

Combining Gradients of Space and Time to Understand Biodiversity Dynamics in the Hawaiian Islands

1 Background

6 Biological diversity is nearing or already past a global tipping point [1]. Beyond this phase transition,
7 the processes regulating biodiversity will change, and the dynamics of their resultant biological systems,
8 from clades to ecosystems, will become non-steady state [1]. Despite the pressing need , our level
9 of understanding of biodiversity dynamics remains rudimentary . We must be able to address how
10 biodiversity has been shaped in the past, what are the expectations as we move into the future, and
11 how will associated ecosystems respond to global change. Phase transitions operate across spatial
12 scales and so we must be able to tackle these questions from plots to biomes in order to detect
13 and understand non-steady state dynamics. Advances in our understanding of specific ecosystem
14 components are idiosyncratic. While remote sensing and distributed biogeochemical monitoring [2, 3]
15 are rapidly advancing ecosystem modeling, similar large scale study of organismal processes, from data
16 generation to theory development, and from genetics to populations and communities, lags behind,
17 especially for “dark taxa” such as arthropods and microbes.

18 Biodiversity results from both evolutionary and historical processes operating at larger spatiotemporal
19 scales and ecological processes operating at smaller scales [4]. Feedbacks between processes along
20 this evolutionary-ecological continuum drive non-steady state biodiversity dynamics [5–7]. The conse-
21 quences of non-steady state dynamics, based on state shifts in the geologic past, will persist for millions
22 of years [8]. Yet we lack approaches that synthesize across scales of space and evolutionary time to
23 understand the consequences of this eco-evolutionary feedback process. The propensity for systems to
24 transition into non-steady states cannot be assessed given current means of synthesizing ideas from
25 ecology with those from evolution. Additionally, a lack of cross-scale biodiversity data (from plots to
26 landscapes and genes to species) combined with a lack of theoretical framework, limit this synthesis.

27 1.1 Theory provides a lens on non-steady state processes

28 Recent theoretical developments have brought mechanistically simplified theory to the forefront of
29 ecological research [9–11]. These simple theories have been critical because they provide robust null
30 models against which to compare real biodiversity patterns in order to rigorously test the importance
31 of specific mechanisms in shaping biodiversity. The maximum entropy theory of ecology [METE; 11]
32 provides one of the most useful null predictive frameworks because it produces many falsifiable patterns
33 (the species abundance distribution, metabolic rate distribution, species area relationship and network
34 structure) and is grounded in the principles of statistical mechanics [11, 12]. METE draws from the
35 probabilistic properties of large, randomly assembled systems [11] and thus its predictions constitute
36 a community in statistical steady state. Statistical steady state means specifically that a system is
37 governed by only a few simple state variables, which constitute a state space, and that no additionally
38 processes limit the systemâŽs ability to freely explore this state space. This precise definition is made
39 more clear in Box 1. Statistical steady state connects to some notions from the literature on ecological
40 equilibrium, specifically the condition of stationarity and ergodicity [13], but is in no way tied [11] to
41 ideas relating to equilibrium as a hypothesized state that ecosystems may attain or be driven away
42 from [e.g. not: 14–16].

43 Deviations from METE allow us to identify ecological systems out of statistical steady state [7, 11].
44 Drivers of such non-equilibrium include rapid assembly following disturbance [11] and constraints
45 imposed by evolutionary history and non-neutral adaptive differences between species that violate
46 the statistical assumptions underlying the principle of maximum information entropy [7]. In order

47 to harness these promising properties of METE as a non-equilibrium diagnostic tool more testing is
 48 needed to understand how exactly the ecological and evolutionary setting of a community predicts its
 49 deviation from METE.

50 We propose to use islands of the Hawaiian archipelago to better understand
 51 how and why ecosystems depart from
 52 steady state, the consequences of these
 53 departures on ecosystem function and
 54 biodiversity dynamics, including nutrient
 55 cycling and invasibility, and finally,
 56 how maximum entropy theory can be
 57 used as a tool to identify these depart-
 58 tures. Remote island archipelagos pro-
 59 vide an opportunity to integrate eco-
 60 logical and evolutionary processes, ad-
 61 vancing our understanding of the regu-
 62 lation of biodiversity through the lens
 63 of theory. This is particularly true
 64 when the component islands are ar-
 65 ranged chronologically, as is found in
 66 “hotspot” islands that form a geologi-
 67 cal age gradient representing snapshots
 68 of community assembly through evo-
 69 lutionary time. Such islands provide
 70 simple and discrete systems, of known
 71 age and varying area, allowing them to
 72 serve as excellent “natural laboratories”
 73 for ecological and evolutionary study in
 74 a regional context [20–22]. Our team
 75 has a strong foundation of research ex-
 76 pertise and experience across the is-
 77 lands on microbes (Brodie), arthropods
 78 (Rominger, Gillespie, Gruner and Kre-
 79 henwinkel), plants (Chase), ecosystems
 80 (Giardina) and the  (Rominger and
 81 Chase).

83 We will characterize the ecological
 84 communities, including their abun-
 85 dance, diversity and network structure,
 86 associated with three critical stages in
 87 nutrient cycling: 1) Living plants, the
 88 arthropods they support and the microbes supported by both; 2) Plant and animal detritus and its
 89 associated arthropod and microbial communities; and 3) Soil communities of arthropods and microbes.
 90 In each of these ecosystem domains we will use the maximum entropy theory of ecology to characterize
 91 departure from statistical steady state. In order to understand the mechanistic causes of these depa-
 92 tures we will also evaluate how deviations from METE can be predicted by the ecology and evolution of
 93 the organisms comprising each community, testing the hypotheses outlined below. We will enable this
 94 line of research by deliberately sampling plants, arthropods and microbes across multiple spatial scales,

Box 1: Statistical Steady State

Whether or not biodiversity dynamics are governed by stable equilibria remains an unsolved question in ecology and evolution [17, 18]. A statistical steady state exists in an ecological community if changes in biodiversity occur slowly and in sync with environmental changes [11]. The existence (or non-existence) of such steady states has wide ranging implications. For example, whether conservation should focus on conventional preservationist paradigms or adaptive management [19] depends on whether biodiversity is largely in statistical steady state or not. Whether biodiversity rapidly and consistently tends towards a steady state determines how species and the communities they form will respond to global environmental change [1].

We posit that two primary classes of non-steady state exist and can be better understood by combining comparative population and phylogenetic insights across multiple species and ecological theory. The first class of non-steady state occurs when a biological assemblage is undergoing succession following disturbance or formation of new habitat; in this case populations of most species in the community and species composition itself will be in flux due to the stochasticity of immigration and small population sizes. In such a situation the assemblage may be expected to eventually converge on a steady state. Recovery from disturbance, range expansion following climate change and primary succession are all potential examples of such non-steady state. The second case occurs when novel mechanisms actively drive an assemblage away from steady state; such mechanisms could include escalatory species interactions or rapid diversification and adaptation in the face of newfound selective pressures. In both cases idealized ecological theory should fail to predict the static biodiversity patterns of the system and departures from population genetic theory should indicate what demographic dynamics are associated with the failure of ecological theory.

95 and across gradients of environment (precipitation and elevation as a surrogate for temperature) and
96 substrate age (as a surrogate for both biogeochemical change and evolutionary development). We will
97 also make use of long term fertilization experiments [23] to evaluate the orthogonal roles of evolution-
98 ary history versus biogeochemical processes in driving biodiversity patterns. Using plants, arthropods
99 and microbes as discrete test cases, representing a breadth of life history strategies across the tree of
100 life, we will test hypotheses (outlined in Box 2) about deviations from statistical steady state based
101 on how organisms persist, adapt and speciate in their environments. In order to understand how com-
102 munities are likely to change in response to non-analog, anthropogenically-driven climate regimes and
103 across spatial scales we will build spatially explicit models that link the mechanistic drivers (e.g. rapid
104 community or population change, and evolutionary novelty) of deviation from statistical steady state
105 to remotely sensed data and detailed ecosystem characterizations taken at the NEON site in Hawaii,
106 and our complementary sampling locations. Our project will contribute theoretical constructs for use
107 across NEON sites and bioinformatic tools to advance the rate and dimensionality of biodiversity data
108 gathered at these sites.

109 2 Proposed Research

110 2.1 Research objectives and hypotheses

111 We will use maximum entropy theory to identify deviation from statistical steady state across environ-
112 mental and evolutionary gradients, and long-term experiments. We will place these deviations in the
113 context of ecological and evolutionary information to understand the mechanistic causes for deviations
114 from statistical steady state and its implications for invasion potential. To forecast these mechanisms
115 and implications into future, non-analog environments we will model the ecological and evolutionary
116 drivers of deviations using remotely sensed environmental variables and detailed field measurements
117 from the NEON site and our complementary sampling sites. These models will be spatially explicit
118 and use the framework of Bayesian hierarchical modeling to incorporate diverse data types. To permit
119 theory testing and modeling across large scales we will develop a novel sequencing and bioinformatics
120 approach to generate massive, multidimensional (i.e. taxonomic and genetic) biodiversity data. We
121 will use this combined approach of novel theory testing and novel data generation to test hypotheses
122 outlined below relating departures from statistical steady state to feedbacks between ecological and
123 evolutionary processes.

124 2.2 Hypotheses

- 125 • Departures from statistical steady state

126 H1 Deviations from METE are largely predicted by age along the chronosequence. These de-
127 viations along the chronosequence will be driven primarily by two processes related to evo-
128 lutionary assembly of biotas: (H2a) primary succession (both by long distance dispersal
129 and speciation) of newly formed habitats; and (H2b) adaptive evolution leading to unique
130 constraints on assembly not consistent with statistical steady state

131 H1a will be more relevant for generalist taxa, especially those that are dispersal limited, on
132 young substrates.

- 133 • We predict greatest deviations for communities dominated by generalist taxa on
134 young substrates
- 135 • We predict a positive correlation between deviations from METE and measures of
136 spatial turnover, both taxonomic and genetic.
- 137 • We predict a negative correlation between the breadth of reconstructed abiotic
138 niches and deviation from METE

- 139 H1b will be most relevant for specialist taxa once they have established intricate evolutionary
 140 relationships with their coexisting species and environments.
- 141 · We predict greatest deviations from METE for communities dominated by specialist
 142 taxa on old substrates
- 143 · We predict a positive correlation between network specialization and deviation from
 144 METE
- 145 · We predict a negative correlation between phylogenetic diversity and deviation from
 146 METE
- 147 H1c Because niche specialization and dispersal limitation both likely result in strong spatial
 148 structuring of communities, measures of spatial turnover and deviations from METE
 149 should be correlated across all ages along the chronosequence
- 150 H1d Because rapid population expansion, population contraction, limited dispersal and local
 151 adaptation all lead to low allelic diversity within populations we predict genetic diversity
 152 to be negatively correlated with deviation from METE
- 153 H2 Deviations from METE are not predicted by environmental variables after accounting for
 154 ecosystem age. This includes the prediction that in long term fertilization experiments,
 155 fertilized communities will conform to the same patterns as their unfertilized control com-
 156 munities of the same age regardless of underlying nutrient availability
- 157 H3 However, with rapidly changing climates we do expect environmental predictors of deviations
 158 from statistical steady state. Specifically, with the creation of novel environments and loss
 159 of existing environments due to changing climate we expect rapid population changes and
 160 exacerbated constraints on movement due to unique evolutionary adaptations to previously
 161 stable environments. Thus we predict novel climatic conditions to drive future deviations
 162 from METE
- 163 H4 We predict that in disturbed systems the only what for statistical steady state to be achieved
 164 is through rapid assembly of novel ecosystems (i.e. communities dominated by highly vagile
 165 invasive taxa). Thus deviations from statistical steady state are expected to promote inva-
 166 sion, while invasion itself will tend to return systems to statistical steady state.
- 167 • Evolution of niches and networks
- 168 H5 We predict that niches will become more constrained across evolutionary time
- 169 H5a Reconstructed niches will be smaller for taxa  endemic to older islands
- 170 H5b Spatial turnover will be stronger across gradients on older islands
- 171 H5c We predict networks will become more specialized across evolutionary time

172 2.3 Significance and Rationale

173 Understanding how environmental change will alter the feedback between ecology and evolution and
 174 drive biodiversity out of statistical steady state is at the core of our proposal. Using METE to capture
 175 statistical steady state and understand deviations from it promises to be a powerful diagnostic tool in
 176 evaluating ecosystems nearing tipping points. Hawaii is an ideal study system to realize this potential
 177 due to its varying chronology (allowing tests of theory in communities of different stages of evolutionary
 178 development) and due to its replicated environmental gradients across this chronology (see Fig. 3. The
 179 NEON site at Puu Maka'ala Natural Area Reserve on Hawaii Island will provide the core measures
 180 needed to quantify the abiotic environment. We will replicate these measurements across gradients of
 181 elevation and precipitation, using ground-truthed remotely sensed measurements to provide both fine
 182 grain and broad-scale environmental data products.

183 The same ability to generate massive amount of environmental data via remote sensing does not exist
 184 for organismal ecology and evolution. As part of our Dimensions in Biodiversity grant, PIs Rominger
 185 and Krehenwinkel are developing laboratory and bioinformatic methods to obtain sequence data, and

¹⁸⁶ estimates of abundance and biomass for thousands to millions of arthropods collected via ecological
¹⁸⁷ sampling. As part of the current proposal this promising new approach will be developed into an open
¹⁸⁸ source lab protocol and software package that can be distributed across all NEON sites.

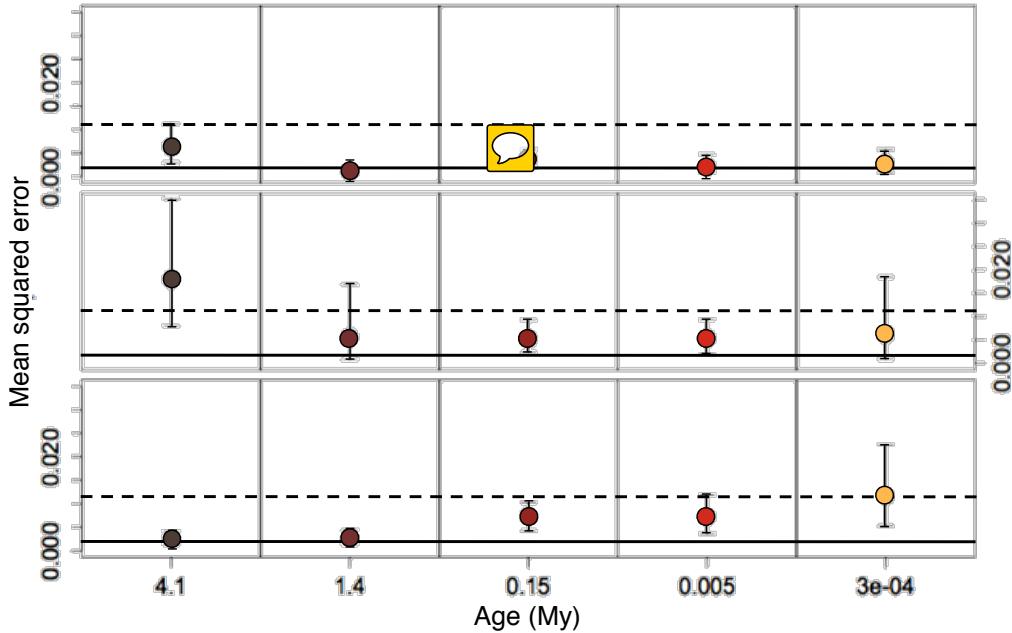


Figure 1: Deviations from METE (measured as mean squared error) across the Hawaiian chronosequence (colors correspond to ages in Fig. 3) for three different arthropod guilds [data from 24]. Note that predatory arthropods have high dispersal ability and low population genetic structure compared to detritivorous arthropods [7], which explains greater deviations from statistical steady state in detritivores, especially in younger aged ecosystems. The poor fit of herbivores in older ecosystems is likely due to their unique evolution with host plants, corroborated in our study of network structure across the chronosequence [7].

¹⁸⁹ Our use of METE as a diagnostic tool has been corroborated in the Hawaiian system with previous
¹⁹⁰ and current work. PI Rominger, with co-PIs Gillespie and Gruner as collaborators and co-authors, has
¹⁹¹ shown that deviations from METE show consistent patterns across the chronosequence for different
¹⁹² arthropod guilds with different life history characteristics (Fig. 1). Additionally, these patterns can be
¹⁹³ predicted by the amount of spatial turnover species composition between sites, and by the proportion
¹⁹⁴ of the community dominated by invasive species (Fig. 2). This work confirms our hypotheses that
¹⁹⁵ deviations from statistical steady state can be predicted by limited dispersal and niche partitioning
¹⁹⁶ (leading to increased spatial turnover) and is related to invisibility of ecosystems. Our proposed work
¹⁹⁷ will extend this approach by explicitly testing more nuanced hypotheses about the role of evolutionary
¹⁹⁸ processes in driving these non-steady state observations and extending these predictions across space
¹⁹⁹ and time with hierachal models.

²⁰⁰ 2.4 Methods

²⁰¹ 2.5 Integration with NEON and sampling design across environmental and age gradients

²⁰² *NEON site.* The goal of NEON is to provide ecological data at multiple spatial and temporal scales.
²⁰³ Our plan is anchored with the Pu'u Maka'ala Natural Area Reserve on the Mauna Loa volcano on
²⁰⁴ the Big Island of Hawaii (19.553° , -155.317° ; Fig. 3), a Core Terrestrial site with the launch date

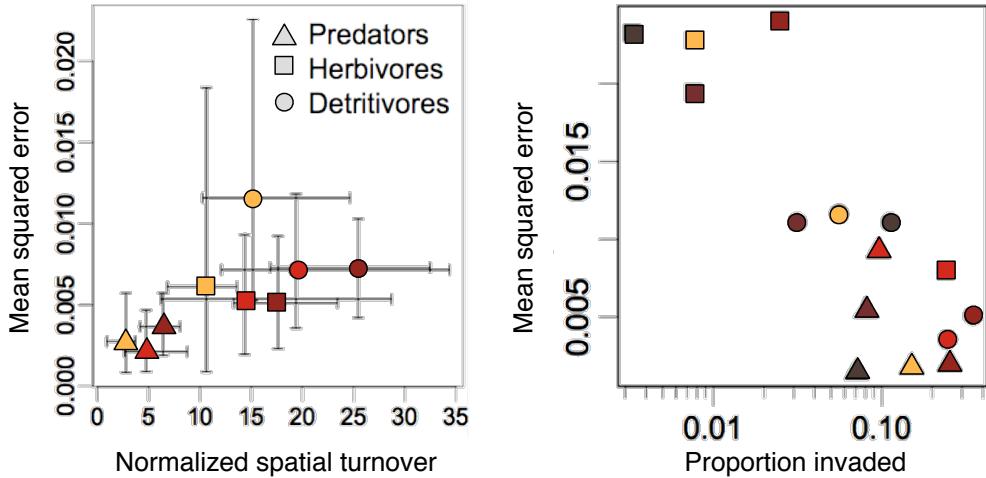


Figure 2: Across arthropod guilds and substrate ages (colors correspond to age in Fig. 3) deviations from METE (measured as mean squared error of METE predictions) are predicted by spatial turnover or species composition (left panel). More heavily invaded systems conform best to METE (right panel) suggesting that invasion acts to bring non steady state systems back into steady state by an influx of highly vagile, generalist invasive species.

205 planned for 2017. The site represents montane wet forest with mostly native vegetation dominated
 206 by the endemic tree, *Metrosideros polymorpha* (Myrtaceae). However, up to 95% of the worldâŽs
 207 terrestrial climates are represented in the greater region of the Hawaiian archipelago [25], and a single
 208 site will fail to characterize this tremendous diversity in climate, habitats and species composition. By
 209 replicating core NEON protocols at carefully selected sites with orthogonal variation in temperature
 210 and precipitation, along a geological chronosequence representing evolutionary time, the Hawaiian
 211 macrosystem will yield the precision of NEON measurements to test ecological theory and to predict
 212 consequences of future changes in climate. We aim to combine data to be collected with data from
 213 sites across the Hawaiian Islands, in order to understand regional-scale ecological processes and how
 214 these respond to change over space and time.

215 *Complementary sites.* We will collect data in an explicit, nested design that allows integration with
 216 the NEON-generated data, while using data from the entire terrestrial region of the Hawaiian Islands
 217 to provide information on processes of several groups of organisms across multiple scales. Data will be
 218 gathered across elevation and precipitation gradients from evolutionarily old, middle aged and young
 219 islands (Kaua'i: 4–5 my; Maui: 1–1.5 my; and Hawai'i: 0.001–0.5 my). On each island we will establish
 220 6 sites (1 ha in size): 3 along a windward (i.e. high precipitation) elevation gradient and 3 along a
 221 leeward (i.e. low precipitation) elevation gradient (Fig. 3). Windward sites will be constrained to be
 222 within 4000–5000 mm annual precipitation, while leeward sites will be constrained to be within 1500–
 223 2500 mm annual precipitation. We will consider an elevation gradient from 900 – 2500 m elevation. On
 224 Hawai'i Island we will use the area adjacent to the Pu'u Maka'ala NEON site as one of these 6 sites.
 225 Each site will consist of 3 replicate plots to insure thorough coverage of local variation. The sampling
 226 locations and design are given in Figure 3.

227 *Sampling approach and collection of organismal data.* We will select sites in clearly defined ohia/koa
 228 montane, wet and mesic forest communities. The rationale here is that (i) Ohia (*Metrosideros poly-*
 229 *morpha*) is the dominant canopy tree in these forests, forming a nearly continuous layer, with patches
 230 of sub-dominant koa (*Acacia koa*) and numerous associated understory trees, shrubs, herbs, and ferns.

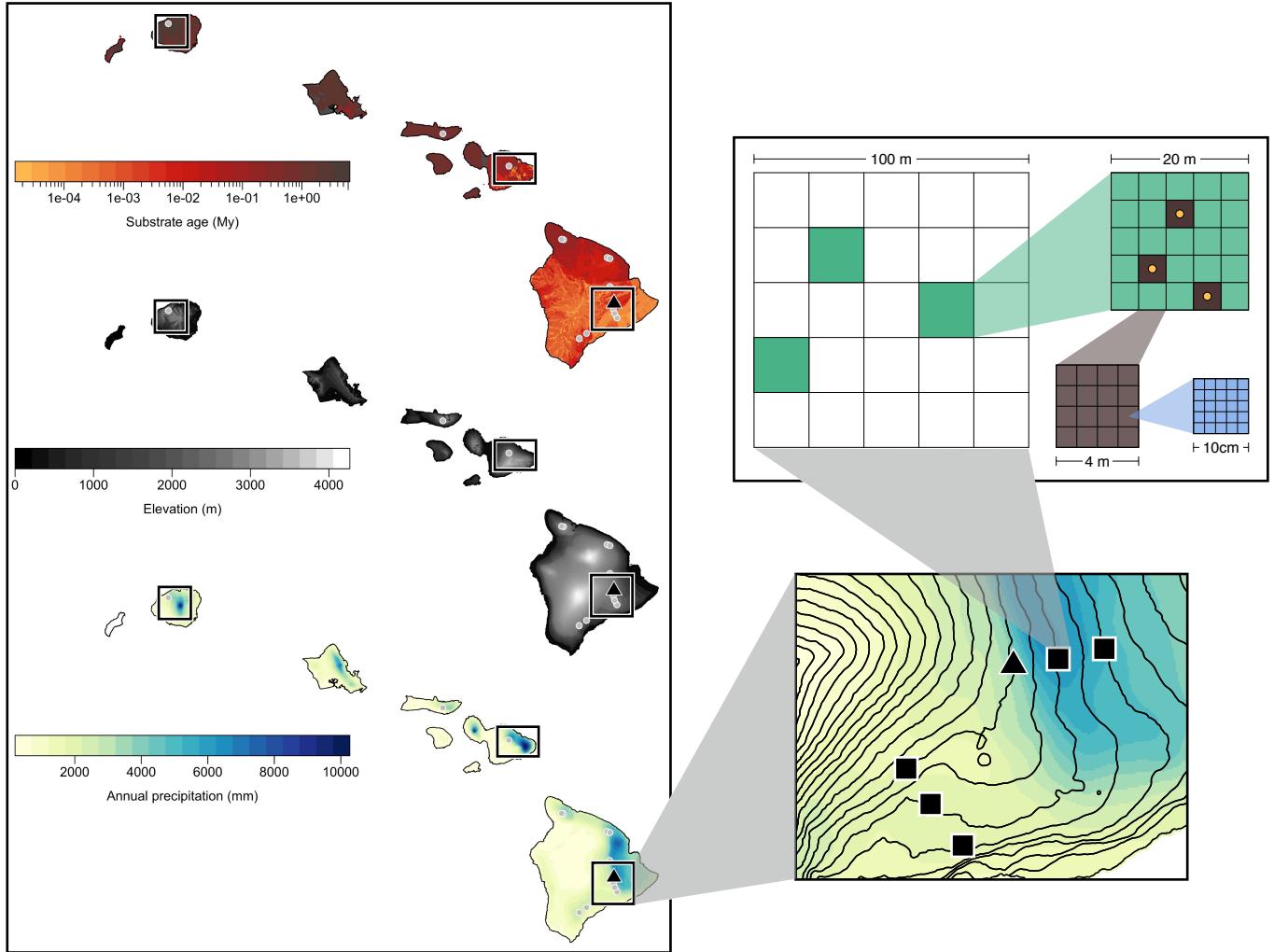


Figure 3: Map of Hawai'i showing chronological age, elevation, and precipitation. Gray dots represent sampling locations with existing data from our Dimensions in Biodiversity project. Triangle corresponds to the NEON site at Pu'u Maka'ala. Regions delineated with rectangles represent proposed areas where sampling sites will be established (6 per island). Black squares in the detail map represent potential sampling sites complementing the NEON site. Sites will be organized in a nested fashion. Green plots represent our stratified sampling plots within a site, brown quadrats correspond to litter and soil sampling quadrats, and the blue 10 cm x 10 cm sub quadrant corresponds to our microbial samples. Orange dots represent locations of temperature and humidity data loggers, in addition to litterfall collection locations.

This forest type (and the presence of *Metrosideros* in particular) has been used as an important landscape feature in our ongoing work through the Hawaii Dimensions of Biodiversity, as it has for a generation of studies on long-term ecosystem development. This constrains sampling to vegetation and soils of similar physiognomy and evolutionary history, while allowing major climatic state factors to vary. (ii) The proposed NEON site is characterized by this forest type. Finally, (iii) *Metrosideros* growth rate, growth form and chemical composition (all detectable by various satellite and airborne spectroscopic techniques [2, 26, 27] reflects the coupled but nonlinear effects of ecosystem age and fertility, which in turn affects the community of organisms in a given forest stand [24, 28]. Differences in plant traits can affect the structure of an entire food web through a series of direct and indirect

240 effects [29, 30].

241 Figure 3 details the proposed layout of our sampling plots. Within each 1-ha site, we will establish three
242 20-m by 20-m plots to be selected as representative of forest height mean, maximum, heterogeneity
243 found in that 1-ha site. Within each 20mx20m plot, we will establish our replicate plots. Each plot will
244 be further gridded into 4 m quadrats (100 in total). Within each quadrant we will record all tree species
245 ≥ 1 cm at breast height. Within three randomly selected quadrats we will also sample all herbaceous
246 species. We will sample all arthropods within each quadrant using timed beating (24 seconds per
247 quadrant). Within the same three randomly selected quadrats we will also extract arthropods using
248 Berlese funnels from litter and soil samples, gridded to 1 m² cells (in keeping with the ground beetles
249 collected at the NEON site). Arthropods will be collected into RNAlater to preserve their DNA and
250 RNA as well as the DNA and RNA of their associated microbes and gut contents. While NEON
251 protocols focus on ground beetles (Carabidae), mosquitoes (Diptera: Culicidae), and ticks (order
252 Ixodida), our study will include all arthropods because ground beetles constitute an eclectic group of
253 lineages, most often arboreal and unevenly distributed across the main islands [31], and there are no
254 native mosquitoes or ticks [32].

255 Microbial richness and abundance will also be sampled in a gridded design. Within three randomly
256 selected quadrats in each plot we will take a soil sample 100 cm in surface area (10 cm by 10 cm) and
257 10 cm deep. In the lab this will be divided into a regular 2 cm grid and each will be sequenced.

258 In all systems, microbial diversity will focus primarily on the Domain Bacteria due to its phylogenetic
259 breadth, and metabolic and respiratory plasticity. Bacterial diversity will be estimated using molecular
260 tools to sequence 16S rRNA gene biomarkers in multiplex using a barcoding approach. DNA extrac-
261 tion and 16S rRNA gene amplification and Illumina sequencing will be carried out according to Earth
262 Microbiome Project standards (<http://www.earthmicrobiome.org/emp-standard-protocols>). An-
263 cillary and meta data collection standards will follow the NEON the soil microbial data collection
264 and metadata tracking worksheet (<http://goo.gl/nE9zPk>). Microbial 16S rRNA gene data will be
265 analyzed according to Shi et al. [33]. Richness will be estimated using both taxonomic (OTUs) and
266 phylogenetic (Faith's phylogenetic distance) metrics. Absolute bacterial abundances will be deter-
267 mined using quantitative PCR as described in [33] while relative abundances of bacterial taxa will
268 be determined based on the fractions of sequence reads assigned to each taxon using adjustments for
269 rRNA gene copy number [34]. In order to relate bacterial taxa to metabolic rate we will use observed
270 relationships between rRNA copy number, genome size and metabolic rate [35].

271 Environmental and biogeochemical data

- 272 1. Plot-level measurements: In each microbial sampling quadrant we will deploy data loggers to
273 record air temperature and moisture content. We will similarly deploy data loggers to record soil
274 temperature and moisture. We will also measure soil physical characteristics, pH, total carbon,
275 nitrogen, phosphorous and sulfur. We will measure monthly litterfall using litter traps as a
276 surrogate for nutrient cycling [36, 37] in addition to litter chemistry (pH, total carbon, nitrogen,
277 phosphorous and sulfur) [3].
- 278 2. Remote Sensing and Measurements of Gases: The NEON site will track fluxes of gases, such as
279 carbon dioxide (CO₂) and water vapor, and collects data about physical and chemical climate
280 conditions, such as temperature, barometric pressure and visible light or Photosynthetically Ac-
281 tive Radiation (PAR). Sensors on the NEON tower systems track fluxes of gases (CO₂, water
282 vapor) and collects data about physical and chemical climate conditions, such as temperature,
283 humidity, wind, and the amount of gas that is exchanged between the atmosphere and the ecosys-
284 tem. Towers extend past the top of the vegetation canopy at each site to allow sensors mounted
285 at the top and along the tower to capture the full profile of atmospheric conditions from the top of

286 the vegetation canopy to the ground. Automated tower sensors collect data continuously to cap-
 287 ture patterns and cycles across various time periods, ranging from seconds to years. Categories of
 288 measurements are physical climate (aerosols, precipitation, radiation, and temperature, pressure
 289 and wind; chemical climate (wet deposition, chemistry, isotopes and scalar concentrations); net
 290 ecosystem exchange: carbon dioxide (CO₂) flux, soil CO flux, water vapor and latent heat flux,
 291 sensible heat, total reactive nitrogen (NO₂) and ozone (O₃).

- 292 3. Airborne Remote Sensing: We will make use of both existing and planned airborne remote
 293 sensing data which can provide information on vegetation composition and land cover and will
 294 be used in particular to examine the complex mosaic of forest structure and composition. The
 295 NEON Airborne Observation Platform (AOP) measures vegetation biochemical and biophysical
 296 properties with spectroscopy, vegetation structure and biomass with LiDAR, and produces high
 297 resolution imagery that can be subject to analyses of land use and relative cover [3].

298 2.6 Modeling evolutionary and environmental drivers of assembly

299 *Maximum entropy theory of ecology across gradients of environment and age* To test our hypothe-
 300 ses relating age, environment and organism/community traits to deviations from METE we will use
 301 the R package `meteR` [developed by Rominger 38] to evaluate the goodness of fit of METE for soil
 302 microbes, arthropods and plants at our sampling sites across gradients of precipitation, elevation and
 303 age. Goodness of fit will be measured as the normalized log likelihood squared [described in 38]. Us-
 304 ing generalized linear models we will evaluate how the goodness of fit varies between major groups
 305 (microbes, arthropods and plants) and as a function of the underlying age and environment of each
 306 site.

307 To further explore the relative importance of age as a proxy for evolution versus biogeochemical en-
 308 vironment we will use VitousekâŽšs long term fertilization experiments to test whether alleviating
 309 nutrient limitations in old and young plots changes the the way in which arthropod and microbial
 310 communities deviate or conform to METE.

311 *Modeling niches, networks and community phylogenetics across space and evolutionary time* We will
 312 develop a Bayesian hierarchical modeling framework to understand how these drivers of deviations
 313 from statistical steady state response to local and regional environments. In all models we will incor-
 314 porate explanatory environmental variables as spatial averages with an exponentially decaying distance
 315 weighted function. Each variable will receive maximum weight at the point location of the specimen
 316 and exponentially less weight as distance from the point location increases. The exponential rate of
 317 decay will be fit as a free parameter in our Bayesian hierarchical model.

318 We will use island age as an explanatory variable interacting with environment to evaluate how the
 319 niche occupancy and network position of each species changes with evolutionary age. Because we will
 320 have phylogenetic data from metabarcoding for all species we will evaluate patterns of niche occupancy
 321 and network position in a phylogenetic framework, testing hypotheses of whether closely related taxa
 322 overlap or diverge in niche occupancy, and whether more recently diverged species tend to be generalists
 323 or specialists.

324 To test whether the niche spaces of taxa change across the chronosequence we will build probabilistic
 325 niche models for all species of plants and arthropods with sufficient data ($n \geq 15$ points per island). We
 326 will use data sources from our gradient plots, plots from our Hawaii Dimensions in Biodiversity project,
 327 digitized museum specimens and species occurrence data made available reporting by the Hawaii Di-
 328 vision of Land and Natural Resources. Because the nature of these data is variable (abundance and
 329 presence-only) we will use Bayesian hierarchical models to combine them into one analysis [39]. We
 330 jointly model the niches of all species in this hierarchical approach.

331 To test how networks evolve across the chronosequence we will quantify network structure using four
 332 complementary approaches: 1) deviation from the maximum entropy predictions; 2) classic ecological
 333 network metrics of nestedness and modularity; 3) network dissimilarity; and 4) network specialization.
 334 We will again take a phylogenetic approach to evaluate how changes in network position of taxa and
 335 changes in overall structure of networks relates to the phylogenetic distance between component taxa.

336 Phylogenetic diversity will itself be modeled as a response to age and environment using the same
 337 Bayesian hierarchical approach as niches and networks.

338 2.7 Projecting deviations from statistical steady state into the future

339 Once we understand the connections between network structure, niche occupancy, population size
 340 change, evolutionary diversification and deviation from statistical steady state, we can use our models
 341 for niches, networks and phylogenetic diversity to project these drivers into the future and predict
 342 where (at a regional scale) statistical steady state will be violated. Using our understanding of how
 343 statistical steady state contributes to invasibility of a community we will also be able to model invasion
 344 risk across scales and into future climate scenarios.

345 2.8 Quantifying evolutionary and macroecological patterns using metabarcoding



346 Next generation sequencing technology has ushered in a revolution in evolutionary biology and ecology.
 347 This revolution has not passed by taxonomy and spurred various new studies in the field of molecular
 348 barcoding. The current leap in sequencing throughput allows to routinely perform barcoding studies
 349 on bulk samples and analyzing whole ecosystems [40–42]. The large scale recovery of species richness,
 350 food web structure, cryptic species, identification of juveniles and hidden diversity, e.g. internal par-
 351 asitoids, promise unprecedented new insights into ecosystem function and assembly [40–43]. While
 352 species richness can be routinely identified by sequencing bulk samples, estimating species abundance
 353 remains challenging [44] and severely limits the application of metabarcoding to many studies. We are
 354 developing wet lab and bioinformatic methods to overcome this issue and revolutionize the generation
 355 of ecological and genetic data. Our pipeline consists of three steps (Fig. 4):

- 356 1. Extraction and sequencing of pooled community samples
- 357 2. Matching the resulting sequences to a reference phylogeny for identification
- 358 3. Using Bayesian hierarchical models to reconstruct unbiased estimates of abundance

359 Step (1) will be released as an open source lab protocol and steps (2-3) will be developed into an open
 360 source R package that allows users to implement these methods in their study systems. We propose
 361 that our open source pipeline can be implemented across NEON sites to generate both taxonomic and
 362 phylogenetic data for focal taxa.

363 Preliminary results from controlled experiments show there is a strong correlation between amount of
 364 DNA and total number of reads; however, this relationship is variable across taxa. A Bayesian model is
 365 able to capture this variability across taxa and thus indicates the success of more general applications
 366 of the modeling approach to field collections.

367 *(1) Extraction and sequencing of pooled community samples.* We will generate sequence information
 368 for mixed arthropod community samples, collected across precipitation gradients on the Hawaiian
 369 Archipelago. The samples will be roughly pre-sorted taxonomically and grouped into different body
 370 size classes to minimize the confounding factors of abundance and body size in determining amount of
 371 DNA per taxon. We will use amplicon sequencing of the COI coding region [42] which has shown
 372 the greatest reliability in preliminary trials.

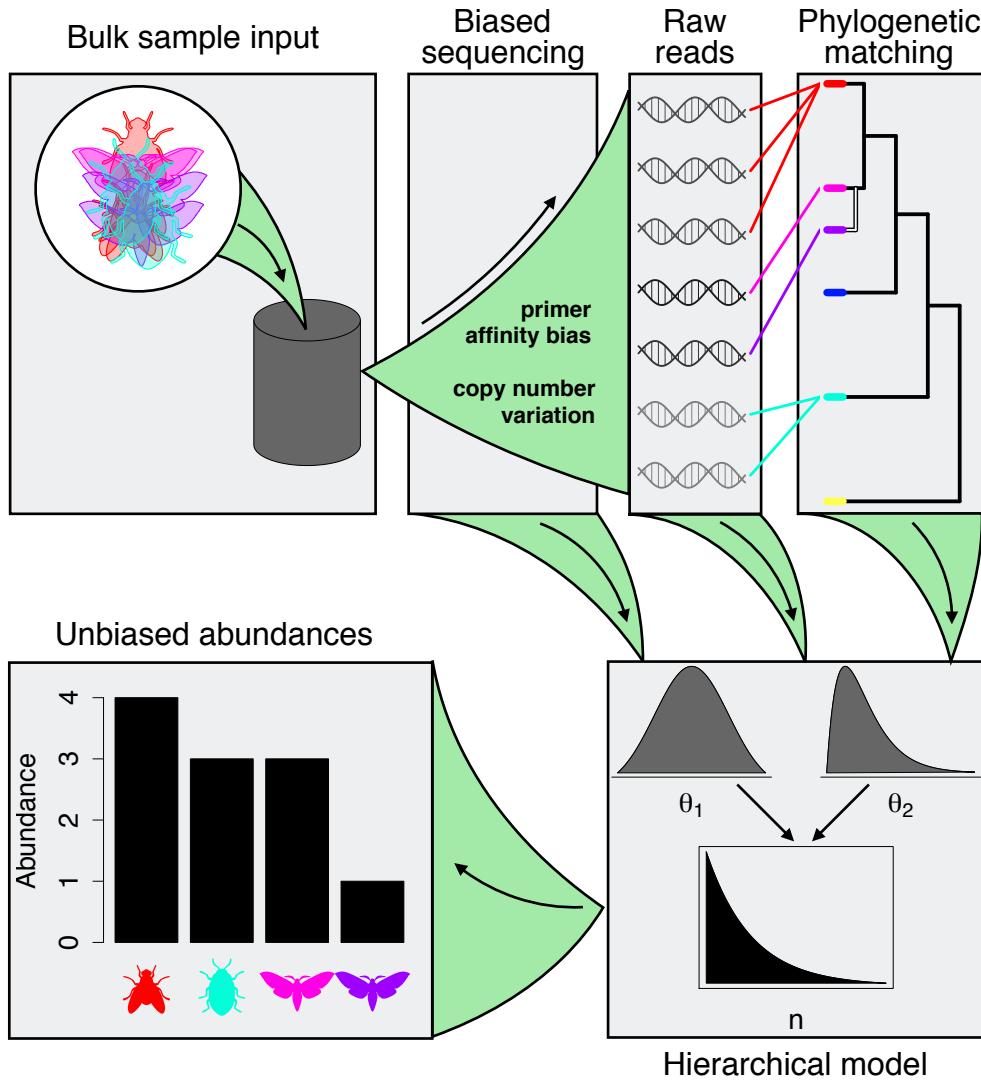


Figure 4: Pipeline for generating and analyzing metabarcoding samples.

(2) *Matching the resulting sequences to a reference phylogeny for identification.* In order to resolve the taxonomy of sequences derived from mixed samples we are developing a library of the barcoding region for species across the Hawaiian archipelago, such that unknown sequences can be phylogenetically matched to the reference library. Sequences not found in the tree of all reference sequences will be grafted and their status as a unique operational taxonomic unit assessed using a cutoff of 3% divergence (Fig. 4). These bioinformatic steps will be included in the R package.

In collaboration with taxonomist and ecologist on Hawaii, we are currently working on generating the barcode reference library for a diverse range of several hundred Hawaiian arthropod taxa. These taxa were sampled across the chronosequence of the Hawaiian Archipelago (Fig. 3). DNA is extracted from each taxon and reference sequence generated for the mitochondrial COI barcoding region. To achieve a comprehensive sampling of the Hawaiian arthropod diversity, samples from environmental gradients (e.g. precipitation) will be included in this reference collection. Such gradients have been shown to have a profound influence on community composition on Hawaii [45].

In order to build a robust phylogenetic backbone for our reference library, the genomic DNA extracts

for all species will be sequenced using the Illumina HiSeq2500. An assembly of the resulting reads promises to generate near complete mitochondrial genomes and nuclear ribosomal clusters of each taxon. To support the Illumina short read assemblies, we will generate long read information by PacBio sequencing. The resulting sequence information will allow us to reconstruct a well resolved community-phylogenetic framework for ecological hypothesis testing. These same specimens will also be used to quantify the microbiomes and feeding habits of hundreds of arthropod species across our sites (discussed further in section “Quantifying networks of microbes, arthropods and plants”).

(3) *Using Bayesian hierarchical models to reconstruct unbiased estimates of abundance* Bayesian hierarchical models permit inference of key quantities (e.g. abundance) while accounting for multiple sources of error and leveraging heterogeneous data types to facilitate inference [46]. The goal of hierarchically modeling metabarcoding data is to estimate the abundances of species while correcting for known biases inherent in amplicon-based sequencing. We will account for bias from copy number variation and primer affinity [44] by directly modeling it, while also using data on the total number of individuals being sequenced, their body sizes, and the phylogenetic relationship between their sequences to constrain the estimates to be more accurate. Furthermore, information from controlled experiments (for example making mock communities of known composition and sequencing those) can be used to constrain prior distributions and obtain even more accurate abundance estimates.

2.9 Quantifying networks of microbes, arthropods and plants

Using the specimens reserved from metabarcoding (i.e. those used to build the reference library and phylogenetic backbone) we will sequence the microbial associates of each species and their gut contents, for herbivorous arthropods. These sequences will allow us to reconstruct the networks between arthropods and their microbial associates as well as herbivorous arthropods and their plant hosts. We will additionally reconstruct microbial networks based on covariance between prevalence of microbial taxa in samples using established approaches [47].