well-studied, to our trait data. For that, we will use model parameters that capture

Process	(symbol)	Neutral	Environment x Trait	Trait x Trait
birth	(b)	equals fitness	equals fitness	positive function of fitness and negative function of competition
death	(d)	constant	constant	constant
immigration	(ν)	constant	positive function of potential fitness in colonizing environment	negative function of potential competition in colonizing community
fitness	(r)	constant	positive function, depending on coefficient ρ, of environmental suitability of trait	constant
environmental filtering	(ρ)	0	strength of environmental filtering	0
competition	(a_{ij})	0	0	positive function of trait-overlap
speciation	(λ)	constant	positive function of environmental suitability of incipient trait	negative function of trait-overlap of incipient species
mutation	(μ)	constant	constant	constant
trait evolution	(σ, θ)	Brownian	Ornstein-Uhlenbeck	Ornstein-Uhlenbeck or Brownian

Table 1: Model parameters.

trait-based interactions with other species [36, 37] and environments [38–40]. Table 1 shows how these processes enter into our model.

We will fit this joint model of multiple biodiversity dimensions (species abundance, phylo- and population genetic, and trait) to empirical data collected across the five DoB projects using both ABC and new machine learning approaches [41]. By evaluating how well the model fits empirical data across systems we will understand whether a parsimonious set of processes encapsulates pervasive patterns in the diversity of life. By quantifying how the specific best-fit parameter values of the model change across systems we will understand the contribution of these different processes to the structuring of biodiversity across disparate systems, and spatial and temporal scales.

2 Proposed Research

2.1 Research objectives and hypotheses

We have three objectives:

R1 To design and implement an informatics platform that can synthesize and facilitate the sharing of diverse data streams associated with DoB projects and similar biodiversity initiatives. We will make this platform open source and usable both through a graphical user interface and an R [43] package. We will implement it across the five DoB projects represented by co-PIs on this current proposal (Hawaii: Rominger and Gillespie; Palau: Dawson; US-China: Soltis; Amazonia: Guralnick, Owens, and Robinson; Atlantic Forest: Carnaval and Hickerson).

R2 To build a joint model for the multiple dimensions of biodiversity—species abundance, phylo- and population genetic, and functional. This model will build off on-going work (by Rominger, Hickerson, Harmon, and Chase) to combine macroecological predictions with phylo- and population

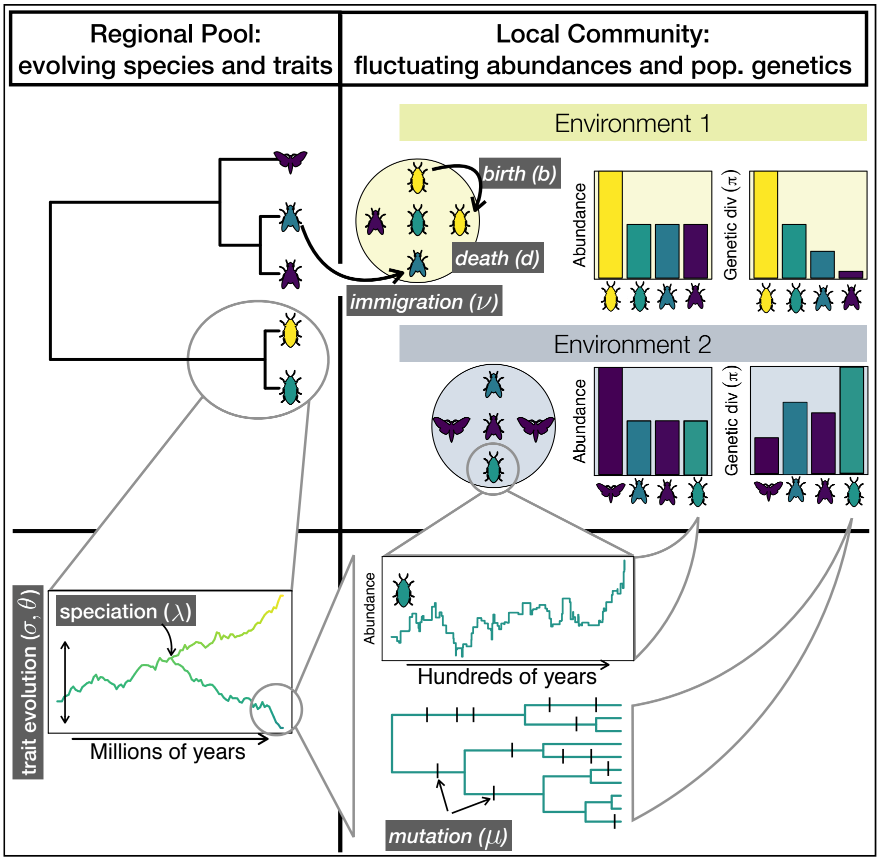

Figure 2: The processes and data types captured by our model. Different species are represented by shape and color; colors of species correspond to trait values. Lighter background colors represent different environments, whose suitability is represented by color matching. All processes are highlighted as white text with a dark background. Processes already implemented in [42] are *italicized*. Regionally, speciation and trait evolution produce functionally novel species. Locally, birth, death, and immigration—which can all depend on trait-based competition and environmental filtering (Table 1)— drive abundance fluctuations, which combined with mutation, produce genetic diversity. The top two panels show a snapshot in the simulation in which the regional pool is static and locally *either* immigration *or* birth replace a death event. The bottom two panels show the respective time dynamics. The modeled processes generate species abundances, population genetic diversity across species, phylogenetic relationships between species, and trait distributions across species.

genetic measures. Model fitting methods based on ABC and machine learning will be created. The entire modeling framework will be released as open source software.

R3 To test this joint model across five diverse systems for which data has been synthesized by the informatics platform. We have included in this proposal five mature DoB projects whose data represent compelling unit tests for our informatics platform, and whose underlying evolution and ecology will test the generality of our model. We have identified data types (e.g. abundance, genetic and/or trait data) where new targeted sampling within these projects will greatly improve our ability to test our model and document the raw diversity of life.

The modeling approach we develop in **R2** allows us to address five non-mutually exclusive hypotheses central to ecology and evolution at an unprecedented quantitative caliber and grounded in real data aggregated (**R1**) from incredibly diverse systems (**R3**). Our approach casts these complex hypotheses as clear correlations (see Fig. 3) between the parameters of our model (Table 1) when fit to data across systems. Because the model encapsulates a range of processes and jointly predicts many data types, interrogating its behavior represents a rigorous, multi-faceted test of hypotheses [18–20]. Outliers to these predictions, or their complete refutation, will shed light on what idiosyncratic system-specific processes drive these deviation. These new insights in turn will inform future work in hypothesis development, modeling, and data gathering.

H1: Isolation fosters diversification. This has long been assumed and documented for some specific systems [34, 44, 45]. However, it remains to be assessed how key isolation is to the striking examples of endemism and radiation at the phylogenetic, genetic, and functional levels, and

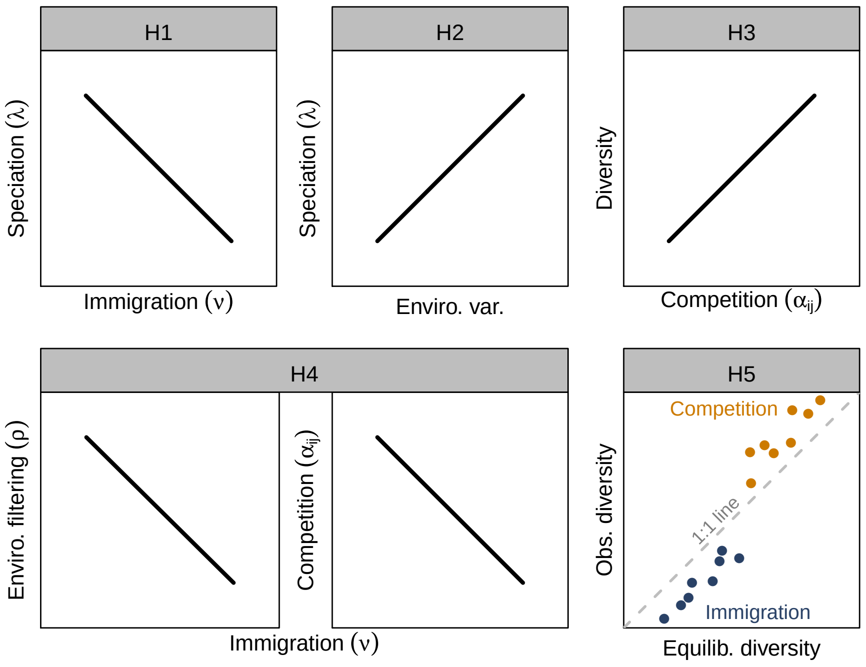

Figure 3: Correlations between inferred parameters (see Table 1) of our model under the assumption that our five hypotheses are true. In **H5**, "competition" and "immigration" refer to communities inferred to be competition-driven versus immigration-driven, respectively.

whether other processes such as strong ecological interactions are also drivers (see H3). Looking across systems, H1 predicts a negative correlation between estimated immigration (ν) and speciation (λ; Fig 3).

H2: Environmentally-driven speciation is important across systems, but depends on intrinsic dispersal ability of taxa. Whether environmental gradients can promote diversification remains an open debate [11, 46, 47]. Our model can test whether interactions between traits and environments is necessary to predict patterns of species abundance, phylo- and population genetic diversity, and trait distributions. After accounting for the effect of H1, this hypothesis predicts a positive correlation between estimated speciation (λ) and the environmental heterogeneity of a study region (Fig. 3), which we can measure spatially and, for many studies, temporally as well.

H3: High diversity systems result from increased strength of species interactions. This is a classic hypothesis about both the latitudinal diversity gradient [12] and ecosystem complexity versus stability [48, 49]. If we find that high diversity systems are systematically better-fit by models with strong trait-based species interactions, this lends strong support to the hypothesis. Looking across systems, we predict a positive correlation between diversity (of all biodiversity dimensions) and the magnitude of estimated trait-based competition (α_{ij}; Fig. 3).

H4: Assembly by immigration (versus *in situ* speciation) leads to weaker species interactions and thus more apparently neutral diversity patterns. If high diversity systems are the result of strong species interactions, do these interactions arise more readily in communities assembled by immigration or *in situ* speciation? We posit the latter, due to increased opportunity for co-evolutionary arms race-like dynamics [50, 51]. Conversely, immigration-assembled communities will draw from pools of the most generalist colonists and thus appear effectively neutral. Thus, this hypothesis predicts a negative correlation between inferred immigration (ν) and the magnitude of parameters capturing purely non-neutral processes (i.e. trait-mediated species-environment [ρ] and species-species [α_{ij}] interactions; Fig. 3).

H5: Most systems have not reached an equilibrium in diversity. Thus historical contingencies are critical to contemporary diversity patterns. We posit that the mechanistic importance of the previous four hypotheses is heightened by systems being out of equilibrium. For systems in equilibrium, the effect size of dynamical processes will be diminished [52]. The time to equilibrium, and whether or not a system has reached it, is something we can quantify. We can then use the best-fit parameters to calculate the hypothetical equilibrium diversity. Across systems we predict discrepancies between observed diversity and the eventual diversity reached at the modeled equilibrium (Fig. 3). Whether this discrepancy is toward higher or lower observed diversity than equilibrium diversity yields interesting additional and complementary hypotheses that we will explore:

> **H5a** Systems constrained by immigration [e.g. arthropod communities on young substrates in Hawaii; 53] will show observed diversities lower than equilibrium.
>
> **H5b** Systems driven by strong interactions will show diversity overshoots in which non-equilibrium diversity is higher than equilibrium diversity because interactions initially drive diversification beyond the intrinsic carrying capacity of a system [54].

2.2 Methods

2.3 Informatics pipeline

Existing informatic platforms and metadata standards now only manage subsets of the data we envision synthesizing: abundance and/or occurrence, trait, genetic, and environmental. Major challenges are (1) allowing for short-term data entry along with long-term data storage and sharing; (2) integrating abundance, sequence, phylogeny, and environmental data into a data management environment primarily intended for specimen data; and (3) educating researchers about proper data management. This last barrier we address with substantial curriculum development, detailed in section 3.1 on training and Broader Impacts.

It is critical to insure that trait, occurrence, and abundance data are standardized properly across all projects. We will use entity-quality relationships to standardize key traits shared across the tree of life, namely body size and metabolic rate. Those traits specific to groups will also be curated using a similar approach. Both individual and species-level traits will be synthesized following a new relational data model developed by Co-PI Guralnick. Abundance data will be standardized using Humboldt Core [55] which captures necessary inventory metadata. Efforts at standardizing content will assure both model-ready data, reproducibility, and re-use for others.

We will build an open source R package that allows users to manage and synthesize their data via three cloud-based and/or open repositories: (1) Google Sheets for data entry and short-term curation; (2) Amazon Relational Database Services (RDS) for mid-term curation; and (3) open repositories detailed below for long-term curation.

Data for such large initiatives as DoB projects necessitate that many participants be involved in the collection and curation of data. Cloud-based platforms are ideal for distributed data collection because they can be accessed remotely and track the version history produced by multiple users. To this end, Rominger, working with undergraduate students in focused mentoring associations, has developed an open source R package [56] that facilitates the entry and error checking of data using R and Google Sheets, either through custom scripts or a graphical user interface built with Shiny [57]. This approach will be extended to facilitate the entry of other data types including abundances, phylogenies, and raw sequence data. For data that cannot be easily recorded in tabular format (e.g. phylogenies), links to those files will be recorded instead.

Once data entry is complete we will house data on Amazon RDS and provide functionality in our R package and Shiny app for data transfer to Amazon RDS. Amazon RDS provides a scalable solution that reduces administration costs and provides access to content via query APIs. It is also easy to connect to such databases via R [58]. We will assure open access to resources via R and the Shiny app to team members and the larger community, thus hastening re-use.

Solutions for open, long-term curation of all data types of interest to our project have already been pioneered. The missing component is practical software to link data from these repositories for a given project. We will fill this gap by making our extended R package work not only with Google Sheets, local files, and Amazon RDS, but with the APIs of data aggregators and curators. We will specifically target GBIF for occurrence data [59], NCBI for sequence data [60, 61], OpenTree for phylogenies [62], Map of Life for abundance data [55, 63], TraitBase for traits [64], and GloBI for biotic interactions data [65]. Importantly, all these curation platforms already have dedicated R packages [66–70] that interface with their APIs, which we will leverage for our newly developed package. We will use these existing resources to create high-level functions that facilitate the publishing of data from Google Sheets, local files, and Amazon RDS to these open repositories. Again, this functionality will be made available both as a standalone R package and as a Shiny graphical user interface.

To facilitate our specific data needs for testing our model of eco-evolutionary community assembly, we need to be able to merge data in three potentially overlapping ways, each merger—or "join"—targeting specific shared attributes across disparate data types. We will use a PostgreSQL [71] backend in our R package to allow for these three joins:

1. Spatial joins: we must be able to collect all relevant data, from occurrences, abundance, and environments, for a specific location or region.
2. Phylogenetic joins: we must be able to collect all relevant data for a clade defined by a sample of its constituent members
3. Data availability joins: we must be able to identify clades and regions for which sufficient data are available and then collate those data.

These joins will allow us to, for example, find all taxa that have abundance data at a given location, query their trait values which may have been measured from specimens at different locations, and compile a phylogenetic tree for those taxa and the files containing their raw sequence data.

2.4 Collection of new data

Because we selected five mature DoB projects, many data resources are already available; therefore our new data collection efforts will focus on filling specific gaps (see Table 2) to ensure continuity of data types across projects. Two classes of traits, relevant across the tree of life, will be compiled for all taxa: (1) body size and metabolic rate [72, 73]; and (2) environmental preference. Environmental preferences are not functional traits unto themselves, but they summarize the influence of many such traits on ecology, biogeography, and evolution [74]. We will be quantify environmental preference using species distribution models, novel applications of which have been developed by the members of our team [e.g.; 3, 9, 27].

Project	**Taxon**	**Abund.**	**Occur.**	**Pop. gen.**	**Phylo.**	**Traits**	**Enviro.**
Hawaii	Arthropods	🟦	🟦	🟦	🟥	🟥	🟦
Palau	Algae	🟦	🟦	🟦	🟥	🟦	🟦
Palau	Invertebrates	🟦	🟦	🟦	🟥	🟦	🟦
Palau	Fish	🟦	🟦	🟦	🟦	🟦	🟦
Atlantic Forest	Plants	🟥	🟦	🟥	🟥	🟥	🟦
Atlantic Forest	Butterflies	🟥	🟦	🟥	🟥	🟦	🟦
Atlantic Forest	Herpetofauna	🟥	🟦	🟦	🟥	🟦	🟦
Amazonia	Plants	🟥	🟦	🟦	🟥	🟥	🟦
Amazonia	Birds	🟥	🟦	🟦	🟦	🟦	🟦
US-China	Plants	🟦	🟦	🟦	🟦	🟦	🟦

Table 2: Data resources available, and planned collection of new data. **Blue cells** correspond to sufficient existing data; **red cells** represent new data to be collected by this proposed project.

Gap-filling field work will complement our abundance, phylogenetic, population genetic, and trait sampling. Our primary source for choosing sites will be the experience of the PIs and the knowledge of local collaborators, supplemented with information from satellite images and remote sensing data. Below we detail the status of specific projects, and what new data are needed to build robust phylogenies, document the multiple dimensions of biodiversity, and test our model.

Hawaii. This project uses the dynamic geomorphology of the Hawaiian Islands to examine the evolution of entire communities of arthropods over extended time. A combination of community ecology approaches with genomics of select evolutionary lineages allows us to determine the importance of changing functional roles of taxa within communities as they differentiate and assemble, and the role of that dynamic community in fostering diversification. Currently missing from this effort, and necessary for the model, are complete data on functional traits and phylogenetic relationships. Body size data already exist for $\approx 10^6$ specimens and trophic interactions exist for a much more limited subset, namely herbivorous hemipterans [53]. We will add trophic interactions for all herbivorous insects following the methods in Rominger et al [53], as well as more general trophic position (e.g. predator, parasitoid, etc.) for all arthropods. In addition, collaborator Boettiger is compiling trophic interaction data for the Hawaii DoB based on field observations and sequencing of gut contents. Genetic resources are quickly accumulating for the Hawaii project following novel metabarcoding practices [75]. We will augment these data, which provide sequences from a single marker for all specimens, with multi-locus sequencing for targeted voucher specimens in order to construct a robust tree for ≈ 2000 Hawaiian arthropod species (≈ 100 of which we will add through this grant). This will be the first time such a mega tree has been constructed for Hawaiian arthropods, many of which have never been phylogenetically treated at all.

Palau. Extensive abundance and trait data (including body size, trophic position and habitat preference) have already been collected. This system also has an aggressive COI barcoding initiative covering all invertebrates (459 vouchered species and subspecies in 205 genera). We will again use multi-locus sequencing of targeted vouchers representing ≈ 100 genera to build a comprehensive phylogeny for invertebrates in Palau's unique marine lakes, augmenting ≈ 1200 barcoded samples. We will additionally analyze ancient COI barcode eDNA using myBaits target capture from down-core sediments followed by Illumina HiSeq sequencing to reconstruct community composition and genetic dynamics through the Holocene. A total of 75 m of cores span the Holocene in 6 lakes, and are in storage at the LacCore facility (U. Minnesota).

Atlantic Forest. We will focus on amphibians (> 100 Hylidae tree frogs within Boana, and the Phyllomedusinae), lizards (within the Ecpleopodinae), butterflies (≈ 50 species in the Biblidinae subfamily and the satyroide clade), and plants (≈ 200 species within the Miconiae, 50 within the Bignoniaceae). For these groups, data on geographic occurrence, population genetics, phylogenetics, and functional traits already exist or can be straightforwardly gathered or enhanced with strong involvement of CUNY undergraduate students and the help of collaborators who have independent funding. A backbone multi-locus phylogeny already exists for these groups, but we will add ≈ 100 each to vertebrates and plants to increase the completeness and resolution of those trees. We will generate barcode data for all available and yet unsequenced individuals of the target groups, following standard protocols [75–77]. For the amphibian data, a strong barcoding effort is already in place. The primary information gaps in the Atlantic Forest are curated data quantifying abundance. To address this, we will (1) merge and curate abundance inventories already available for plants (a table of abundance of woody plants is available for > 20 sites), butterflies (7 sites, available through collaborator Freitas), amphibians and reptiles (5 sites, collaborator Rodrigues); (2) complete five replicate surveys to complement the vertebrate abundance data, covering both lowland and montane areas of the forest; and (3) compile data on the history of biocollections for these groups to build a model of occupancy from existing occurrence data. These spatially broad occupancy estimates can be combined with the spatially narrow but detailed abundance data and used in our modeling framework.

Amazonia. We will focus on birds and two plant lineages—Lecythidaceae and Miconieae—for which the most complete data exist. For the plants, we will compile publicly-available abundance data from forest inventory plots [78–80] to augment the detailed occurrence, genetic, and functional trait data that have already been collected for this group. To make use of the many other species inventoried in forest plots, we will also target voucher specimens to generate new sequence data that can be added to existing resources for plant community phylogenetics [81, 82]. For birds, we will perform targeted mist-netting and point surveys to fill geographical and taxonomic gaps in existing abundance data from museum records, estimates from the eBird database, and long-term survey data provided by co-PI Robinson and collaborator Loiselle (who is independently funded).

US-China. The US-China project focuses on levels of diversity across the classic eastern Asia-eastern North American floristic disjunction. Detailed sampling of plant genetic, trait, and phylogenetic diversity at six sites in eastern North America and 5 sites in China will be compiled from existing sources and collected anew by this funded project. Remote sensing data for 3 of the 6 North American sites are currently available. Each site is linked to existing abundance data from NEON (US) and CERN (China) plots, with additional existing abundance data for woody plants provided by the Forest Inventory and Analysis Program (US Forest Service) and partner organizations in China, who are independently funded. We seek to collect sequence data for plant species at four additional NEON sites in North America to improve sampling across the large geographic and climatic gradient and to add remote sensing data for additional NEON sites as the data become publicly available. We will add trait data from publicly available sources for any new species sampled. Occurrence data will be obtained from public databases. Phylogenies for 30 plant genera from across the disjunction will be completed in the funded project.

2.5 Theory development

Hickerson and his PhD student Overcast, who will be funded on the current proposal, have extended the island-mainland neutral model [83] to include coalescence of alleles under neutral mutation, thereby allowing joint predictions of species abundance distributions (SADs) and aggregate population genetic metrics such as genetic diversity (π) across all species (Fig. 2). Rominger, Hickerson, Overcast, and collaborators Harmon and Chase will extend this work further to complete objective **R2** by including the process of speciation [33], the evolution of traits, and ecological interactions (between species, and between species and their environments) based on those traits. This work was initiated by a working group between these researchers at the Santa Fe Institute, and will continue to be independently fostered by working groups among these researchers at the German Centre for Integrative Biodiversity Research.

Trait evolution will follow a continuous Markov process (see Table 1) for continuous traits, and be thresholded [84] for discrete traits. Trait-environment interactions will depend on the distance between an individual's trait value and the optimal trait value for the environment, the greater the distance the less the fitness of that individual. Trait-based species-species interactions will be again based on the distance between traits, in this case the closer the trait values of two individuals, the stronger their Lotka-Voltera [85] competition. This full model (Fig. 2, Table 1) will jointly predict species abundance data, population genetic and phylogenetic data, and trait distributions.

Because available population genetic data across many individuals in a community are limited to extremely reduced genomic regions, and often mined for synonymous SNPs, we will initially model neutral mutations. However, demography will be driven by potentially non-neutral ecological dynamics, and thus population genetic measures, despite being mutationally neutral, may still be far from the predicted drift-mutation balance of a constant population.

Analytical solutions to such models might be available, but only in the asymptotic limit of equilibrium [e.g., 15, 22]. Because we are interested in potentially non-equilibrial histories (hypothesis **H5**) as a driver of biodiversity patterns, we will use efficient methods to simulate model predictions under different scenarios (from neutral to trait-based assembly) and use a combination of approximate Bayesian computation (ABC) and machine learning methods to infer which models with what specific parameter values are best supported by the data. Model fitting is discussed in section 2.6.

Figure 2 summarizes the simulation model which incorporates two key time scales and spatial scales. Across short timescales, slow rates of macroevolution and very large population sizes render diversity at the largest spatial scale, the regional pool, approximately constant. Abundances in the much smaller local community fluctuate due to a birth-death-immigration process [86], modified under the non-neutral models to include trait-based interactions (Table 1). The fluctuating local population determines the temporal trajectory of effective population sizes N_e [87] which, together with mutation rate and colonization event timings, determine the distribution of genetic diversities across species. Across long time scales new species arise and traits evolve.

We will use well known mathematical results for assembly by birth, death, and immigration [35, 86] to implement efficient forward time simulations. These forward-time assemblage histories will constrain the coalescent simulation of genetic data [87, 88]. We will use *msprime* [89] to implement this coalescent approach, which requires the time trajectory of effective population sizes (N_e) for each species and the mutation rate μ as input.

2.6 Model fitting

All processes in Table 1 are expressed by parameters which must be fit to data. Because we are interested in the non-analytically tractable non-equilibrium dynamics of eco-evolutionary community assembly (**H5**) simple likelihood functions are not attainable. Therefore we will use simulation-based approaches to model fitting, including new advances in machine learning and hierarchical ABC. Fitting our model that jointly predicts four data axes that span the three dimensions of biodiversity under many mechanistic processes is precisely the means by which we test our hypotheses about the drivers of diversity (**R3**).

While ABC has been standard for inference of complex models [32], we are also interested in exploring machine learning approaches because they enable the efficient use of high-dimensional data and parameters without specific knowledge of the joint probability distribution. Unlike available ABC methods that can become intractable in high-dimension applications with many summary statistics, supervised machine learning methods perform best in such cases [90, 91].

For our ABC and machine learning approaches we will explore a range of summary statistics that contain information from all four data axes (species abundance, phylogenies, population genetics, and trait distributions) while flexibly allowing some axes to be incomplete. Incomplete data can be accounted for in two ways: (1) summary statistics pertaining to the missing data can be masked, effectively integrating over the possible values those data could take; or (2) a sampling model can be added to allow for the inclusion of the sparse data that do exist. The latter solution is particularly relevant when considering the occurrence data available from many studies. In these cases, we will simulate actual abundance data, but then coarse-grain those abundances to occurrences according to a binomial sampling process with detection probability p. Using such occupancy type models for occurrence data is not unprecedented [92]. However, in our approach we will be able to even more accurately estimate p because we will augment studies of only occurrence data with new surveys of actual abundances. Where occurrence points and surveys overlap in geographic regions, we will be able to precisely fit the value of p, which then will be carried by the model to all regions where occurrence data are available.

Based on preliminary work by Rominger, Renyi entropy (equal to the logarithm of the Hill number) summary statistics of SADs in a machine learning context can be as good or better than pure likelihood approaches in parameter estimation. We will therefore begin with Renyi entropies of SADs [93], distributions of genetic diversities across species, phylogenetic diversity [94], and trait distributions as our summary statistics. We will evaluate whether Renyi entropies are sufficient or if additional statistics improve fit.

2.7 Hypothesis testing

Fitting our model to real data from diverse ecosystems will allow us to understand whether a common set of mechanisms can be seen as universally driving patterns of diversity (**R3**). We will test our five main hypotheses by evaluating how the best-fit model parameters vary across taxa and systems. Figure 3 shows how correlations among estimated parameter values and between parameter values and intrinsic characteristics of regions and taxa can confirm or refute our hypotheses. Critically, not all systems are represented by the same taxa, nor do they cover the same spatial scale. Thus in searching for correlations between model parameters we are prepared to account for random effects of taxon and project.

Evaluating whether our five biogeographic regions have achieved equilibrium (**H5**) requires added analytical nuance. For that, we will take two approaches. In the case of Hawaii and Palau, the age of ecosystems is knowable based on the geology of the system. We will thus explore explicitly how the processes inferred by our model themselves change through time, and how the community as a whole approaches, but potentially does not reach, equilibrium. In the case of Hawaii this is achieved by looking across the chronosequence [53]. In the case of Palau this is achieved by looking at the sediment core from which we will extract taxonomic, functional, and genetic/phylogenetic data.

The proportion of equilibrium achieved [sensu 83], will also be a parameter in our simulation models (i.e. it will determine for how long a given simulation is run). By allowing this parameter to be fit to all our datasets, not just those from Hawaii and Palau, we can understand whether the community of interest has reached equilibrium. If we estimate that it has not, we can then calculate what equilibrium diversity would be under the best fit model. Comparing these observations to the real data will allow us to understand whether the non-equilibrium conditions produce higher or lower diversity across the four data axes that map onto the three dimensions of biodiversity, and whether the specific assembly process is predictive of diversity overshoots or undershoots relative to equilibrium (**H5a-b**).

3 Broader Impacts

Our proposed project contributes to broader society in three quite different ways: (1) training and mentoring in data science, including specifically for underrepresented groups; (2) contributing open source software for science and conservation; and (3) adding to the Tree of Life.

3.1 Training and mentoring

We will work with Data Carpentry and Software Carpentry (“the Carpentries”) to design new workshop content for heterogeneous biodiversity data. The Carpentries are a non-profit organization that teaches hands-on, evidence-based workshops. We will use our newly developed content in a national workshop (for ≈ 30 participants) to be hosted by the Co-PIs at the University of Florida, who have demonstrated success in hosting such workshops. The data science principles taught in the workshop will be practiced by students—who will receive support to attend— as they prepare, curate, and publish the new data that will be generated for our proposed research.

To make this educational material as broadly accessible as possible, the biodiversity data science curriculum will be developed as a massive open online course (MOOC) hosted on the Santa Fe Institute’s online education platform. This content will be distributed under a Creative Commons license. The creation of this course will leverage existing infrastructure and expertise at the Santa Fe Institute.

3.2 Open source software for science and conservation

3.3 Informatic platform

We will provide open-source data management and sharing software that fills the gaps of existing biodiversity data tools, developing a powerful and intuitive software that can help research groups manage, curate, share, and publish their diverse data streams associated with large, multi-dimension, multi-institution projects exemplified by DoB. This tool will also allow multiple user experiences, from programmatic interface through an R package, to a web-based graphical user interface enabled through Shiny.

3.4 Eco-evolutionary synthesis modeling

Our modeling framework, including simulations under different model specifications, model fitting, and model selection, will be made available as an open source R package. With this modeling power at hand, users will be able to gain completely new insights into their study systems. Whether or not communities are assembled primarily by *in situ* evolution or *ex situ* immigration strongly determines their response to anthropogenic pressures [95] and optimal conservation management [96]. We will highlight this importance in an open access applications manuscript which we will seek to publish in a conservation-oriented journal. We will specifically seek publicity for this work such that its message, and our open source software, can be brought to the attention of conservation practitioners.

3.5 Adding to the Tree of Life

Expanding the Tree of Life is a community-wide priority, as we ought to document diversity before human actions erase some of the most remarkable realizations of the evolutionary process. The majority of our data collection is aimed at adding robust tips to the Tree of Life, complementing other ongoing projects [62].

4 Timeline and Project Management

In the table below we detail the timeline for major milestones of our project.

Activity	Yr 1	Yr 2	Yr 3	Yr 4	Yr 5
Organizational working group at Santa Fe Institute (SFI)	■				
Informatics pipeline	■	■			
Carpentries content development		■			
Carpentries workshop		■			
SFI MOOC development			■		
Filling data gaps (abundance, genetics, traits)	■	■			
Model development	■	■	■		
Model validation and fitting with simulated data			■	■	
Model-based inference with real data				■	■
Capstone workshop at SFI					■

4.1 Project Team Coordination

We have assembled a diverse team of researchers from theoreticians to systematists to field ecologists. Rominger will lead coordinating efforts across institutions both remotely and during logistical and synthesis meetings at the Santa Fe Institute. Remote meetings will be held at least monthly to discuss progress on independent tasks and on synthesis. Our larger group will be organized into three core teams: (1) systems data gathering team, (2) informatics team, and (3) theory development team. The **systems team** will compile existing and collect new data from our five biogeographic regions: Rominger and Gillespie (Hawaii), Dawson (Palau), Carnaval and Michelangeli (Atlantic Forest), Owens and Robinson (Amazonia), and Soltis (US-China)), along with their independently funded collaborators. The **informatics team** (Rominger, Guralnick, and Owens with funding for a scientific programmer to be housed at University of Florida) will coordinate with both the systems and theory teams to build an informatics platform that satisfies both groups' needs. The informatics team will also be responsible for developing the biodiversity data science training curriculum in collaboration with the Carpentries and the Santa Fe Institute Office of Education. The **theory team** (Rominger, Hickerson, and collaborators Chase and Harmon) will build our novel modeling framework that jointly predicts species abundances, phylo- and population genetics, and trait distributions. They will also coordinate with the systems team to ensure the model's biological realism.

5 Results from Prior Funding

Carnaval. DEB 1343578: Dimensions US-BIOTA-Sao Paulo: A multidisciplinary framework for biodiversity prediction in the Brazilian Atlantic forest hotspot; 2013–2018; $1,991,480. **Intellectual Merits.** The team is developing models of spatial patterns of biodiversity in the megadiverse Atlantic Forest of Brazil. They generate and integrate: (1) environmental data from novel remote sensing-based datasets and paleoenvironmental archives; (2) locality, phylogenetic, genomic, and trait data from plants and animals; and (3) new methods that incorporate big genomic and environmental data into predictive models. The project has generated 80+ publications to date, including high impact journals [2, 97–104]. **Broader Impacts**. The project contributed to training of five Ph.D. students (2 female, 2 Latin-American), 11 M.Sc. students (10 female, 9 underrepresented minority), 7 undergraduates (all female/URM), and 5 post-docs (2 female, 2 Latin-American).

Dawson. OCE-1243970: Dimensions: Collaborative Research: Do parallel patterns arise from parallel processes; 2013–2017 plus 1-yr no cost extension; $1,369,982**.** **Intellectual Merit**: We investigated taxonomic, genetic, and functional diversity of microbes and macrobiota in widespread yet under-explored marine lake ecosystems, which present independent eco-evolutionary "natural experiments." We are revealing how parallel processes in space and time influence genetic, phylogenetic, and functional diversity. We surveyed over 14000 points across 15 lakes in Palau, sampling 9000 specimens, and barcoding of ≈ 1200 invertebrates and algae. We are finding (1) marine lakes conform with island theory, (2) microbes and macrobiota can share similar biogeographic patterns, (3) abiotic lake dynamics are driven by regional climate and local factors, and (4) life-history and chance strongly influence population structure in invertebrates. Currently available research products: [105–110]. Published datasets: BCO-DMO Project 2238 datasets. **Broader Impacts**: 4 female graduate students (incl. 1 Pacific islander), 2 female lab techs, and a citizen scientist trained on the project. Advised local resource managers on conservation.

Gillespie. DEB 1241253 Dimensions: A community level approach to understanding speciation in Hawaiian lineages. 2013–2019; $1,181,407. **Intellectual Merit**. This project aims to integrate evolutionary and ecological theory to transform understanding of the dynamics underlying community assembly. The synergy between the two approaches is made possible through the use of a habitat chronosequence, provided by the dynamic geomorphology of the young islands of the Hawaiian archipelago. We selected 6 replicates in each of 12 sites and are processing thousands of arthropod specimens while creating an mtDNA barcode library and testing metabarcoding approaches. From these data we are estimating macroecological metrics and conducting food web analysis. For focal lineages, genomic data is providing information on population differentiation over the island chronology. Papers to date: [53, 75, 111–121]. **Broader Impacts**. Research is integrated into education; trained 9 postdocs, 10 graduate students, 44 undergraduates, and one high school student.

Guralnick. DEB-1262610: Collaborative Research: ABI Development: Advancing Map of Life's Impact and Capacity for Sharing, Integrating, and Using Global Spatial Biodiversity Knowledge; $560,861; 2013–2018 plus one year no-cost extension. **Intellectual Merit**: The Map of Life project focuses data integration based from a variety of sources, including biocollections, species attribute data, expert assessment of range boundaries, surveys, and inventories. This work strongly enhances methods to publish, discover, openly link, and create new knowledge, in order to further assessment of global biodiversity. Papers to date: [63, 122–127]. **Broader Impacts**. The support has yielded web platforms that provide data and knowledge and provided support for 2 Ph.D. students.

Hickerson. DEB 1253710: Career: Dynamic models of isolation and admixture for community-scale population genomic inference; 2013–2017; $667,074 plus 1-yr no cost extension. **Intellectual Merits**. Hickerson and colleagues developed inference methods to infer the aggregate spatial history of species assemblages given genome-wide data from multiple taxa. The grant has supported the publication of 22 peer reviewed articles [42, 128–148]. **Broader Impacts**. The grant has supported two PhD students, two postdocs, one undergraduate researcher who has gone on to pursue a PhD in population genetics and an extensive educational outreach component.

Owens. DBI 1523732: Postdoctoral Fellowship in Biology: Out of the Tropics and Out of the Drawer: Integrative Analysis of the Tropical Diversity Gradient from Museum Collections of New World Swallowtail Butterflies; 2015–2017; $138,000. **Intellectual Merit.** I imaged and processed over 1300 butterfly specimens housed at four institutions through a digitization pipeline I designed for this project. Results, leading to one publication to date [9], suggest abiotic niche conservatism and temperate-to-tropical dispersal are important, under-appreciated sources of diversity in the tropics. **Broader impacts.** I engaged worldwide volunteers through the Notes From Nature platform to transcribe specimen label data. I have also participated in several public outreach events at the Florida Museum of Natural history to demonstrate how data derived from specimens can inform our understanding of macroecology.

Soltis. EF-1115210/DBI-1547229. Digitization HUB: A Collections Digitization Framework for the 21st Century/iDigBio Phase 2; $12,661,986/$15,486,747; 2011–2016 and 2016–2021. **Intellectual Merit**. iDigBio is the national coordinating center for digitization of biodiversity collections, the goal of which is to make data for millions of biological specimens available in electronic format. iDigBio currently serves over 108 million specimen records and over 22 million media records, providing information on taxonomy, geographic location, collector, date of collection, etc., and notes on phenology, habitat, molecular resources, and other features, as well as images and vocalizations. These diverse data promote integrative biodiversity research, contain untapped trait data, and provide an immense baseline for assessing impacts of climate change, invasive species, and other environmental issues. The project has produced over 20 papers to date, including in *Nature, BioScience, Proceedings of the IEEE International Conference on e-Science, New Phytologist, Philosophical Transactions of the Royal Society B, American Journal of Botany, Cladistics*, and *ZooKeys*, among others **Broader Impacts.** More than 100 workshops and hackathons have fostered training in digitization workflows, use of digitized data in research, applications in citizen science, education, and outreach. Soltis has trained 2 post-docs, 6 graduate students, and 5 undergraduates.