EENS sections

Linking null models in evolution and ecology with next generation sequencing data to illuminate non-equilibrium dynamics of biodiversity

Introduction

Particularly in the current era of dramatic and often unprecedented disturbances to ecological and evolutionary processes, one of the key objectives for biodiversity scientists is to understand and predict how eco-evolutionary systems respond to perturbations. Eco-evolutionary systems are intrinsically complex systems, in the sense that they are made up of numerous components (genes, individuals, species) interacting via a multitude of interrelated, often stochastic and/or nonlinear, pathways. Rather than attempting to model and predict the behavior of these systems in exacting mechanistic detail (e.g. the exact population dynamics of individual species), it is therefore often pragmatic to focus instead on the emergent, macroscopic properties of these systems (such as the distribution of abundance among all species).

The macroscopic focus is especially promising for understanding how disturbances affect eco-evolutionary systems. Macroscopic properties tend to be relatively insensitive to minor changes at lower levels of organization, due in part to the phenomenon of *staistical equilibrium*. In complex systems in general, when a large number of random processes combine to generate a phenomenon (such as the shape of the species abundance distribution), the fine-scale details of these processes can smooth or cancel each other out and result in consistent, predictable macroscopic outcomes that converge with the outcomes that emerge from simple null models. Through statistical averaging, these macroscopic phenomena are relatively insensitive to subtle variations or perturbations at lower levels of organization. However, a sufficiently strong perturbation can temporarily overwhelm statistical equilbrium and produce macroscopic outcomes that deviate from the expectations of statistical null models. These deviations can be used to detect and diagnose strong perturbations operating in complex systems. This conceptual framework is the foundation of, for example, applications of the principle of maximum entropy in statistical mechanics.

Efforts to use deviations from statistical null theories in ecology and evolution have demonstrated that this is a useful approach. Rather than assuming any one mechanism dominates the assembly of populations into a community, these theories assume all mechanisms could be valid, but their unique influences have been lost to the enormity of the system and thus the outcome of assembly is a community in statistical equilibrium (Harte, 2011; Pueyo et al., 2007). The mechanistic agnosticism is what makes statistical theories useful nulls. These statistical theories are also consistent with niche-based equilibria (Neill et al., 2009; Pueyo et al., 2007) if the complicated, individual or population level models with many mechanistic drivers were to be upscaled to entire communities. In ecology, the Maximum Entropy Theory of Ecology (METE) has been shown to successfully predict the shape of numerous macroscopic distributions given only a restricted set of "state variables" (species richness, total abundance, and optionally total metabolic flux) and minimal assumptions about ecological process. In contrast, systems that have recently undergone major disturbances - such as forests that have been clear-cut, or oceanic island communities that have only recently emerged - deviate markedly from the statistical null expectations derived from METE. In population genetics, deviations from neutral genetic drift are routinely used to detect and diagnose disturbances such as population bottlenecks.

Historically, applications of statistical equilibrium in ecology and evolution have proceeded separately - i.e. the fields of macroecology () and population genetics - but integrating ecological and evolutionary statistical nulls can provide new insights not possible through focusing on either of these phenomena separately

. While macroecology has revealed several ecological patterns that have provided useful insight into general processes and the importance of statistical nulls in ecology, it has become increasingly clear that the patterns traditionally entertained in macroecology do not, on their own, contain sufficient information to distinguish among competing hypotheses. In particular, macroecology has largely adopted a time-averaged perspective that is not capable of accounting for temporal or historical dynamics. Incorporating population genetics with macroecology would provide additional data dimensions for statistical inference, and would provide a *historical* perspective that is critical for understanding how past disturbances influence contemporary patterns. On the other hand, ecological population genetics has historically focused on the dynamics of a single focal species or population, and only indirectly incorporated larger scale processes (but see ...).

So far, progress towards integrating statistical null models in macroecology with community genetics has shown promise (...) but has been largely constrained by the lack of a unified theoretical framework for population genetic and macroecological patterns, on one hand, and the lack of empirical data incorporating both organismal abundance and community genetic data, on the other. In the absence of a quantitative framework linking macroecological and population genetic dynamics, the field has instead used retroactive modifications to macroecological models () or explored qualitative hypotheses about the co-variation of, for example, species and genetic diversity (). However, recent advances in simulation modeling of *joint* ecological and evolutionary processes open up new ground for developing and testing rigorous predictions for how various types of equilibrial and non-equilibrial dynamics are expected to manifest at *both* organismal and genetic levels (...). Simultaneously, next-generation sequencing technologies will soon make it economical and efficient to collect community-wide genetic data to complement organismal abundance data. The data produced via these methods are naturally suited to testing theoretical predictions for joint genetic and community dynamics, and can provide inferential leverage for diagnosing present and past non-equilibral processes not discernible using present approaches.

Here, we illustrate how statistical null models of joint population genetic and ecological dynamics, coupled with community genetic sequencing data, can illuminate historical and contemporary trajectories of eco-evolutionary systems. We present a theoretical framework for interpreting deviations from statistical equilibrium at ecological and/or population genetic levels, and use simulation models to demonstrate how different scenarios of non-equilibrial dynamics fit into this framework. We demonstrate using this framework with empirical data on species' abundances and genetic diversity, and illustrate a prospective bioinformatic pipeline for applying this framework to next-generation sequencing data. We hope that these theoretical and methodological advances will inspire future steps integrating community genetic sequencing with statistical null models to understand and predict trajectories of ecological and evolutionary change in the Anthropocene.

Disturbances and deviation from statistical equilibrium

Ecosystems experience regular disturbances and subsequent periods of re-organization, which can occur on ecological time-scales, such as primary succession, or evolutionary time scales, such as evolution of novel innovations that lead to new ecosystem processes (Erwin, 2008). We hypothesize that these regular disturbances can lead to periods or even cycles of non-equilibrium in observed biodiversity patterns. We illustrate these scenarios in the phase space of equilibrium and non-equilibrium states shown in Figure 1. Some scenarios, including a complete cycle through all four phases, cannot be observed without a time machine, but by using community-level genetic data to examine deviations from ecological and evolutionary statistical null models, we can identify where in this space a focal system is located and use this to infer its past and future dynamics.

Explanation of states and transitions in Figure 1

Phase I:

Phase II:

Phase III:

Phase IV:

Cycles between the phases

Simulations Simulation modeling of equilibiral and perturbed eco-evolutionary dynamics allows us to develop quantitative versions of the qualitative dynamics described in Figure 1. Here, we use simulation models to illustrate transitions between Phases X and Y.

Simulation modeling methods

Results of sims

Leveraging next-generation sequencing data

The theoretical framework presented here uses joint data on community-wide genetic structure and species' relative abundances. While such datasets are relatively rare in the literature at present, emerging technologies for community metabarcoding will soon make it logisically tractable and economical to collect wide-scale data of exactly this type.

One major barrier to using joint community genetic and species abundance data derived from metabarcoding is the challenge of estimating species' abundances - rather than simply presence/absence - from metabarcoding data. One solution to this is to conduct organismal surveys associated with genetic data sampling, to either wholly substitute or simply ground-truth abundances estimated from genetic data. Here we present a second option: using model-free abundance estimation to retrive species' abundances from genetic data.

Specifically, [model-free abundance estimation supplement]. We propose a pipeline (Fig. I) where raw reads are generated and assembled into a phylogeny using standard approaches, and potentially aided by additionally available sequence data in a super tree or super matrix approach. The numbers of sequences assigned to each terminal tip are then used in a Bayesian hierarchical model which seeks to estimate the true number of organisms representing each terminal tip, accounting for sequencing biases originating from, e.g. primer affinity and copy number differences between taxa. Information on phylogenetic relatedness can inform modeled correlations in biases between taxa (Angly et al., 2014); e.g. copy number is known to be phylogenetically conserved at least in microbes). This approach is particularly tailored to metabarcoding data. In a potentially powerful extension, and thanks to the proposed Bayesian framework, information from sequencing experiments that seek to calibrate metabarcoding studies (e.g., (Krehenwinkel et al., 2017)) can be used to build meaningfully informative priors and improve model accuracy. Through a simulation study (described in the supplement) we show that true underlying abundances can be accurately estimated (Fig. II).