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

## Synopsis

## 1 Background

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

Biodiversity results from both evolutionary and historical processes operating at larger spatiotemporal scales and ecological processes operating at smaller scales (Lessard et al. 2012). Feedbacks between processes along this evolutionary-ecological continuum drive non-steady state biodiversity dynamics [**?** **?** **?** **?** ]. The consequences of non-steady state dynamics are profound for ecosystem function [**?** ], and based on state shifts in the geologic past, the consequence for evolution will persist for millions of years [**?** ]. Yet we lack approaches that synthesize across scales of space and evolutionary time to understand the consequences of this eco-evolutionary feedback process. The propensity for systems to transition into non-steady states cannot be assessed given current means of synthesizing ideas from ecology with those from evolution. Additionally, a lack of cross-scale biodiversity data (from plots to landscapes and genes to species) combined with a lack of theoretical framework, limit this synthesis.

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

Recent theoretical developments have brought mechanistically simplified theory to the forefront of ecological research [2**?** **?** , 1]. These simple theories have been critical because they provide robust null models against which to compare real biodiveristy patterns in order to rigorously test the importance of specific mechanisms in shaping biodiversity. The maximum entropy theory of ecology [METE 1] provides one of the most useful null predictive frameworks because it produces many falsifiable patterns (the species abundance distribuiton, metabolic rate distribution, species area relationship and network structure) and is grounded in the principles of statistical mechanics [1**?** ]. METE draws from the probabilistic properties of large, randomly assembled systems [1] and thus its predictions constitute a community in statistical steady state. Statistical steady state means specifically that a system is governed by only a few simple state variables, which constitute a state space, and that no additionally processes limit the systemâĂŹs ability to freely explore this state space. This precise definition is made more clear in Box 1. Statistical steady state connects to some notions from the literature on ecological equilibrium, specifically the condition of stationarity [**?** ] and egodicity [**?** ], but is in no way tied [1] to ideas relating to equilibrium as a hypothesized state that ecosystems may attain or be driven away from [].

Deviations from METE allow us to identify ecological systems out of statistical steady state [1**?** ]. Drivers of such non-equilibrium include rapid assembly following disturbance [1] and constraints imposed by evolutionary history and non-neutral adaptive differences between species that violate the statistical assumptions underlying the principle of maximum information entropy [**?** ]. In order to harness these promising properties of METE as a non-equilibrium diagnostic tool more testing is needed to understand how exactly the ecological and evolutionary setting of a community predicts its deviation from METE.

**Box 1: Statistical Steady State**

Whether or not biodiversity dynamics are governed by stable equilibria remains an unsolved question in ecology and evolution (Rabosky 2009; Quental and Marshall 2013; Rabosky et al. 2015; Harmon and Harrison 2015). A statistical steady state exists in an ecological community if changes in biodiversity occur slowly and in sync with environmental changes (Harte 2011). The existence (or non-existence) of such steady states has wide ranging implications. For example, whether conservation should focus on conventional preservationalist paradigms or adaptive management (Levin 1999) 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 (Barnosky et al. 2012).

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.

We propose to use islands of the Hawaiian archipeligo to better understand how and why ecosystems depart from steady state, the consequences of these departures on ecosystem function and biodiversity dynamics, including nutrient cycling and invasibility, and finally, how maximum entropy theory can be used as a tool to identify these departures. Remote island archipelagos provide an opportunity to integrate ecological and evolutionary processes, advancing our understanding of the regulation of biodiversity through the lens of theory. This is particularly true when the component islands are arranged chronologically, as is found in "hotspot" islands that form a geological age gradient representing snapshots of community assembly through evolutionary time. Such islands provide simple and discrete systems,of known age and varying area, allowing them to serve as excellent "natural laboratories" for ecological and evolutionary study in a regional context (Simon 1987; Chadwick et al. 2007; Gillespie and Clague 2009). Our team has a strong foundation of research expertise and experience across the islands on microbes (Brodie), arthropods (Rominger, Gillespie, Gruner and Krehenwinkel), plants (Chase), ecosystems (Giardina) and theory (Rominger and Chase).

We will characterize the ecological communities, including their abundance, diversity and network structure, associated with three critical stages in nutrient cycling: 1) Living plants, the arthropods they support and the microbes supported by both; 2) Plant and animal detritus and its associated

arthropod and microbial communities; and 3) Soil communities of arthropods and microbes. In each of these ecosystem domains we will use the maximum entropy theory of ecology to characterize departure from statistical steady state. In order to understand the mechanistic causes of these departures we will also evaluate how deviations from METE can be predicted by the ecology and evolution of the organisms comprising each community, testing the hypotheses outlined in Box 2. We will enable this line of research by deliberately sampling plants, arthropods and microbes across multiple spatial scales, and across gradients of environment (precipitation and elevation as a surrogate for temperature) and substrate age (as a surrogate for both biogeochemical change and evolutionary development). We will also make use of long term fertilization experiments [**?** ] to evaluate the orthogonal roles of evolutionary history versus biogeochemical processes in driving biodiversity patterns. Using plants, arthropods and microbes as discrete test cases, representing a breadth of life history strategies across the tree of life, we will test hypotheses (outlined in Box 2) about deviations from statistical steady state based on how organisms persist, adapt and speciate in their environments. In order to understand how communities are likely to change in response to non-analog, anthropogenically-driven climate regimes and across spatial scales we will build spatially explicit models that link the mechanistic drivers (e.g. rapid community or population change, and evolutionary novelty) of deviation from statistical steady state to remotely sensed data and detailed ecosystem characterizations taken at the NEON site in Hawaii, and our complementary sampling locations. Our project will contribute theoretical constructs for use across NEON sites and bioinformatic tools to advance the rate and dimensionality of biodiversity data gathered at these sites.

# 2 Proposed Research

# 3 Research objectives and hypotheses

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

*Hypotheses*

- Departures from statistical steady state
  - H1 Deviations from METE are largely predicted by age along the chronosequence. These deviations along the chronosequence will be driven primarily by two processes related to evolutionary assembly of biotas: (H2a) primary succession (both by long distance dispersal and speciation) of newly formed habitats; and (H2b) adaptive evolution leading to unique constraints on assembly not consistent with statistical steady state
    - H1a will be more relevant for generalist taxa, especially those that are dispersal limited, on young substrates.

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

H1b will be most relevant for specialist taxa once they have established intricate evolutionary relationships with their coexisting species and environments.

- · We predict greatest deviations from METE for communities dominated by specialist taxa on old substrates
- · We predict a positive correlation between network specialization and deviation from METE
- · We predict a negative correlation between phylogenetic diversity and deviation from METE

H1c Because niche specialization and dispersal limitation both likely result in strong spatial structuring of communities, measures of spatial turnover and deviations from METE should be correlated across all ages along the chronosequence

H1d Because rapid population expansion, population contraction, limited dispersal and local adaptation all lead to low allelic diversity within populations we predict genetic diversity to be negatively correlated with deviation from METE

H2 Deviations from METE are not predicted by environmental variables after accounting for ecosystem age. This includes the prediction that in long term fertilization experiments, fertilized communities will conform to the same patterns as their unfertalized control communities of the same age regardless of underlying nutrient availiblity

H3 However, with rapidly changing climates we do expect environmental predictors of deviations from statistical steady state. Specifically, with the creation of novel environments and loss of existing environments due to changing climate we expect rapid population changes and exacerbated constraints on movement due to unique evolutionary adaptations to previously stable environments. Thus we predict novel climatic conditions to drive future deviations from METE

H4 We predict that in disturbed systems the only what for statistical steady state to be achieved is through rapid assembly of novel ecosystems (i.e. communities dominated by highly vagile invasive taxa). Thus deviations from statistical steady state are expected to promote invasion, while invasion itself will tend to return systems to statistical steady state.

- Evolution of niches and networks

H5 We predict that niches will become more constrained across evolutionary time

H5a Reconstructed niches will be smaller for taxa endemic to older islands

H5b Spatial turnover will be stronger across gradients on older islands

H5c We predict networks will become more specialized across evolutionary time

### 3.1 Significance and Rationale

Understanding how environmental change will alter the feedback between ecology and evolution and drive biodiversity out of statistical steady state is at the core of our proposal. Using METE to capture statistical steady state and understand deviations from it promises to be a powerful diagnostic tool in evaluating ecosystems nearing tipping points. Hawaii is an ideal study system to realize this potential due to its varying chronology (allowing tests of theory in communities of different stages of evolutionary development) and due to its replicated environmental gradients across this chronology (see Fig. **??**. The NEON site at Puu Makaala Natural Area Reserve on Hawaii Island will provide the core measures

needed to quantify the abiotic environment. We will replicate these measurements across gradients of elevation and precipitation, using ground-truthed remotely sensed measurements to provide both fine grain and broad-scale environmental data products.

The same ability to generate massive amount of environmental data via remote sensing does not exist for organismal ecology and evolution. As part of our Dimensions in Biodiversity grant, PIs Rominger and Krehenwinkel are developing laboratory and bioinformatic methods to obtain sequence data, and estimates of abundance and biomass for thousands to millions of arthropods collected via ecological sampling. As part of the current proposal this promising new approach will be developed into an open source lab protocol and software package that can be distributed across all NEON sites.

Our use of METE as a diagnostic tool has been corroborated in the Hawaiian system with previous and current work. PI Rominger, with co-PIs Gillespie and Gruner as collaborators and co-authors, has shown that deviations from METE show consistent patterns across the chronosequences for different arthropod guilds with different life history characteristics (see Box 3). This work needs to be extended to other members of the ecosystem to understand its generality and the mechanistic drivers of deviations from METE, and thus statistical steady state, need to be better understood and modeled into the future to understand how ecosystems will respond to changing climates.

## 3.2 Methods

### 3.3 Integration with NEON and sampling design across environmental and age gradients

*NEON site.* The goal of NEON is to provide ecological data at multiple spatial and temporal scales. Our plan is anchored with the Pu'u Maka'ala Natural Area Reserve on the Mauna Loa volcano on the Big Island of Hawaii (19.553, -155.317), a Core Terrestrial site with the launch date planned for 2017. The site represents montane wet forest with mostly native vegetation dominated by the endemic tree, Metrosideros polymorpha (Myrtaceae). However, up to 95% of the worldâĂŹs terrestrial climates are represented in the greater region of the Hawaiian archipelago, and a single site will fail to characterize this tremendous diversity in climate, habitats and species composition. By replicating core NEON protocols at carefully selected sites with orthogonal variation in temperature and precipitation, along a geological chronosequence representing evolutionary time, the Hawaiian macrosystem will yield the precision of NEON measurements to test ecological theory and to predict consequences of future changes in climate. We aim to combine data to be collected with data from sites across the Hawaiian Islands, in order to understand regional-scale ecological processes and how these respond to change over space and time.

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

*Sampling approach and collection of organismal data.* We will select sites in clearly defined ohia/koa montane, wet and mesic forest communities. The rationale here is that (i) Ohia (Metrosideros polymorpha) is the dominant canopy tree in these forests, forming a nearly continuous layer, with patches of sub-dominant koa (Acacia koa) and numerous associated understory trees, shrubs, herbs, and ferns. 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 (Asner et al. 2006; Asner and Martin 2009; Asner et al. 2011)) reflects the coupled but nonlinear effects of ecosystem age and fertility, which in turn affects the community of organisms in a given forest stand (Crews et al. 1995; Gruner 2007b). Differences in plant traits can affect the structure of an entire food web through a series of direct and indirect effects (Gruner et al. 2005; Bukovinszky et al. 2008).

Figure **??** details the proposed layout of our sampling plots. Within each 1-ha site, we will establish three 20-m by 20-m plots to be selected as representative of forest height mean, maximum, heterogeneity found in that 1-ha site. Within each 20mx20m plot, we will establish our replicate plots. Each plot will be further gridded into 4 m quadrats (100 in total). Within each quadrat we will record all tree species $\geq$ 1 cm at breast height. Within three randomly selected quadrats we will also sample all herbaceous species. We will sample all arthropods within each quadrat using timed beating (24 seconds per quadrat). Within the same three randomly selected quadrats we will also extract arthropods using Berlese funnels from litter and soil samples, gridded to 1 $m^2$ cells (in keeping with the ground beetles collected at the NEON site). Arthropods will be collected into RNAlater to preserve their DNA and RNA as well as the DNA and RNA of their associated microbes and gut contents. While NEON protocols focus on ground beetles (Carabidae), mosquitoes (Diptera: Culicidae), and ticks (order Ixodida), our study will include all arthropods because ground beetles constitute an eclectic group of lineages, most often arboreal and unevenly distributed across the main islands, and there are no native mosquitoes or ticks.

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

In all systems, microbial diversity will focus primarily on the Domain Bacteria due to its phylogenetic breadth, and metabolic and respiratory plasticity. Bacterial diversity will be estimated using molecular tools to sequence 16S rRNA gene biomarkers in multiplex using a barcoding approach. DNA extraction and 16S rRNA gene amplification and Illumina sequencing will be carried out according to Earth Microbiome Project standards (http://www.earthmicrobiome.org/emp-standard-protocols/). Ancilliary and meta data collection standards will follow the NEON the soil microbial data collection and metadata tracking worksheet (http://goo.gl/nE9zPk).

Microbial 16S rRNA gene data will be analyzed according to Shi et al (2015). Richness will be estimated using both taxonomic (OTUs) and phylogenetic (FaithâĂŹs phylogenetic distance) metrics. Absolute bacterial abundances will be determined using quantitative PCR as described in Shi et al (2015) while relative abundances of bacterial taxa will be determined based on the fractions of sequence reads assigned to each taxon using adjustments for rRNA gene copy number (Kembel et al 2012).

In order to relate bacterial taxa to metabolic rate we will use observed relationships between rRNA copy number, genome size and metabolic rate (DeLong et al, 2010). Species area relationships (SARs) for bacterial communities will be calculated from volumes of 0.1 cm3, 0.5 cm3, 1cm3, 10cm3.

*Environmental and biogeochemical data*

1. Plot-level measurements: In each microbial sampling quadrat we will deploy data loggers to record air temperature and moisture content. We will similarly deploy data loggers to record soil temperature and moisture. We will also measure soil physical characteristics, pH, total carbon, nitrogen, phosphorous and sulfur. We will measure monthly litterfall using litter traps as a surrogate for nutrient cycling [**?** ] in addition to litter chemistry (pH, total carbon, nitrogen, phosphorous and sulfur).
2. Remote Sensing and Measurements of Gases: The NEON site will track fluxes of gases, such as carbon dioxide (CO2) and water vapor, and collects data about physical and chemical climate conditions, such as temperature, barometric pressure and visible light or Photosynthetically Active Radiation (PAR). Sensors on the NEON tower systems track fluxes of gases (CO2, water vapor) and collects data about physical and chemical climate conditions, such as temperature, humidity, wind, and the amount of gas that is exchanged between the atmosphere and the ecosystem. Towers extend past the top of the vegetation canopy at each site to allow sensors mounted at the top and along the tower to capture the full profile of atmospheric conditions from the top of the vegetation canopy to the ground. Automated tower sensors collect data continuously to capture patterns and cycles across various time periods, ranging from seconds to years. Categories of measurements are physical climate (aerosols, precipitation, radiation, and temperature, pressure and wind; chemical climate (wet deposition, chemistry, isotopes and scalar concentrations); net ecosystem exchange: carbon dioxide (CO2) flux, soil CO flux, water vapor and latent heat flux, sensible heat, total reactive nitrogen (NO2) and ozone (O3).
3. Airborne Remote Sensing: We will make use of both existing and planned airborne remote sensing data which can provide information on vegetation composition and land cover and will be used in particular to examine the complex mosaic of forest structure and composition. The NEON Airborne Observation Platform (AOP) measures vegetation biochemical and biophysical properties with spectroscopy, vegetation structure and biomass with LiDAR, and produces high resolution imagery that can be subject to analyses of land use and relative cover.

### 3.4 Modeling evolutionary and environmental drivers of assembly

*Maximum entropy theory of ecology across gradients of environment and age* To test our hypotheses relating age, environment and organism/community traints to deviations from METE we will use the R package meteR [developed by Rominger **?** ] to evaluate the goodness of fit of METE for soil microbes, arthropods and plants at our sampling sites across gradients of precipitation, elevation and age. Goodness of fit will be measured as the normalized log likelihood squared [described in **?** ]. Using generalized linear models we will evaluate how the goodness of fit varies between major groups (microbes, arthropods and plants) and as a function of the underlying age and environment of each site.

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

*Modeling niches, networks and community phylogenetics across space and evolutionary time* We will develop a Bayesian hierarchical modeling framework to understand how these drivers of deviations from statistical steady state response to local and regional environments. In all models we will incorporate explanatory environmental variables as spatial averages with an exponentially decaying distance weighted function. Each variable will receive maximum weight at the point location of the specimen

and exponentially less weight as distance from the point location increases. The exponential rate of decay will be fit as a free parameter in our Bayesian hierarchical model.

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

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

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

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

### 3.5 Projecting deviations from statistical steady state into the future

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

### 3.6 Quantifying evolutionary and macroecological patterns using metabarcoding

Next generation sequencing technology has ushered in a revolution in evolutionary biology and ecology. This revolution has not passed by taxonomy and spurred various new studies in the field of molecular barcoding. The current leap in sequencing throughput allows to routinely perform barcoding studies on bulk samples and analyzing whole ecosystems (Taberlet et al. 2012; Leray & Knowlton 2015; Gibson et al. 2014; Ji et al. 2013). The large scale recovery of species richness, food web structure, cryptic species, identification of juveniles and hidden diversity, e.g. internal parasitoids, promise unprecedented new insights into ecosystem function and assembly (Krehenwinkel et al. 2016; Shokralla et al. 2015; Shokralla et al. 2012; Kress et al. 2015; Kartzinel et al. 2015). While species richness can be routinely identified by sequencing bulk samples, estimating species abundance remains challenging (Elbrecht & Leese, 2015) and severely limits the application of metabarcoding to many studies. We are developing wet lab and bioinformatic methods to overcome this issue and revolutionize the generation of ecological and genetic data. Our pipeline consists of three steps (Fig. **??**):

1. Extraction and sequencing of pooled community samples
2. Matching the resulting sequences to a reference phylogeny for identification

3. Using Bayesian hierarchical models to reconstruct unbiased estimates of abundance

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

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

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

*(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 cuttoff of 3% divergence (Fig. **??**). 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. **??**). 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 (Zimmermann & Vitousek 2012).

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 [**?** ]. 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 [] 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 (Fig. **??**). 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.

### 3.7 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 [**?** ].