The Statistical Mechanics of Biodiversity

by

Andrew James Rominger

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Environmental Science, Policy and Management

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John Harte, Co-chair
Professor Rosemary Gillespie, Co-chair
Professor Charles Marshall

Spring 2016

The dissertation of Andrew James Rominger, titled The Statistical Mechanics of Biodiversity, is approved:

Co-chair _____ Date _____

Co-chair _____ Date _____

_____ Date _____

University of California, Berkeley

The Statistical Mechanics of Biodiversity

Abstract

The Statistical Mechanics of Biodiversity

by

Andrew James Rominger

Doctor of Philosophy in Environmental Science, Policy and Management

University of California, Berkeley

Professor John Harte, Co-chair

Professor Rosemary Gillespie, Co-chair

Invasive brag; forbearance.

I dedicate this thesis to my guides and sources of solace through the process of completing my Ph.D: to my mother Judy Rominger and my godmother Marsha Keener; to my partner Linden Schneider; and to the kolea *(Pluvialis fulva)* and giant koa *(Acacia koa)* of Hawaii.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to first and foremost thank my co-Chairs, Drs. John Harte and Rosemary Gillespie; their dedication and support to fostering my research and advocating for my contribution to the scientific community has been the defining feature of my graduate experience. I would also like to thank my outside member Dr. Charles Marshall who's scientific curiousity and practicle advice have been critical in forging my path as a researcher.

This work would not have been possible without the support and teamwork of many friends and collaborators. I would like to thank especially Dr. Pablo Marquet, Dr. Miguel Fuentes, Dr. Daniel Gruner, Dr. Cory Merow, Dr. George Roderick, Jun Lim, Dr. Kari Goodman and Dr. Matthew Van Dam for their comraderee and tireless collaboration. I would like to thank Dr. Manisha Anantharaman, Dr. Naim Darghouth, Dr. Danielle Christianson, Erica Newman, Dr. David Hembry, Jade Zhang and Meredith Jabis for their friendship and support.

This work is rooted in the diversity of nature, particularely the unique biota of Hawaii. I am grateful to those who have dedicated their lives to preserving and restoring that unique legacy, especially Pat Bily, Ed Misaki and Russell Kallstrom of The Nature Conservancy Hawaii, Rhonda Loh and Raina Kaholoa'a with the National Park Service, and Cynthia King and Betsy Gagne with the Hawaii Department of Land and Natural Resources.

I would also like to acknowledge my funders including the National Science Foundation Graduate Research Fellowship, NSF DEB 1241253, the Moore Foundation, and the Department of Environmental Science, Policy and Management and the Walker Fund at UC Berkeley.s

Introduction

stub

Chapter 1

Punctuated non-equilibrium and niche conservatism explain biodiversity fluctuations through the Phanerozoic

1.1 Background

Biodiversity has not remained constant nor followed a simple trajectory through geologic time [1–6]. Instead, it has been marked by fluctuations in the number of extant taxa, both positive in the case of net origination or negative in the case of net extinction. Major events, such as adaptive radiations and mass extinctions have received special attention [7, 8], but fluctuations of all sizes are ubiquitous [2, 5, 9].

Several approaches have been taken to study the complex trajectory of paleo-biodiversity ranging from the hypothesis that biological systems self-organize to the brink of critical phase-transitions [10, 11] to invocations of non-linear environmental perturbations [12] and escalatory co-evolutionary interactions [13]. New data and analyses have not supported any of these hypotheses at the scale of the entire Phanerozoic marine invertebrate fauna [5, 14, 15]. Other studies have modeled the mean trend in diversity as tracking a potentially evolving equilibrium [2, 5, 6, 16] and yet ignore the potential role of stochasticity and non-equilibrium dynamics in producing observed patterns [4, 9, 17]. As such, we still lack a synthetic theory of evolving biodiversity through the fossil record. Here we use a simple model of evolution in an abstract niche space derived from universal non-equilibrium processes to predict, with great accuracy, the complex distribution of pervasive diversity fluctuations throughout the marine Phanerozoic.

Despite the heterogeneity of explanations of Phanerozoic biodiversity, consensus has emerged on three properties of macroevolution: *i*) gross ecological and life history attributes of clades (i.e. groups of related species descending from a common ancestor) are often maintained, a phenomenon known as niche conservatism [18, 19]; *ii*) long periods of niche conservatism are interrupted by adaptive diversification and exploration of new eco-

logical niche space [19–21]; and *iii*) as a consequence of the interaction between their life history characteristics and the dynamics of the environments they inhabit [22] different clades experience different rates of morphological evolution, speciation and extinction [2, 3, 23, 24].

Observed bursts of adaptive radiation leading to novel morphologies in the fossil record led Eldredge and Gould to their hypothesis of punctuated equilibrium [20]. Here we show that this punctuation is actually akin to the "super statistical theory" of non-equilibrium dynamics in statistical physics [25]. Super-statistics [25] proposes that non-equilibrial systems can be decomposed into locally equilibrial sub-systems. The distribution of equilibria across sub-systems determines the dynamics of the complete system [25]. When these sub-systems are superimposed the resulting system can no longer be described by a single equilibrial model. In the context of macroevolution we propose that a clade with conserved life history characteristics corresponds to a locally equilibrial sub-system. If a certain region of niche space can only contain a finite diversity of taxa [16, 23, 26, 27] then diversity within clades should fluctuate stochastically about this equilibrium due to random origination and extinction. The magnitude of these macroevolutionary rates should be a function of the life history and ecological characteristics that define that region of niche space. Larval type [28], body plan [17], body size [29], range size [29, 30] and substrate preference [19] have all been shown to influence such rates. Thus different regions of niche space, and the clades occupying them, will experience different magnitudes of stochastic fluctuation in diversity. As clades occasionally split to fill new regions of niche space their punctuated diversification determines the non-equilibrium nature of the entire biota. Here we show that these properties of macroevolution are sufficient to explain the complex fluctuations of marine invertebrate diversity through the Phanerozoic.

In statistical mechanics, local sub-systems can be defined by a simple statistical parameter β often corresponding to inverse temperature. In the context of macroevolution we define β as the inverse variance of a homogenous origination-extinction process, which will capture all relevant information about a clade's diversification under such a process. The limit distribution of time averaged fluctuations in clade k's diversity through time should be approximately Gaussian with variance $1/\beta_k$ (Appendix A.1). We posit that just as in statistical systems, non-equilibrium dynamics can arise from the mixing of the dynamics of many locally equilibrial subsystems. For marine Phanerozoic diversity this corresponds to mixing the dynamics of many clades, all being described by their unique β_k values. Three exemplar dynamics taken from a bias-corrected (see methods section) aggregation of the Paleobiology Database (PBDB) [5] are shown in Figure 1.1. To predict the super-statistical behavior of the entire marine invertebrate Phanerozoic fauna we must integrate over all possible local equilibria that each clade could experience. The distribution of β_k values describes the probability that a given clade, chosen at random, will occupy a region of niche space characterized by that inverse variance value. The form of this stationary distribution of β_k values could shed interesting light on the biological processes that lead different clades to different equilibria, as discussed below. Figure 1.1 shows the shape of this stationary distribution estimated from bias-corrected PBDB [5] data.

To uncover the super-statistical nature of the marine invertebrate Phanerozoic fauna

we analyze the distribution of diversity fluctuations using two canonical databased of fossil biodiversity, the PBDB [5] and Sepkoski's compendium [31] of fossil marine invertebrates (results from Sepkoski's compendium are presented in Appendix A.2.3). We filter PBDB data to include only well preserved marine invertebrates following previously published collection inclusion criteria [5, 6]. We account for detection bias in the PBDB using an extension of the "three timer" correction [5]. "Three timer" correction accounts for the rate of failure to observe a genus, estimated by the number of times a gap occurs in its occurrence history. We extend this correction by also employing a new publication bias correction to help eliminate bias from preferential publication of novel taxa (see methods section). Results obtained from this correction strategy are similar to other published methods (Fig. A.2). Fluctuations within a clade are computed as the difference in standing diversity between two time intervals, or equivalently the number of originations minus extinctions in one interval.

Phanerozoic biodiversity can be deconstructed and grouped into clades, or sub-systems, in several different ways. Lacking a full phylogenetic hypothesis for all marine invertebrates we use taxonomic classifications to identify potential sub-systems. Taxa ideally represent groups of organisms that descend from a common ancestor and share similar ecologically and evolutionary relevant morphological traits [32, 33]. For Phanerozoic marine invertebrates, Holman [24] has shown that variance in diversity dynamics is less between taxa belonging to the same order than taxa in different orders, indicating that the taxonomic level of orders is a likely candidate for sub-system delineation. To evaluate the optimal taxonomic level for sub-system designation, we test our superstatistical theory using taxonomic levels from order through class to phylum. Additionally we compare our results to randomized taxonomies to confirm that observed patterns are not an artifact of arbitrary classification but instead represent real biologically relevant differences between clades.

1.2 Methods

1.2.1 Paleobiology Database data download and filtering

Data were downloaded from the Paleobiology Database (PBDB; www.pbdb.org) on 28 May 2013. Collections were filtered using the same approach as Alroy [5] to insure that only well preserved marine invertebrate occurrences were used in subsequent analysis resulting in 221202 genus occurrences. These were further filtered to exclude those occurrences without order-level taxonomy and those collections with age estimate resolutions outside the 10my default bins of the PBDB resulting in 189516 occurrences left for analysis. To avoid basic sampling concerns we excluded the earliest Cambrian and the Cenozoic.

To focus attention on the variance of fluctuations we center each clade's fluctuation distribution. Because "equilibrium" in the statistical mechanical sense means a system undergoes coherent, concerted responses to perturbation the mean trend line is of less interest than deviations from it. We also note that most clades are already close to centered and so centering has little influence on clade-specific fluctuation distributions.

1.2.2 Three-timer and publication bias correction

We use a new and flexible method, described below, to correct for known sampling biases in publication-based specimen databases [5, 6]. We were motivated to use this method because rarefaction has been shown to under-perform compared to the more data-intensive shareholder quorum subsampling (SQS) method [6]. However, subsampling cannot be applied to small orders (i.e. the majority) because SQS becomes increasingly unreliable as sample size decreases [6]. We therefore develop a simple method based on first correcting for detection bias using the "three timer" correction [5] in which the rate of failure to observe a genus is estimated by the number of times a gap occurs in the occurrence history of each genus. To eliminate further bias due to preferential publication of novel taxa we divide observed number of genera per order per time period by the expected number of genera given publications in that time period. The expected number is calculated by regressing the log-transformed number of genera on log-transformed number of publications. There is a weak trend toward higher diversity with more publications (Fig. A.1) meaning that the most important correction comes from the three timer correction.

Our new method effectively re-scales each genus occurrence from 0/1 (absent/present), to a weighted number continuously ranging between 0 and 1. This method achieves similar results to more computationally intensive sub-sampling procedures [5, 6]. We directly compare our predicted time series of global genus diversity with results derived from SQS [6] and the raw data (Fig. A.2). Our method shows minor differences with the SQS prediction, However, these discrepancies do not have impact the distribution of fluctuations (Fig. A.2) and super-statistical analysis on uncorrected PBDB data (see section A.2.1) produces a similar result to the analysis on corrected PBDB data presented in the main text.

1.2.3 Numerical methods

To fit our super-statistical prediction we use the method of least squares instead and maximum likelihood. When building the prediction for $P(x)$ by calculating order-level Gaussian distributions and integrating over them, we use least squares to fit the variance term to each order. We do so because orders potentially show asymmetries in their distribution of fluctuations. Least squares is more flexible in fitting such distributions compared to maximum likelihood which will always estimate the empirical variance as the best-fitting parameters.

We also estimate $P(x)$ directly from the raw data using maximum likelihood to compare the fit of our super-statistical prediction and that of a simple Gaussian distribution using AIC. To calculate a likelihood-based confidence interval on our prediction we bootstrapped the data, subsampling fluctuations with replacement from all orders combined.

1.3 Results and Discussion

At the order level in both the sampling-corrected PBDB (Fig. 1.1) and Sepkoski's compendium (Appendix A.2.3), fluctuations in genus diversity are well described by a Gaussian

distribution (Fig. 1.1). Gaussian fluctuations would result from a homogeneous origination-extinction process under the condition of independence between orders. Independence could result from neutral-like processes [34], where the dynamics of one taxon are unaffected by those of another, or from dampening mechanisms that stabilize complex networks of interacting taxa [35]. This is in direct contrast to the instability hypothesis underlying the self-organized criticality theory of paleo-biodiversity [10, 11].

We estimate the distribution of β_k's simply as the maximum likelihood distribution describing the set of inverse variances for all orders. In both the PBDB and Sepkoski's compendium, Phanerozoic marine invertebrate orders clearly follow a Gamma distribution in their β_k values (Fig. 1.1). While multiple processes could lead to a Gamma stationary distribution (e.g. [36]), one interesting possibility is a mean-reversion process [36]. Mean reversion could be a consequence of niche conservatism if β_k values are associated with a clade's physiological and life history traits, themselves evolving via Ornstein-Uhlenbeck-like exploration of an adaptive landscape [36, 37].

Using the observation of within order statistical equilibrium and Gamma-distributed β_k parameters we can calculate, without further adjusting free parameters, the distributions of order-level fluctuations for the entire marine Phanerozoic, $P(x)$, as

$$P(x) = \int_0^\infty p_k(x \mid \beta) f(\beta) d\beta \tag{1.1}$$

where $p_k(x \mid \beta)$ is the distribution of fluctuations within an order and $f(\beta)$ is the stationary distribution of inverse variance in the magnitude of order-level fluctuations in diversity. This leads to a non-Gaussian, fat-tailed prediction for $P(x)$ which matches both the PBDB and Sepkoski data closely (Fig. 1.2 and Appendix A.2.3).

To quantitatively evaluate how well the super-statistical prediction matches the data we constructed a 95% confidence envelope from bootstrapped maximum likelihood estimates for $P(x)$. Observed fluctuations fall within this 95% confidence envelope (Fig. 1.2), indicating that the data do not reject the super-statistical prediction. For further comparison, we fit a Gaussian distribution to the observed fluctuations, which corresponds to the equilibrium hypothesis that all orders conform to the same statistics. Using Akaike Information Criterion (AIC) we find that observed fluctuations are considerably better explained by the super-statistical prediction than by the Gaussian hypothesis ($\Delta\text{AIC} = 11285.18$). Thus, as expected under the super-statitical hypothesis, the fat tailed distribution of fluctuations arise from the superposition of independent normal statistics for fluctuations within orders.

Computing the distribution of fluctuations using classes instead of orders leads to a substantially poorer fit to the observed data (Fig. A.4). We quantify this shift with the Kolmogorov-Smirnov statistic, which changes from 0.041 in order to 0.062 in classes (Fig. 1.3). However, if super-statistical theory explains the data, this worsening fit with increasing taxonomic scale is expected as the different classes are not well defined subsystems in their fluctuation dynamics. Instead, classes aggregate increasingly disparate groups of organisms, and thus effectively mix their associated Gaussian fluctuations, meaning that one statistic should no longer be sufficient to describe class-level dynamics. Our analysis indicates that

orders are evolutionarily coherent and independent entities, with all subsumed taxa sharing key ecological and evolutionary attributes allowing them to diversify concertedly and independently from other orders. Both the good fit at the order level and worsening fit at higher taxonomic levels is confirmed in Sepkoski's compendium, which also allows analysis of phylum-level patterns (Fig. A.5).

To further test the evolutionary coherence of orders we conducted a permutation experiment in which genera were randomly reassigned to orders while maintaining the number of genera in each order. For each permutation, we calculated the super-statistical prediction and the Kolmogorov-Smirnov statistic. The permutation simulates a null model in which common evolutionary history is stripped away (genera are placed in random orders) but the total number of observed genera per order is held constant. Repeating this permutation 500 times yields a null distribution of Kolmogorov-Smirnov statistics that is far separated from the observed value (Fig. 1.3) suggesting the good fit at the order level is not merely a statistical artifact of classification but carries important biological information.

1.4 Conclusion

Our analysis makes no assumption that orders should correspond to super-statistical subsystems, but identifies them as the appropriate level for marine invertebrates. Holman [24] has also shown that orders are "evolutionarily coherent" in that subtaxa within orders share common diversification dynamics. As we show, orders differ only in the variances of their diversity fluctuations (Fig. 1.1).

Our study is the first to demonstrate that complex patterns in the sequence of origination and extinction events in the fossil record are the result of a simple underlying process analogous to the statistical mechanisms by which complexity emerges in large physical [38] and social systems [39]. We do so by identifying the biological scale at which clades conform to equilibrial dynamics, which could result from the process of niche conservatism. We then show that punctuated shifts to different equilibria between clades, a consequence of punctuated exploration of niche space by newly evolving clades, leads to a characteristically non-equilibrial distribution of diversity fluctuations when the marine Phanerozoic fauna is viewed as a whole macro-system.

Our work highlights the importance of both niche conservatism and punctuated adaptive radiation in producing the statistical behavior of the Phanerozoic; our theory thus provides new motivation for identifying the eco-evolutionary causes of innovations between lineages and how those innovations are eventually conserved within lineages. Armed with an understanding of the statistical behavior of diversification we can go on to examine mechanisms underlying additional patterns in the mean trend of biodiversity through the Phanerozoic. In particular, clades have been shown to wax and wane systematically through time [4, 9], a pattern that we cannot explain with super-statistics alone.

Super-statistics could also be applied to other areas of evolution and macroecology. For example new phylogenetic models already consider heterogeneous rates of diversification

[e.g. 40]. The super-statistics of clades in adaptive landscapes could provide a means to build efficient models that jointly predict morphological change and diversification. This framework could also provide a new paradigm in modeling the distributions of diversity, abundance and resource use in non-neutral communities. Non-neutral models in ecology are criticized for their over-parameterization [41], yet a persistent counter argument to neutral theory [34] is the unrealistic assumption of ecological equivalency [42] and poor prediction of real dynamics [43]. If ecosystems are viewed as the super-position of many individualistically evolving clades, each exploiting the environment differently and thus obeying a different set of non-equivalent statistics, then diversity dynamics could be parsimoniously predicted while incorporating real biological information on ecological differences between taxa.

Acknowledgments

We thank Charles Marshall, Joseph Felsenstein, John Harte, Rosemary Gillespie and John Alroy for helpful discussion. Michael Foote provided a digitized copy of Sepkoski's compendium. AJR thanks funding sources Fulbright Chile and the National Science Foundation Graduate Research Fellowship Program; MAF thanks FONDECYT 1140278; PM thanks support from Grant PFB-023 (CONICYT) and ICM-P05-002).

References

1. Raup, D. M., Sepkoski Jr, J. J., *et al.* Mass extinctions in the marine fossil record. *Science* **215,** 1501–1503 (1982).
2. Sepkoski, J. J. A Kinetic Model of Phanerozoic Taxonomic Diversity. III. Post-Paleozoic Families and Mass Extinctions. *Paleobiology* **10,** 246–267 (1984).
3. Gilinsky, N. L. Volatility and the Phanerozoic decline of background extinction intensity. *Paleobiology,* 445–458 (1994).
4. Liow, L. H. & Stenseth, N. C. The rise and fall of species: implications for macroevolutionary and macroecological studies. *Proceedings of the Royal Society B: Biological Sciences* **274,** 2745–2752 (2007).
5. Alroy, J. *et al.* Phanerozoic Trends in the Global Diversity of Marine Invertebrates. *Science* **321,** 97–100 (2008).
6. Alroy, J. The Shifting Balance of Diversity Among Major Marine Animal Groups. *Science* **329,** 1191–1194 (2010).
7. Benton, M. Diversification and extinction in the history of life. *Science* **268,** 52–58 (1995).
8. Erwin, D. H. The end and the beginning: recoveries from mass extinctions. *Trends in Ecology and Evolution* **13,** 344–349 (1998).

9. Quental, T. B. & Marshall, C. R. How the Red Queen Drives Terrestrial Mammals to Extinction. *Science* (2013).

10. Bak, P. & Sneppen, K. Punctuated equilibrium and criticality in a simple model of evolution. *Phys. Rev. Lett.* **71,** 4083–4086 (1993).

11. Solé, R. V., Manrubia, S. C., Benton, M. & Bak, P. Self-similarity of extinction statistics in the fossil record. *Nature* **388,** 764–767 (1997).

12. Newman, M. E. J. & Roberts, B. W. Mass Extinction: Evolution and the Effects of External Influences on Unfit Species. *Proceedings of the Royal Society of London B* **260,** 31–37 (1995).

13. Vermeij, G. J. *Evolution and Escalation* (Princeton University Press, Princeton, N.J., 1987).

14. Kirchner, J. W. & Weil, A. No fractals in fossil extinction statistics. *Nature* **395,** 337–338 (1998).

15. Madin, J. S. *et al.* Statistical Independence of Escalatory Ecological Trends in Phanerozoic Marine Invertebrates. *Science* **312,** 897–900 (2006).

16. Rabosky, D. L. Ecological limits and diversification rate: alternative paradigms to explain the variation in species richness among clades and regions. *Ecology Letters* **12,** 735–743 (2009).

17. Erwin, D. H. Novelties that change carrying capacity. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* **318,** 460–465 (2012).

18. Roy, K., Hunt, G., Jablonski, D., Krug, A. Z. & Valentine, J. W. A macroevolutionary perspective on species range limits. *Proceedings of the Royal Society B: Biological Sciences* **276,** 1485–1493 (2009).

19. Hopkins, M. J., Simpson, C. & Kiessling, W. Differential niche dynamics among major marine invertebrate clades. *Ecology letters* **17,** 314–323 (2014).

20. Eldredge, N. & Gould, S. J. Punctuated equilibria: an alternative to phyletic gradualism. *Models in paleobiology* **82,** 115 (1972).

21. Newman, C., Cohen, J. & Kipnis, C. Neo-darwinian evolution implies punctuated equilibria. *Nature* **315,** 400–401 (1985).

22. Vrba, E. S. Macroevolutionary Trends: New Perspectives on the Roles of Adaptation and Incidental Effect. *Science* **221,** 387–389 (1983).

23. Simpson, G. Horotely, Bradytely, and Tachytely. *The Major Features of Evolution,* 313–337 (1953).

24. Holman, E. W. Some evolutionary correlates of higher taxa. *Paleobiology,* 357–363 (1989).

25. Beck, C. & Cohen, E. Superstatistics. *Physica A: Statistical Mechanics and its Applications* **322,** 267–275 (2003).

26. Gavrilets, S. & Vose, A. Dynamic patterns of adaptive radiation. *Proceedings of the National academy of Sciences of the United States of America* **102,** 18040–18045 (2005).

27. Price, T. D. *et al.* Niche filling slows the diversification of Himalayan songbirds. *Nature* **509,** 222–225 (2014).

28. Jablonski, D. Species Selection: Theory and Data. *Annual Review of Ecology, Evolution, and Systematics* **39,** 501–524 (2008).

29. Harnik, P. G. Direct and indirect effects of biological factors on extinction risk in fossil bivalves. *Proceedings of the National Academy of Sciences* **108,** 13594–13599 (2011).

30. Foote, M., Crampton, J. S., Beu, A. G. & Cooper, R. A. On the bidirectional relationship between geographic range and taxonomic duration. *Paleobiology* **34,** 421–433 (2008).

31. Sepkoski, J. J. *A compendium of fossil marine animal families* (Milwaukee Public Museum, Milwaukee, WI, 1992).

32. Mayr, E. Numerical phenetics and taxonomic theory. *Systematic Zoology* **14,** 73–97 (1965).

33. Erwin, D. H. Disparity: morphological pattern and developmental context. *Palaeontology* **50,** 57–73 (2007).

34. Hubbell, S. P. *The unified neutral theory of biodiversity and biogeography (MPB-32)* (Princeton University Press, 2001).

35. Brose, U., Berlow, E. L. & Martinez, N. D. Scaling up keystone effects from simple to complex ecological networks. *Ecology Letters* **8,** 1317–1325 (2005).

36. Cox, J. C., Ingersoll Jr, J. E. & Ross, S. A. A theory of the term structure of interest rates. *Econometrica: Journal of the Econometric Society,* 385–407 (1985).

37. Butler, M. A. & King, A. A. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *The American Naturalist* **164,** 683–695 (2004).

38. Beck, C. Superstatistics in hydrodynamic turbulence. *Physica D: Nonlinear Phenomena* **193,** 195–207 (2004).

39. Fuentes, M. A., Gerig, A. & Vicente, J. Universal Behavior of Extreme Price Movements in Stock Markets. *PLoS ONE* **4,** e8243 (2009).

40. Rabosky, D. L. LASER: a maximum likelihood toolkit for detecting temporal shifts in diversification rates from molecular phylogenies. *Evolutionary bioinformatics online* **2,** 247 (2006).

41. Rosindell, J., Hubbell, S. P. & Etienne, R. S. *The Unified Neutral Theory of Biodiversity and Biogeography* at Age Ten. *Trends in ecology & evolution* **26,** 340–348 (2011).

42. Chave, J. Neutral theory and community ecology. *Ecology letters* **7,** 241–253 (2004).

43. Ricklefs, R. E. The unified neutral theory of biodiversity: do the numbers add up? *Ecology* **87,** 1424–1431 (2006).

Figures

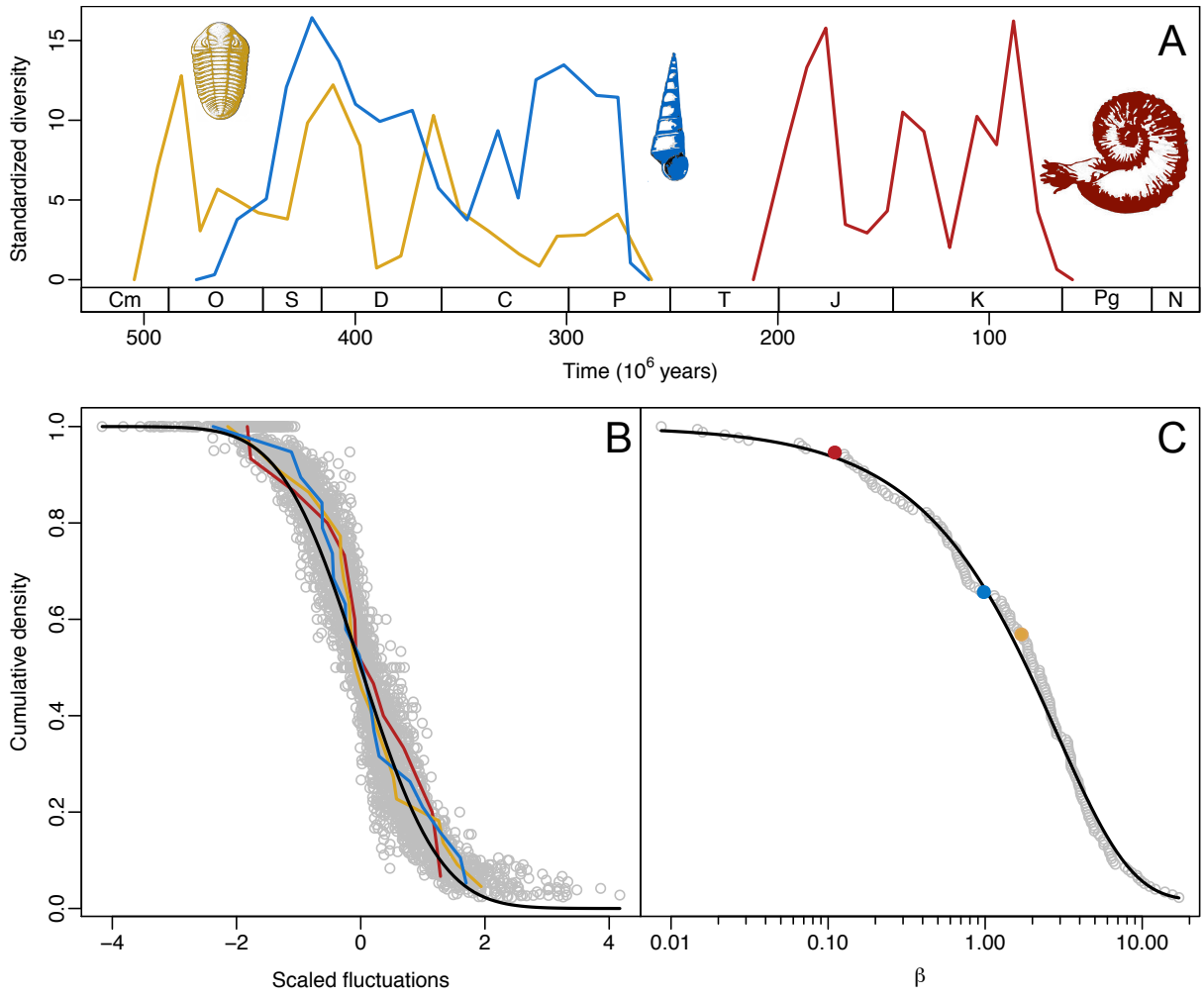

Figure 1.1: The distributions of within-order fluctuations in genus diversity shown for the trajectories of three exemplar orders (A) and shown as an empirical cumulative density aggregated across all orders (B). To display all orders simultaneously we simply collapse their fluctuation distributions by dividing by their standard deviations. If orders conform to the Gaussian hypothesis their scaled fluctuations should fall along the cumulative density line of a normal N(0, 1) distribution, as shown in (B). In (C) the distribution of inverse variances β_k across all orders matches very closely to a Gamma distribution (black line); exemplar orders are again highlighted.

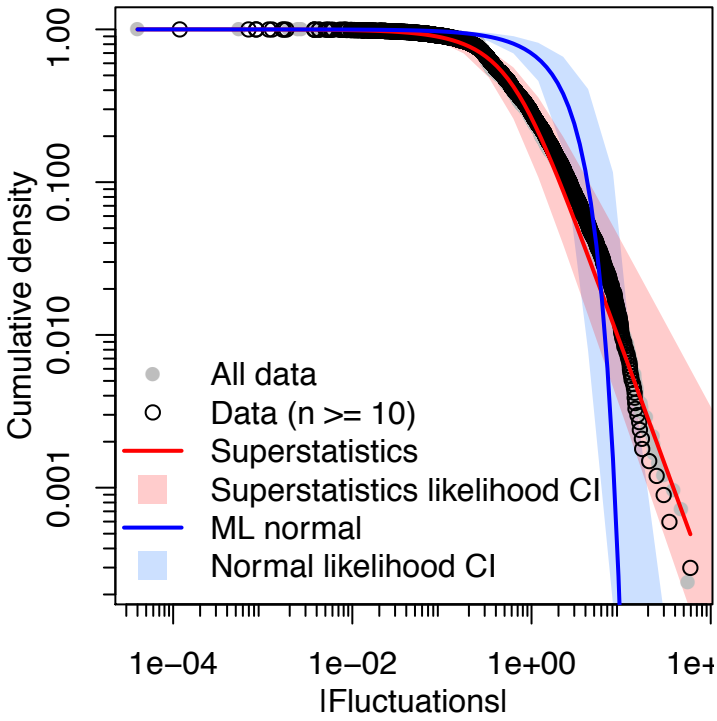

Figure 1.2: Distribution of fluctuations in genus diversity within orders of marine invertebrates in the Paleobiology Database [5] after bias correction. The distribution is fat-tailed as compared to the maximum likelihood estimate of the normal distribution (blue line). At the order level the empirical distribution of fluctuations are well described by our super-statistical approach, both when computed from integrating over the distribution of observed variances (red line) and when fit via maximum likelihood (95% confidence interval; red shading).

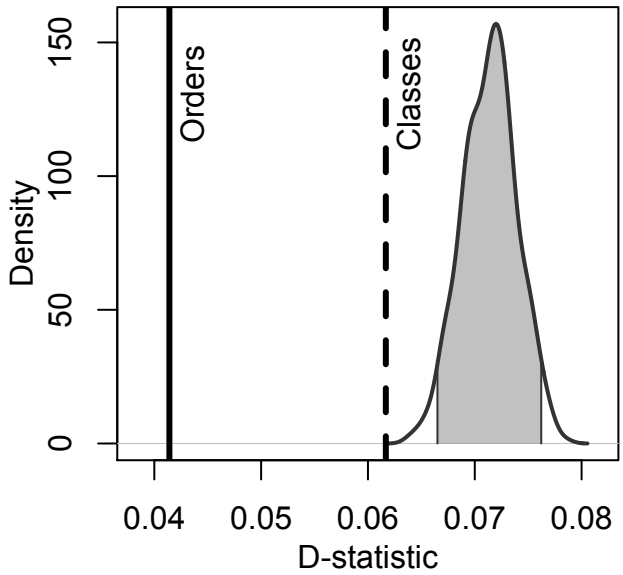

Figure 1.3: Distribution of Kolmogorov-Smirnov (KS) statistics from randomly permuting genera within orders (gray shading represents 95% confidence interval). Solid black line is observed KS statistic at the order level, while the dashed black line shows the observed KS statistic at the class level.

Chapter 2

Community assembly on isolated islands: macroecology meets evolution

2.1 Introduction

Current biodiversity is a product of speciation, extinction and dispersal, contingent on the ecological interactions of organisms with their biotic and abiotic environment. The evolutionary history leading to the assembly of any given ecological community must in some way shape current ecological assemblages. However, because the processes of evolution and ecology occur on different temporal and spatial scales, disentangling the relative influence of local ecological mechanisms from historical evolutionary processes on patterns of community structure remains a central challenge [1].

The evolutionary processes of speciation and extinction are classically viewed as constraints on regional species pools, occurring in a manner largely removed from local ecology [2–4]. Conversely, ecological mechanisms tend to be viewed as packing standing diversity into local communities through consumption, competition, facilitation and, more recently, neutral ecological drift [2, 5–7]. While recent theoretical advances have provided greater insight into ecological drift [2, 8], niche partitioning [5], competition, predation [7] and species interaction networks [9, 10], these insights typically do not contain realistic evolutionary assumptions [11] or ignore them entirely.

Insights into the genetic, biogeographic and selective mechanisms leading to diversification have also emerged based on inference from current patterns of species, genetic or phylogenetic diversity [e.g. 4, 12]. However, it is not possible to use current static patterns to infer the temporal dynamics of either the evolutionary mechanisms or their ecological consequences, nor can we understand what constitutes meaningful change in a system without a baseline for comparison. Here we show how testing idealized ecological theories—such as the unified neutral theory [2] or the maximum entropy theory of ecology [13]—on archipelagos composed of islands formed in a discrete geological sequence can help identify the shifting balance and feedback between fast-acting, local ecological mechanisms, and longterm,

large-scale evolutionary processes in determining ecological community structure. Islands having different ages of formation, along with discrete volcanoes within islands, provide the opportunity to study diversification of species and the assembly of communities in different stages. Ecological theory provides an idealized null baseline against which to compare observed patterns.

2.1.1 Hotspot oceanic archipelagos as model systems

Hotspot oceanic islands are opportune model systems for studying the interplay of local ecological mechanisms and the evolutionary drivers of biodiversity patterns. Due to their sequential formation as the tectonic plate moves over a volcanic hotspot, such island systems offer a range of spatial and temporal scales over which to analyze the outcomes of ecological and evolutionary processes [14]. While many archipelagos around the world share these biotic and geological properties, the Hawaiian archipelago provides a particularly useful system for study because its linear geological chronology [15], ecosystem developmental trajectories [16] and phylogeographic patterns of biodiversity are each well characterized [17]. Moreover, studies of species diversity across the islands have revealed patterns that are non-uniform across the island chronosequence with marked differences among lineages [e.g. 18, 19] that can be used to test for biologically meaningful differences among lineages that might drive their disparate diversification patterns.

2.1.2 Development of genetic structure

High levels of dispersal and associated gene flow among localities limit the extent to which populations can diverge genetically. However, when gene flow is low, distinct populations in different localities are free to diverge through local selective pressures and drift, which can lead to diversification [20] Thus, the magnitude of genetic connectivity among populations provides a measure of the relative importance of dispersal-driven assembly (dictated by processes removed from the local setting) in contrast to assembly by local (*in situ*) diversification in determining community composition. Using the chronosequence of the Hawaiian archipelago, we can analyze populations from multiple sets of taxa across trophic guilds occurring in geological contexts from young to old. We predict that dispersal-driven (ecological) processes will dominate in community assembly in young habitats, with the importance of in situ (evolutionary) processes increasing with habitat age. If evolutionary processes are not important, we predict that communities should reach a statistical steady state through ecological processes alone [13]. If, as we expect, evolutionary processes become increasingly important in community assembly over time, we would expect to find associated deviations from an ecological null model of community assembly, provided by idealized ecological theory. Differences in population structure among taxa or trophic groups could indicate whether sufficient time has passed along the chronosequence for the group of interest to experience significant evolutionary pressures.

2.1.3 Macroecological metrics and idealized ecological theory

By their nature, unified theories of biodiversity [e.g. 2, 13] provide a simplified view of ecology, but deviations from theory can provide insights into which particular ecological patterns require additional biological mechanisms for their explanation [13]. The maximum entropy theory of ecology [METE; 13] in particular provides predictions of species abundance distributions, speciesarea relationships and metabolic rate and network linkage distributions for idealized ecological communities in which the behavior of a system is governed by a simple set of state variables. The principle of maximum information entropy (MaxEnt), from which the METE is derived, is an established inference procedure that has yielded accurate predictions of diverse patterns in fields as varied as thermodynamics (Jaynes, 1957), economics [21], forensics [22], imaging technologies [23] and, more recently, ecology [e.g. 13, 24, 25]. MaxEnt works by seeking the least-biased prediction of a distribution of interest (e.g. the distribution molecular velocities in the case of thermodynamics or of species abundances in the case of ecology) while constraining that prediction to be consistent with state variables describing the macroscopic attributes of the system (e.g. temperature or the total number of species and individuals). These are the most ignorant possible predictions about the system. Thus, studying the unique ecological conditions and evolutionary histories of real-world systems that deviate from the conditions predicted from maximizing information entropy can provide insights into the processes driving ecological systems away from the statistical steady state [13].

Ecological networks are complex systems forming hierarchical structures to which the principle of MaxEnt has recently been applied [13, 26] and are a prime study focus because networks of interacting species embody both the ecology of trophic links and evolutionary processes such as co-evolution [6, 27–29]. Thus they present an opportune starting place to study ecological and evolutionary feedbacks. The distribution of linkages in ecological networks can test whether plantanimal interaction networks assemble neutrally or through deterministic processes such as co-evolution of traits involved in foraging [30]. Analysis of other network metrics such as modularity (the degree to which species interact in semi-autonomous modules) and nestedness (the degree of asymmetry in interaction between specialists and generalists) can further illuminate the underlying eco-evolutionary processes driving patterns of species interactions [6, 27, 28]. In nested networks, species with fewer interactions (i.e. more specialized species) will interact with a subset of the species with which generalists interact. In this way interaction nestedness is mathematically equivalent to island nestedness (in which islands that are less species rich are subsets of islands that are more species rich). However, we only consider network nestedness here.

To gain insights into community assembly as it happens, we propose an integrative framework that harnesses advances in both evolutionary and ecological theory, placed in the context of age-structured archipelagos. Mechanistically simplified ecological theories such as the METE [13] can be used as powerful null models; deviations from theoretical expectations can flag biological phenomena that warrant further study. Here we demonstrate how community-level data from age-structured island systems, combined with population genetic

and phylogenetic data, can test the extent to which the evolutionary histories behind such communities drive their deviation from theoretical expectations. We provide an initial test of this concept using a synthesis of published data on arthropod lineages in the Hawaiian islands. We provide metrics of ecological and evolutionary dynamics across communities from settings that range in geological age from 500 years to 5 Ma. We estimate taxon-specific timelines for the development of population genetic structure for both herbivores and predators and couple these results with macroecological measures of community structure, using predictions from statistical steady-state and ecological network theory to provide insights into changes in community structure over the extended timeframe provided by the island chronosequence.

2.2 Methods

2.2.1 Dispersal-driven processes to in situ differentiation across the island chronosequence

To evaluate the balance between regional immigration and the potential for local differentiation, we measured how molecular variation is partitioned among populations within species across locations of known substrate age on the islands of Hawaii and Maui (Fig. 2.1). We compiled published [DNA sequences, amplified fragment length polymorphism (AFLPs) and allozymes] and new data sets for a diversity of native Hawaiian arthropod groups that represent a spectrum of trophic levels (Table1). New sequences were included for sap-feeding Hemiptera group *Nesosydne* planthoppers [COI; data generated following the protocols in Goodman *et al.* [31]; GenBank accession numbers: KT023113–KT023179] and *Trioza psyllids* [COI, cytB; data generated following protocols in Percy [32]; GenBank accession numbers: KR108061KR108144]. Samples were from the focal sites described below for the ecological analysis, as well as from other locations across Hawaii and Maui. These data were used to provide an estimate of how arthropod populations have accumulated genetic population structure within the focal sites of different geological age.

We used analysis of molecular variance (AMOVA) to examine how genetic variation is partitioned at two scales of population structure: among sites within volcanoes and among volcanoes on both the island of Hawaii and the islands of the Maui Nui complex (Maui, Molokai, Lanai). All analyses of allozyme and DNA sequence data were performed in `Arlequin` v.3.5 [33] using the AMOVA procedure to compute F_{ST}, a measure of genetic variance, or, where possible, Φ_{ST}, an F_{ST} analogue that incorporates genetic sequence information. The *Laupala* AFLP data were analyzed using `tfpga` v.1.3 [34], using the same hierarchical approach of comparing within and among volcanoes as described above. To provide a temporal framework for the population differentiation analysis we assembled divergence-dating information from the literature for as many of the taxa as possible.

To explicitly test the association between landscape age and the potential for in situ genetic divergence we analyzed how within-site F_{ST} varies with the geological age of volcanoes

on the islands of Hawaii and Maui Nui. For each volcano we calculated F_{ST} or Φ_{ST} [33] for each taxon among sites within volcanoes. This analysis assumes that volcano age parallels habitat age, allowing more or less time for the presence of the populations.

2.2.2 Ecological metrics across the island chronosequence

To investigate how ecological patterns change as communities age, we selected four focal sites across the chronosequence and island ages (two on the island of Hawaii, one on Maui and one on Kauai; Fig. 2.1) of approximately 12 km^2 (each was defined as a point with a 2 km radius buffer). Focal sites were selected to have similar forest composition (dominated by *Metrosideros polymorpha*; Myrtaceae), elevation (1100–1400 m) and rainfall (mean annual precipitation 2000–3000 mm). We then constructed bipartite interaction networks between native herbivorous Hemiptera species and native plants at each of the study sites. Bipartite networks describe the topology of ecological interactions between two guilds of organisms (e.g. herbivores and their plant hosts). Quantitative information on the relative importance of interaction links can be incorporated into network analyses [35]. However, currently available data are restricted to binary networks: those that describe the potential for interaction between any two species but not the relative frequency of that interaction to each species.

We compiled species lists of all native herbivorous Hemiptera for each focal site from published species accounts. Species accounts and other published sources were used to determine the presence, probable presence, or probable absence of each species at each of our four focal sites. A documented presence was defined as a known specimen collected at the focal site; a probable presence was defined as a species whose abiotic tolerances and known geographic range overlap with a focal site but no known specimen exists confirming its presence. Probable absence was assumed when the criteria for presence or for probable presence are not met. Two sets of species lists for each focal site were compiled: a conservative data set composed of only documented presence occurrences and a less conservative data set that also included probable presences.

Host plants for each species of Hemiptera were determined from published species accounts. Data on host plant use at each specific site were not available so we assumed that if a known host plant were present at a site it would eventually be used. Host plant occurrence in the focal sites was determined using distribution models for 1158 species of Hawaiian plants [36]. Each focal site was spatially joined in a geographic information system with all coincident plant distribution models that fell within its boundaries. Two sets of resulting focal site-specific networks were constructed: one using the conservative data set of Hemiptera species presences and the other using the less conservative data set.

We hypothesized that potentially complex evolutionary feedbacks contributing to community assembly should result in departures from the predicted ecological statistical steady state. We used the METE [13, 26] to compute the statistical steady state for the distribution of the number of host plants used by each Hemiptera species (hereafter referred to as degree distribution). To evaluate how well the METE predicts the data we simulated METE-conforming communities having the same number of species and links as observed.

We then calculated the log-likelihood of each simulated data set and compared the resultant distribution of log-likelihoods under the hypothesis that the METE is true with the observed log-likelihood. This comparison is identical in approach to a z-score test using a Monte Carlo simulation to estimate the sampling distribution of log-likelihoods. `R` scripts [v.3.1.1; 37] used for METE estimation and Monte Carlo methods are available in Appendix B.2. To investigate how speciation may in part drive network patterns and deviations from those predicted by idealized ecological theory, we analyzed the number of links assigned to each Hemiptera species (the degree distribution) separately for single-island endemics (those species found on only one island and thus probably derived from in situ diversification) versus multi-island endemics (those species found on multiple islands). Although multiple processes can lead to a species being a single-island endemic [38], such taxa provide a proxy for how much speciation occurs within islands. To compare species degree distributions between single-island endemics and multi-island endemics across sites of different ages we conducted a generalized linear model with binomial error, treating site identity as a categorical predictor. Binomial errors effectively account for network size due to the bounded support of the binomial distribution.

To understand how other network properties change with age of the ecosystem substrate, we calculated two widely used descriptive network metrics across sites–nestedness and modularity. Nestedness describes the degree of asymmetry of species interactions connecting specialists and generalists [6, 39]. We calculated nestedness using the NODF metric [40] as implemented in the `R` package `vegan` [41] and modularity using a variety of algorithms implemented in the `R` package `igraph` [42]. These metrics are not directly comparable across networks of different size and connectance [39], so for each metric in each network we calculate z-scores using a null model that randomizes network structure while maintaining certain aggregate network properties (Ulrich et al., 2009). These z-scores are calculated as the difference between the observed network metric minus the mean of the null model divided by the null model standard deviation, or $(x_{obs}\bar{x}_{sim})/SD_{sim}$. Because z-scores can be highly sensitive to the choice of null model [39] we implemented both a probabilistic null model [6] and a null model that strictly constrains the degree distributions of plants and herbivores [39]. The probabilistic null uses the frequency of interactions as the probability that a randomized link gets assigned to that cell in the interaction matrix [6]; thus the probabilistic null constrains row and column sums in probability but not absolutely.

2.3 Results

2.3.1 Dispersal-driven processes to in situ differentiation across the island chronosequence

The AMOVA revealed significant genetic population structure from the smallest to the largest spatial scales examined, all within a very recent timeframe. For mitochondrial loci, statistically significant molecular variation partitioned among sites within volcanoes ranged from

0.037 to 0.92 and among volcanoes from 0 to 0.30. Corresponding variation at multilocus nuclear loci among sites within volcanoes ranged from 0.21 to 0.58 and among volcanoes from 0.04 to 0.34. Taxa in the lower trophic levels (herbivorous sap-feeding Hemiptera: planthoppers and psyllids) had as much or more molecular variation partitioned at the among-site, within-volcano level than the among-volcano level, while the predatory spiders were less structured at localities within volcanoes compared with among them (Table 1). The analysis of genetic population structure across the chronosequence of localities revealed a similar pattern. The herbivores show high genetic population structure among localities even on young volcanoes (Fig. 2.2). By contrast, predatory spiders exhibited little genetic population structure within sites on the same volcano; this was higher among volcanoes, with values increasing with age across the chronosequence.

The observed levels of genetic divergence have evolved rapidly in many cases. For example, for species from the island of Hawaii for which phylogenetic data provide divergence times, estimates of dates of species divergence range from 0.5–4 Ma, with additional within-species genetic divergence having developed subsequently (Table 1). That some of these estimates are older than the known age of the Big Island suggests that genetic divergence pre-dates their colonization to Hawaii, or alternatively that estimates include sampling error. For the one species where population genetic data were used to estimate divergence times between populations, herbivorous *Nesosydne* planthoppers, it was determined that populations diverged as little as 2600 years ago [31, Table 1].

2.3.2 Ecological metrics across the island chronosequence

The degree distribution of Hemiptera species varied across the chronosequence with both the youngest and oldest sites deviating most from the statistical steady-state maximum entropy predictions (Fig. 2.3). In the intermediate-aged site of Kohala, deviations are not significantly different from the predictions of maximum entropy. The generalized linear model revealed significant differences between the degree distributions of single-island endemics (species whose distributions are restricted to only one island) versus archipelagic endemics that are found across multiple islands (Fig. 2.3). Single-island endemics show significantly lower degree distributions overall (i.e. more specialization) compared with more generalist species found across multiple islands. Furthermore, single-island endemics use more host plant species on the intermediate-aged Maui site. The slightly younger Kohala shows increased generalization for both single-island endemics and archipelago endemics. However, when considering the degree distribution defined by trophic links to plant genera instead of plant species, the pattern of increased generalization holds for Kohala, but endemics on Maui no longer show a difference in their degree distributions from other island endemics. This change in pattern suggests that increased generality of Maui endemics may be driven by increased plant species diversity within genera on that island.

Network nestedness decreased with habitat age while modularity increased (Fig. 2.4). This trend was recovered in networks constructed from both more and less stringent geographic criteria (Fig. B.2). Choice of null model changed the magnitude of modularity and

the sign of nestedness z-scores; however, the relative pattern of decreasing nestedness and increasing modularity remained across the different null models used to standardize network metrics (Fig. B.1). The patterns were also robust to sampling intensity, as demonstrated by a rarefaction analysis (Fig. B.3).

2.4 Discussion

2.4.1 Development of genetic population structure at different trophic levels

The analysis of available genetic data presented here indicates that divergence is occurring within the islands at small spatial scales and over short time periods (Table 2.1, Fig. 2.2). Furthermore, the scale of population structure varies with trophic position, with structure developing in sap-feeding herbivore lineages at smaller scales (and hence shorter timeframes in the context of the chronosequence) compared with detritivorous crickets and predatory spiders (Table 2.1, Fig. 2.2). Structure within species may allow populations to take independent evolutionary trajectories, especially when aided by other evolutionary processes acting differentially across species geographic ranges. A variety of factors have been associated with the genetic divergence of populations and species in the lineages described here, including combinations of genetic drift associated with geographic isolation [31, 32, 43, 44], adaptation associated with competition, predation and mutualism [45–47] and sexual signaling [43, 48–50].

The *Nesosydne* planthoppers provide evidence that some period of geographic isolation preceded the divergence of sexual signals [31, 50]. Shifts in plant host use are also associated with diversification in this group [46]. In a phylogenetic study of a radiation of sap-feeding *Nesophrosyne* (Cicadellidae) leafhoppers, species divergence was associated with host plant specialization between 1 and 5 Ma, but only with geography on the younger island [51]. Our network analysis shows that specialization and modularity are more pronounced on Maui than on Hawaii (Figs. 2.3 and 2.4), consistent with the phylogenetic results from *Nesophrosyne*. Available dating analyses of other arthropod taxa indicate that population genetic structure can develop in much less than 1 Myr (Table 1), and suggest that landscape fragmentation processes (e.g. lava flows) may dominate the earliest stages of diversification across taxa in the Hawaiian islands. Other taxa at low trophic levels, such as the herbivorous *Trioza psyllids*, detritivorous *Laupala* crickets and fungivorous Drosophila, show similar signals of geographic isolation combined with ecological and sexual processes driving genetic divergence and diversification across sites as young as those on Hawaii [32, 43, 44, 48, 49]. By contrast, spiders, which are predatory, develop genetic discontinuities at larger spatial and temporal scales with a strong signature of increasing structure with age of the chronosequence [52, Table 1]. Further work is needed to assess the generality of this pattern of slower genetic differentiation in predators compared with herbivores.

2.4.2 Macroecological metrics: network structure and steady state

Across the Hawaiian archipelago, nestedness appears to decrease generally with site age, and is highest on the geologically youngest volcano, Kilauea. High nestedness on Kilauea may arise with high immigration of new species with high probabilities to eat or be eaten by the generalist species already present at the site [6]. However, despite high nestedness on Kilauea, and thus the potential for neutral colonization-driven assembly, this site did not conform to the statistical steady-state predication of the METE. The observed deviations from the METE at Kilauea appear to be largely driven by a surplus of singleton links (Fig. 2.3), which may reflect a state of incomplete assembly, possibly by lower species richness of the plant and herbivore biotas. Conversely, at Kohala, at intermediate age (150 ka), observations were not significantly different from the METE predictions. We posit that the reason why theoretical predictions fit Kohala so well is that the site has had sufficient time to undergo ecological succession and thus arrive at a statistical steady state, but is still too young to be affected by ecological specialization and rapid in situ diversification associated with host plants on older islands.

Interestingly, the communities on the older Maui and Kauai sites show strong deviations from the METE expectations (Fig. 2.4). The METE is agnostic about which mechanisms determine the values of the state variables that lead to its macroecological predictions (Harte, 2011). It does not account for the evolutionary history of biological systems. Thus, one possible explanation for the strong deviations from the METE expectations, compared with observations at our intermediate-aged site (Kohala), is that while the ages of Maui and Kauai are sufficient for evolutionary assembly driven by specialization and diversification on host plants, the older age of these islands may have led to range contractions and possibly extinction of plant species on the oldest island of Kauai (Whittaker et al., 2008).

Our results show decreased nestedness and increased modularity on Maui and Kauai. Co-evolution between interacting species should lead to greater modularity [27, 28]. However, the influence of certain network properties, such as nestedness, on stability is still unknown, and so theoretical predictions of how network properties should change over evolutionary time, generally, are lacking. Theoretical and empirical studies have suggested that nestedness may or may not promote stability [53, 54]. Furthermore, almost all studies of food webs have focused primarily on single or short ecological time spans of network development that do not span as much evolutionary time as is included here [e.g. 55]. Food webs are dynamic emergent entities, with broad topological characteristics that may change dramatically over time [e.g. 56]. To our knowledge, our study represents the first to evaluate network topology over larger temporal scales, and we argue that age-structured landscapes such as the Hawaiian archipelago are promising for resolving long-standing debates on the causes and consequences of network properties such as nestedness.

We found that single-island endemics were always more specialized than multiple-island endemics. Although dietary breadth has been positively associated with geographic range size [57], the direction of causality is unclear [58]: while dietary breadth may allow some

species to colonize other islands, it may also be driven by adaptation to exploit locally abundant hosts across a large range. Nevertheless, both scenarios are consistent with the hypothesis that in situ formation of single-island endemics may be the product of co-evolution and specialization. At the Kohala site, which showed the best fit to maximum entropy theory, single-island endemic and multiple-island endemic species alike showed increased generalization (i.e. a higher degree, or more links; Fig 2.3), while at the youngest site of Kilauea, specialist single-island endemics may be limited by low plant diversity and thus appear more specialized (Fig. 2.3). Conversely at the oldest site on Kauai, where plant diversity is high [59], single-island endemics are again associated with decreased degree and thus genuine specialization (Fig. 2.3). On Maui, single-island endemics show statistically significant increases in generalization, but this pattern disappears when analyzing the data at the resolution of plant genera, thus suggesting that Hemiptera species endemic to Maui may benefit from the diversification of plant species within genera.

2.4.3 Future research

The data and analyses presented here describing insect and plant communities across a chronosequence of habitats in Hawaii generate testable hypotheses concerning the relative importance of ecological and evolutionary processes in community assembly. Our work to date suggests the overarching hypothesis that ecological processes dominate community assembly in younger environments, with evolutionary processes becoming increasingly important as communities age. We can also make predictions about the sequence of community assembly based on proposed mechanisms.

In younger communities we predict characteristics of ecological assembly, with species resembling random samples through immigration from regional source pools. Thus, metrics describing these communities will approach expectations of an ecological statistical steady state. An exception will be communities that are still undergoing the initial stages of primary succession, which will change rapidly through time and represent nonrandom samples of source pools. We also predict that these communities will exhibit a nested network structure, assuming new species will eat or be eaten by the generalist species already present in the community, as suggested by previous work on nestedness [6] and by our finding that widespread species tend to be generalists (Fig. 2.4). Following the same logic, in older communities we expect to see characteristics of evolutionary assembly, dominated by processes such as adaptive exploration of niche space, giving way to speciation. Thus, we predict increasing specialization and modularity with time [6, 27, 28] as reflected by age across the chronosequence.

Ecological data: assembly of species into communities

In order to build a more rigorous understanding of the assembly process in both younger and older communities, fine-grained sampling of all macroscopic arthropod taxa is needed from a large number of sites across the island chronosequence. This will allow an assessment of

changes in overall species composition and diversity across all players in the time-calibrated landscape [18]. Such data will allow us to test entire arthropod communities for deviations from METE predictions of statistical steady state [13] across substrates of different ages. For example, predators, whose assemblages are likely to be more dominated by immigration and ecological assembly (Fig. 2.2), may never show strong deviations from METE predictions, whereas herbivores could show increasing deviation with age in agreement with the network results of this paper (Fig. 2.3).

Evolutionary data: diversification within species

The current study demonstrates that taxa from different trophic guilds differ in the scale at which differentiation occurs and highlights the importance of fragmentation of the landscape in facilitating differentiation. Future work will be aimed at gathering data for additional focal taxa within this system, spanning different trophic levels. We will use these data to understand taxonomic and functional differences in the rate of differentiation, to assess the roles of genetic fusion and fission and the spatial scale over which they are important in fostering diversification [60], and to detail the relative rates of speciation and extinction across the island chronosequence.

2.5 Conclusions

We have shown how a chronosequence can be used to understand biodiversity dynamics across an ecologicalevolutionary continuum. Focusing on entire communities of arthropods in the Hawaiian islands allows us to incorporate predictions from idealized ecological theories to understand eco-evolutionary feedbacks and generate predictions about how entire communities develop over an extended time. Such an approach may prove fruitful for investigating the separate and interactive roles of ecological and evolutionary drivers of community assembly using age-structured systems as a simplified natural experiment, as exemplified by oceanic archipelagos.

We have demonstrated how taxa in the lower trophic levels developed genetic structure even in the youngest habitats of the observed chronosequence and at smaller spatial scales (Table 2.1, Fig. 2.2). Thus, lower trophic levels are affected by in situ processes of diversification very early in the chronosequence, compared with higher trophic levels, though in situ processes become more important over time in the latter. Network nestedness decreased while modularity increased with age (Fig. 2.4), again indicating a possible shift from assembly driven by ex situ immigration early on to one based on in situ diversification, such as in co-diversification of insect herbivores with host plants [6, 27]. That single-island endemics (probably the product of in situ diversification) show more specialization at older sites than more broadly distributed species (those taxa more likely to be initial colonists; Fig. 2.3) also supports this hypothesis.

This study provides a framework for using chronologically arranged oceanic island systems to examine the interplay between evolutionary and ecological processes in shaping biodiversity. Our initial results provide a clear hypothesis that ecological processes dominate community assembly in younger environments, with evolutionary processes becoming more important as communities age. We demonstrate how this approach can provide insights into the development of communities over ecologicalevolutionary time, and the dynamic feedbacks involved in assembly.

Acknowledgements

We are indebted to many scientists and land managers in Hawaii who have provided access to the lands: Pat Bily (The Nature Conservancy of Hawaii), Melissa Dean, Christian Giardina, and Tabetha Block (Hawaii Experimental Tropical Forests), Betsy Gagne (Natural Area Reserve System), Lisa Hadway and Joey Mello (Department of Forestry and Wildlife Hilo), Cynthia King and Charmian Dang (Department of Land and Natural Resources) and Rhonda Loh (Hawaii Volcanoes National Park). We thank Robert Ricklefs and two anonymous referees for thoughtful commentary. We are very grateful to Guida Santos, Richard Field and Robert Ricklefs for inviting us to contribute to this Special Issue. The research was supported by the National Science Foundation DEB 1241253.

References

1. Ricklefs, R. E. A comprehensive framework for global patterns in biodiversity. *Ecology Letters* **7,** 1–15 (2004).
2. Hubbell, S. P. *The unified neutral theory of biodiversity and biogeography (MPB-32)* (Princeton University Press, 2001).
3. Cavender-Bares, J., Kozak, K. H., Fine, P. V. & Kembel, S. W. The merging of community ecology and phylogenetic biology. *Ecology letters* **12,** 693–715 (2009).
4. Wiens, J. J., Pyron, R. A. & Moen, D. S. Phylogenetic origins of local-scale diversity patterns and the causes of Amazonian megadiversity. *Ecology Letters* **14,** 643–652 (2011).
5. Tilman, D. Niche tradeoffs, neutrality, and community structure: a stochastic theory of resource competition, invasion, and community assembly. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 10854–10861 (2004).
6. Bascompte, J. & Jordano, P. Plant-animal mutualistic networks: the architecture of biodiversity. *Annu. Rev. Ecol. Evol. Syst.* **38,** 567–93 (2007).
7. Borer, E. T. *et al.* Herbivores and nutrients control grassland plant diversity via light limitation. *Nature* (2014).

8. Rosindell, J. & Phillimore, A. B. A unified model of island biogeography sheds light on the zone of radiation. *Ecology Letters* **14,** 552–560 (2011).

9. Williams, R. J. & Martinez, N. D. Simple rules yield complex food webs. *Nature* **404,** 180–183 (2000).

10. Brose, U., Williams, R. J. & Martinez, N. D. Allometric scaling enhances stability in complex food webs. *Ecology Letters* **9,** 1228–1236 (2006).

11. Ricklefs, R. E. The unified neutral theory of biodiversity: do the numbers add up? *Ecology* **87,** 1424–1431 (2006).

12. Jetz, W., Thomas, G., Joy, J., Hartmann, K. & Mooers, A. The global diversity of birds in space and time. *Nature* **491,** 444–448 (2012).

13. Harte, J. *The Maximum Entropy Theory of Ecology* (Oxford University Press, 2011).

14. Warren, B. H. *et al.* Islands as model systems in ecology and evolution: prospects fifty years after MacArthur-Wilson. *Ecology Letters* **18,** 200–217 (2015).

15. Price, J. & Clague, D. How old is the Hawaiian biota? Geology and phylogeny suggest recent divergence. *Proc R Soc Lond B Biol Sci* **269,** 2429–2435 (2002).

16. Vitousek, P. M. *Nutrient cycling and limitation: Hawai'i as a model system* (Princeton University Press, 2004).

17. Wagner, W. & Funk, V. *Hawaiian Biogeography Evolution on a Hot Spot Archipelago* (Smithsonian Institution Press, Washington, DC, 1995).

18. Gruner, D. S. Geological age, ecosystem development, and local resource constraints on arthropod community structure in the Hawaiian Islands. *Biological Journal of the Linnean Society* **90,** 551–570 (2007).

19. Gillespie, R. in *Guide to Ecology* (eds Levin, S. *et al.*) 143–152 (Princeton University Press, Princeton, NJ, 2009).

20. Slatkin, M. Gene flow and the geographic structure of natural populations. *Science* **236,** 787–792 (1987).

21. Golan, A., Judge, G. & Miller, D. Maximum entropy econometrics: Robust estimation with limited data (1996).

22. Roussev, V. in *Advances in digital forensics vi* 207–226 (Springer, 2010).

23. Gull, S. & Newton, T. Maximum entropy tomography. *Applied optics* **25,** 156–160 (1986).

24. Phillips, S. J., Anderson, R. P. & Schapire, R. E. Maximum entropy modeling of species geographic distributions. *Ecological modelling* **190,** 231–259 (2006).

25. Dewar, R. C. & Porté, A. Statistical mechanics unifies different ecological patterns. *Journal of Theoretical Biology* **251,** 389–403 (2008).

26. Williams, R. J. Simple MaxEnt models explain food web degree distributions. *Theoretical Ecology* **3,** 45–52 (2010).

27. Donatti, C. I. *et al.* Analysis of a hyper-diverse seed dispersal network: modularity and underlying mechanisms. *Ecol. Lett.* **14,** 773–781 (2011).

28. Nuismer, S. L., Jordano, P. & Bascompte, J. Coevolution and the architecture of mutualistic networks. *Evolution* **67,** 338–354 (2013).

29. Thompson, J. N., Schwind, C., Guimarães, P. R. & Friberg, M. Diversification through multitrait evolution in a coevolving interaction. *Proceedings of the National Academy of Sciences* **110,** 11487–11492 (2013).

30. Vázquez, D., Poulin, R., Krasnov, B. & Shenbrot, G. Species abundance and the distribution of specialization in host-parasite interaction networks. *Journal of Animal Ecology* **74,** 946–955 (2005).

31. Goodman, K., Welter, S. & Roderick, G. Genetic divergence is decoupled from ecological diversification in the Hawaiian Nesosydne planthoppers. *Evolution* **66,** 2798–2814 (2012).

32. Percy, D. Radiation, diversity and host plant interactions among island and continental legume-feeding psyllids. *Evolution* **57,** 2540–2556 (2003).

33. Excoffier, L. & Lischer, H. E. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources* **10,** 564–567 (2010).

34. Miller, M. Tools for Population Genetic Analysis (TFPGA), 1.3: a Windows program for the analysis of allozyme and molecular population genetic data. Distributed by the author: http://www.marksgeneticsoftware.net/tfpga.htm. (1997).

35. Vázquez, D. P., Blüthgen, N., Cagnolo, L. & Chacoff, N. P. Uniting pattern and process in plant-animal mutualistic networks: a review. *Annals of Botany* **103,** 1445–1457 (2009).

36. Price, J. P. *Mapping plant species ranges in the Hawaiian Islands: developing a methodology and associated GIS layers* (US Department of the Interior, US Geological Survey, 2012).

37. R Development Core, T. R: A language and environment for statistical computing. *Vienna, Austria* (2013).

38. Whittaker, R. J., Triantis, K. A. & Ladle, R. J. A general dynamic theory of oceanic island biogeography. *Journal of Biogeography* **35,** 977–994 (2008).

39. Ulrich, W., Almeida-Neto, M. & Gotelli, N. J. A consumer's guide to nestedness analysis. *Oikos* **118,** 3–17 (2009).

40. Almeida-Neto, M., Gumarães, P., Gumarães, P., Loyola, R. & Ulrich, W. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos* **117,** 1227–1239 (2008).

41. Oksanen, J. *et al. vegan: Community Ecology Package* (2013). <http://CRAN.R-project.org/package=vegan>.

42. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **Complex Systems,** 1695 (2006).

43. Mendelson, T. C. & Shaw, K. L. Sexual behaviour: rapid speciation in an arthropod. *Nature* **433,** 375–376 (2005).

44. O'Grady, P. M. *et al.* Phylogenetic and ecological relationships of the Hawaiian Drosophila inferred by mitochondrial DNA analysis. *Molecular phylogenetics and evolution* **58,** 244–256 (2011).

45. Gillespie, R. Community assembly through adaptive radiation in Hawaiian spiders. *Science* **303,** 356–359 (2004).

46. Roderick, G. K. & Percy, D. in *Specialization, Speciation, and Radiation. Evolutionary Biology of Herbivorous Insects* (ed Tilmon, K.) 151–161 (University of California Press, Berkeley, 2008).

47. Brewer, M. S., Carter, R. A., Croucher, P. J. & Gillespie, R. G. Shifting habitats, morphology, and selective pressures: Developmental polyphenism in an adaptive radiation of Hawaiian spiders. *Evolution* **69,** 162–178 (2015).

48. Percy, D. & Kennedy, M. Psyllid communication: acoustic diversity, mate recognition and phylogenetic signal. *Invertebrate Systematics* **20,** 431–445 (2006).

49. Magnacca, K. N., Foote, D. & O'Grady, P. M. A review of the endemic Hawaiian Drosophilidae and their host plants. *Zootaxa* **1728,** 1–58 (2008).

50. Goodman, K., Kelley, J., Welter, S., Roderick, G. & Elias, D. Rapid diversification of sexual signals in Hawaiian Nesosydne planthoppers (Hemiptera: Delphacidae): the relative role of neutral and selective forces. *Journal of evolutionary biology* **28,** 415–427 (2015).

51. Bennett, G. M. & O'Grady, P. M. Historical biogeography and ecological opportunity in the adaptive radiation of native Hawaiian leafhoppers (Cicadellidae: Nesophrosyne). *Journal of Biogeography* **40,** 1512–1523 (2013).

52. Roderick, G., Croucher, P., Vandergast, A. & Gillespie, R. Species differentiation on a dynamic landscape: shifts in metapopulation and genetic structure using the chronology of the Hawaiian Archipelago. *Evolutionary Biology* **32,** 192–206 (2012).

53. Allesina, S. & Tang, S. Stability criteria for complex ecosystems. *Nature* **483,** 205–208 (2012).

54. Suweis, S., Simini, F., Banavar, J. R. & Maritan, A. Emergence of structural and dynamical properties of ecological mutualistic networks. *Nature* **500,** 449–452 (2014).

55. Albrecht, M., Riesen, M. & Schmid, B. Plant–pollinator network assembly along the chronosequence of a glacier foreland. *Oikos* **119,** 1610–1624 (2010).

56. Yeakel, J. D., Guimarães, P. R., Bocherens, H. & Koch, P. L. The impact of climate change on the structure of Pleistocene food webs across the mammoth steppe. *Proceedings of the Royal Society of London B: Biological Sciences* **280,** 20130239 (2013).

57. Lewinsohn, T. M., Novotny, V. & Basset, Y. Insects on plants: diversity of herbivore assemblages revisited. *Annual Review of Ecology, Evolution, and Systematics,* 597–620 (2005).

58. Slatyer, R. A., Hirst, M. & Sexton, J. P. Niche breadth predicts geographical range size: a general ecological pattern. *Ecology letters* **16,** 1104–1114 (2013).

59. Kitayama, K. & Mueller-Dombois, D. Vegetation changes along gradients of long-term soil development in the Hawaiian montane rainforest zone. *Vegetatio* **120,** 1–20 (1995).

60. Gillespie, R. & Roderick, G. Geology and climate drive diversification. *Nature* **509,** 207–298 (2014).

Figures

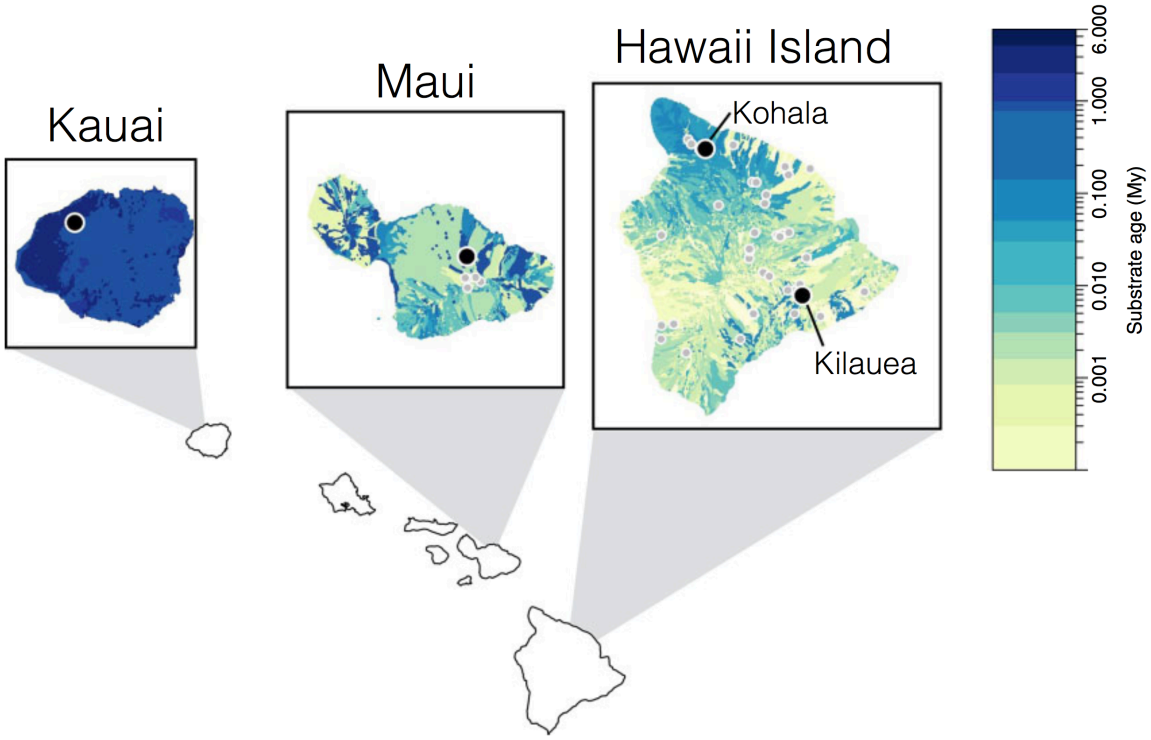

Figure 2.1: Map of substrate age (millions of years, My) of the islands of Kauai, Maui and Hawaii. Colours correspond to substrate age from young (light) to old (dark). Focal sites are shown as black circles (on Hawaii, Kohala is in the north, Kilauea in the south) while sampling sites for genetic data are represented by grey circles.

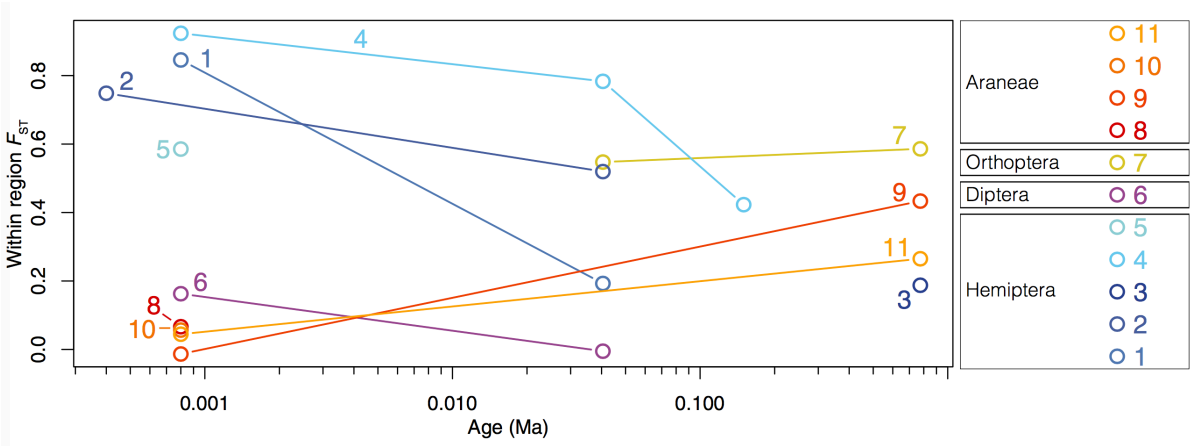

Figure 2.2: Population genetic structure (Φ_{ST} for all taxa except *Laupala* for which we used F_{ST}) among sites within volcanoes with volcano age for insects and spiders. Calculations were based on mitochondrial DNA only (see Table 1 for details). The plant-feeding groups, specifically the sap-feeding Hemiptera, show higher genetic structure among sites on young volcanoes relative to older volcanoes, whereas detritivores (crickets), fungivores (*Drosophila*) and in particular predators (spiders) show little structure on young volcanoes. For spiders, substantial structure develops only later in the chronosequence, for example on Maui at approximately 1 Ma. Numbers refer to different species: 1, *Nesosydne chambersi*; 2, *Nesosydne raillardiae*; 3, *Nesosydne bridwelli*; 4, *Trioza* HB; 5, *Trioza* HC; 6, *Drosophila sproati*; 7, *Laupala cerasina*; 8, *Tetragnatha anuenue*; 9, *Tetragnatha brevignatha*; 10, *Tetragnatha quasimodo*; 11, *Theridion grallator*.

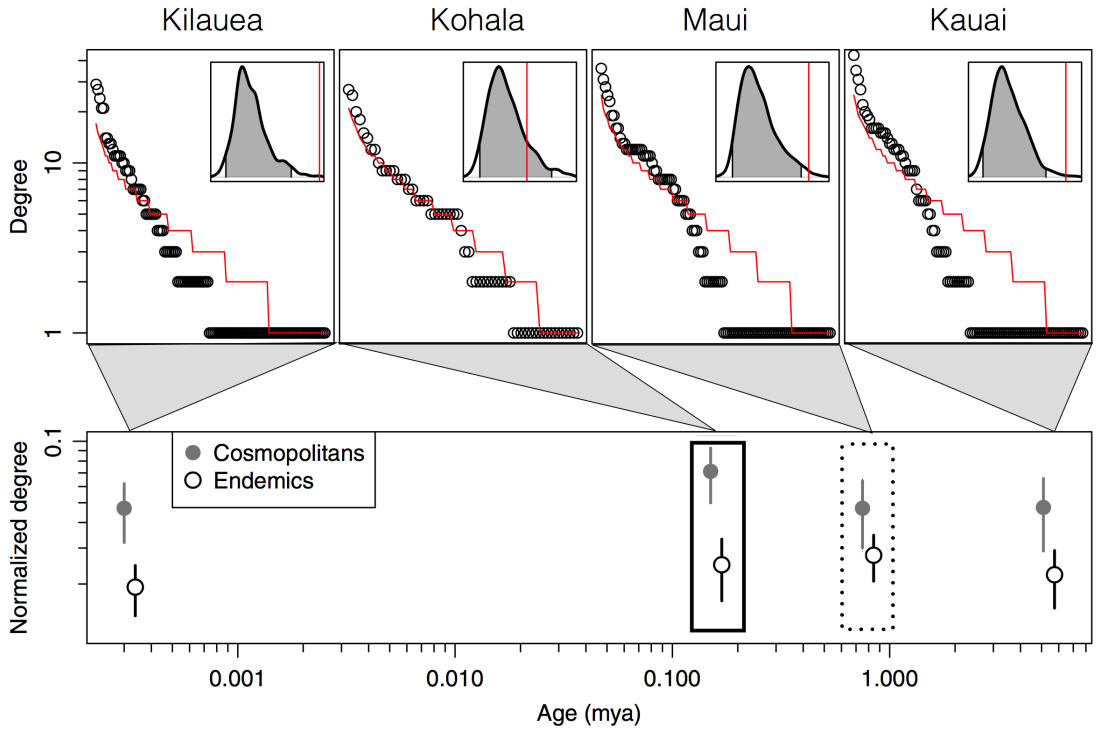

Figure 2.3: Patterns in degree distributions across sites, comparing archipelago-wide endemics (cosmopolitans) with single-island endemic (Endemics) taxa. The top panels show that networks deviate most from the predictions of the maximum entropy theory of ecology on the youngest and oldest sites. Inset figures show the distribution of simulated mean squared errors; if the vertical/red line falls within the grey region (95% confidence interval) the data are not significantly different from the predictions of maximum entropy theory. All sites except Kohala deviate from the predications. The bottom panel shows the number of links for endemics versus cosmopolitans. Endemics show lower linkage overall, but significantly increase on the intermediate-aged site Maui (highlighted with dotted box). Kohala shows increased linkage overall (highlighted with a solid box).

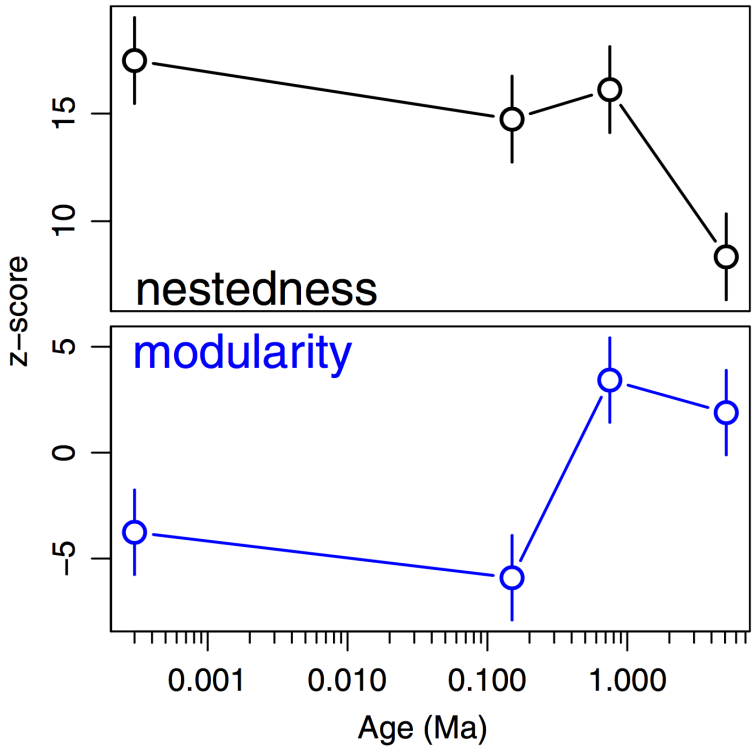

Figure 2.4: Trends in network metric nestedness and modularity through time. Nestedness decreases while modularity increases. Error bars represent 95% confidence intervals from a null model simulation.

Tables

Group	Functional group	Species	Data (no. of individuals: length/no. of markers)	Between volcanoes[1] % variation (*P*-value)	Within volcanoes[1] % variation (P-value)	Timeframe
Planthoppers	Herbivore	*Nesosydne chambersi*[2]	mtDNA: COI (187:653)	0.05***	0.77***	Within-species divergence[11] = 2600 [95% HPD (1.2–35.1) × 10^3], and 20,100 [95% HPD (7.4–135.0) × 10^3] years for two population pairs[2]
			nucDNA: microsatellites (292:13)	0.04***	0.21***	
	Herbivore	*Nesosydne raillardiae*[3]	mtDNA: COI (33:581)	0.26***	0.49***	na
	Herbivore	*Nesosydne bridwelli*[3]	mtDNA: COI (34:677)	na	0.18**	na
Psyllids	Herbivore	*Trioza* HB[3,4]	mtDNA: COI and cytB (29:857)	–0.14***	0.92***	na
	Herbivore	*Trioza* HC[3,4]	mtDNA: COI and cytB (17:857)	0.17**	0.53**	na
Fly	Fungivore	*Drosophila sproati*[5]	mtDNA: COII (232:570)	0.11***	0.81***	Max age[12] = 1.15 (95% HPD 0.75–1.5) Ma[8]
Cricket	Detritivore	*Laupala cerasina*[6]	nucDNA: AFLP (631)	0.30***	0.58***	na
Spiders	Predator	*Tetragnatha anuenue*[7]	mtDNA: COI (162:607)	0.23***	0.041***	Max age[12] = 3.0 (95% HPD 2.5–4.5) Ma[10]
	Predator	*Tetragnatha brevignatha* (Hawaii Island)[7]	mtDNA: COI (54:605)	0.16*	0.00	Max age[12] = 4.0 (95% HPD 3.0–4.75) Ma[10]
	Predator	*Tetragnatha quasimodo*[7]	mtDNA: COI (149:439)	0.09***	0.037***	Max age[12] = 2.5 Ma[10]
			nucDNA: allozymes (46:9)	0.34***	na	
	Predator	*Theridion grallator*[7]	mtDNA: COI (209:1270)	0.30***	0.05***	Node age[13] = 0.56 (95% HPD 0.37–0.75) Ma[9]
			nucDNA: allozymes (224:8)	0.19***	na	
	Predator	*Ariamnes* spp.[7]	mtDNA: COI (8:420)	0.05	na	na

mtDNA, mitochondrial DNA; nucDNA, nuclear DNA; HPD, highest posterior density; na, no information available.
***$P < 0.001$, **$P < 0.05$, *$P < 0.10$.
[1]mtDNA and microsatellites are calculated as Phi-st; AFLPs and allozymes are calculated as F_{ST}.
[2]Goodman *et al.* (2012).
[3]This paper.
[4]These *Trioza* species are in the process of being described; HB and HC are their provisional identifiers (D.M.P., in prep).
[5]Eldon *et al.* (2013).
[6]Mendelson and Shaw (2005).
[7]Roderick *et al.* (2012).
[8]Magnacca and Price (in press).
[9]Croucher *et al.* (2012).
[10]Casquet *et al.* (2015).
[11]Within-species divergence, estimated using Isolation with Migration (IM: see Goodman *et al.* 2012)
[12]Max age = the node age of the phylogenetic split between this species and its sister species, calculated using divergence dating analyses performed in BEAST. In most cases, this will be an overestimate of the node age of the species itself, but is the best information we have at present.
[13]Node age = age of the most recent common ancestor of the monophyletic group on the island of Hawaii, estimated using BEAST.

Table 2.1: Results of the analyses of molecular variance (AMOVA) that partitions molecular genetic variation among volcanoes and among sites within volcanoes for arthropod lineages found within the study sites on the island of Hawaii. Where estimates of divergence through molecular dating are available for the taxa, they are presented to show the timeframe within which this genetic structure has developed.

Chapter 3

`meteR`: An R package for testing the Maximum Entropy Theory of Ecology

Abstract

1. Macroecological patterns appear to follow consistent forms across a range of natural systems, however the origin of their regularity remains obscured. The Maximum Entropy Theory of Ecology (METE) predicts macroecological patterns of abundance, metabolic rates and their distribution within communities and across space using an information theoretic approach. METE's success in predicting empirical patterns demands that we further press the theory's predictions to determine how (or whether) predictability depends on attributes of the system and the temporal, spatial and biological scales at which we study it.

2. METE predicts multiple macroecolgical metrics using statistical idealizations from information theory; thus confronting METE with data represents a strong test of the underlying biological mechanisms that could drive real communities away from statistical idealizations. METE has remained somewhat inaccessible due to its highly mathematical nature and a lack of software for model construction/evaluation. To remedy this, we have developed an `R` package implementation of METE.

3. Our open source (GNU General Public License v2) `R` package, `meteR` (version 1.0; `cran.r-project.org/web/packages/meteR`), (1) directly calculates all of METE's predictions from a variety of data formats; (2) automatically handles approximations and other technical details; (3) provides high-level plotting and model comparison functions to explore and interrogate models.

4. With these tools in hand, ecologists can more readily test the predictions of METE for their data sets. By facilitating tests of METE, we expect that a better understanding of its strengths and limitations will emerge. A better understanding of the strengths and limitations of METE will offer insight into how biological mechanisms and statistical constraints combine to drive macroecological patterns.

3.1 Background on the Maximum Entropy Theory of Ecology

Macroecology [1] seeks to predict patterns in the distribution of individuals within species, across body sizes, and over space. These patterns can vary with spatial, temporal and taxonomic scale which makes their regularities challenging to detect. Macroecological patterns can be described quantitatively in the form of well-defined metrics such as species abundance distributions (SAD), body size distributions (e.g. distribution of metabolic rates [=power] across individuals, or the individual power distributions; IPD), and species-area relationships (SAR). Macroecological theory attempts to predict the mathematical form of these metrics from combinations of explicit mechanisms and statistical assumptions. Harte and colleagues have developed a Maximum Entropy Theory of Ecology (METE) grounded in information theory, which predicts the form of most of the macroecology metrics found in the literature, needing very limited empirical data as input, and with no adjustable parameters [2–4].

Maximizing information entropy (MaxEnt) is a general inference procedure for solving a class of problems involving inference of the least biased mathematical form of a probability distribution (e.g., the species abundance distribution) given some prior knowledge that can be expressed in the form of constraints on that distribution (e.g., the mean abundance across all species). One special case of MaxEnt familiar to many ecologists is its use in machine learning and species distribution modeling [SDM; 5], though there are a variety of applications in ecology [6, 7] and other sciences [e.g. 8–10]. The conceptual goal of the MaxEnt approach in macroecology is to build predictions that are not sensitive to potentially arbitrary model parameter choices and represent a statistical idealization of a community that is near steady state. The solution to the MaxEnt problem entails finding the form of a distribution, $p(n)$ (e.g. the SAD), that maximizes the information entropy function: $I = -\sum_n p(n)log(p(n))$ under the constraints imposed by prior knowledge on $p(n)$ (e.g. mean abundance). Maximization is carried out by the method of Lagrange multipliers. The resulting distribution is the function that is as smooth as possible given the constraints, and thus reflects no information other than the prior constraints [11, 12].

The development of METE parallels the method used to derive the laws of classical equilibrium statistical mechanics and thermodynamics by Jaynes [11], in which constraints arise from knowledge of the state variables of the system: for example, the total energy, volume, and number of molecules in a container of gas. In METE, the state variables used to predict the metrics of macroecology are the area, A_0, of an ecosystem at any spatial scale at which census data exist, the total number of species S_0, censused in that area, the total number of individuals, N_0, across the S_0 species, and the total rate of metabolic energy use, E_0, by the N_0 individuals. With the constraints that arise from the ratios of these state variables, the maximum information entropy condition predicts the mathematical forms of macroecological metrics (Table 3.1). Two distributions are at the core of the theory. The first is a joint distribution (the *Ecosystem Structure Function*; ESF) over abundance, n, and metabolic rate, ϵ: $R(n, \epsilon|S_0, N_0, E_0)$. $R \cdot d\epsilon$ is the probability that a species randomly selected

from the census has abundance n, and an individual randomly selected from that species has metabolic energy requirement in the interval $(\epsilon, \epsilon + d\epsilon)$. The second core distribution is the *Spatial Structure Function* $\Pi(n|A, n_0, A_0)$ (SSF; mirroring the ESF). If a species has n_0 individuals in an area A_0, then $\Pi(n)$ is the probability that it has n individuals in an area A within A_0. From these two core distributions many of the metrics commonly studied in macroecology can be derived: for example the species abundance distribution arrises by integrating $R(n, \epsilon)$ over ϵ and the distribution of metabolic rates (or body size under metabolic scaling theory [13] arrises by summing over n. The species area relationship can be derived by combining the species abundance distribution with Π at multiple scales. These derivations are detailed in Harte [4] and Table 3.1 provides a summary.

3.1.1 Stronger tests of METE

Tests of METE have been successful in a wide range of systems [2–4, 14, 15], however a great deal of further testing is needed to establish METE's generality, strengths and weaknesses. Existing tests have focused on SADs or SARs while tests of metabolic rate distributions are rare [but see 4, 15, 16]. However, the most illuminating cases are perhaps those where METE fails [4], as these will demonstrate what attributes of a community drive it away from the simple steady state into alternate, more complex or transient states. Additionally, macroecological predictions other than the SAD and SAR have received relatively little attention, and their generality and correlates of their successes and failures are unknown. Two categories of failed predictions stand out. The first was originally considered a success: the theory predicts an inverse relationship between metabolic rate (body size) and abundance, called 'energy equivalence' or the Damuth rule [17]. Yet considerable data analysis reveals numerous exceptions to this prediction [18]. Second, and more fundamentally, the theory fails to accurately predict empirical patterns in ecosystems undergoing relatively rapid change [4, 15]. Examples are rapidly diversifying habitats on newly formed islands, or ecosystems recovering from recent disturbance and undergoing relatively rapid succession, as for example in the aftermath of fire. Because the deviation between the patterns predicted by theory and empirical data from disturbed systems appears to itself follow a systematic pattern [Rominger *et al.*, in prep.; 15], clues exist for how to extend the static theory to the dynamic realm.

To promote the broad testing of METE across varied systems we have developed an open source (GNU General Public License v2) `R` package, `meteR` (version 1.0; `cran.r-project.org/web/packages/meteR`) that calculates all of METE's predictions in an object-oriented framework. We envision that `meteR` will be valuable for testing the scale dependence and equilibrium assumptions of METE's predictions. That is, METE is not constrained to work at any specific spatial or temporal resolution, or for any particular taxonomic unit. While METE is typically applied to species coexisting in a single snap shot in time in a single plot, these conventions are imposed by data availability and not any fundamental attributes of the theory. One could readily study genus or family patterns across regional or subplot extents. `meteR` allows rapid generation of many models built at different scales or across different clades to

better examines METE's successes and failures. We encourage in particular exploration of how well METE predictions work across gradients of disturbance, ecosystem age, latitude, elevation, phylogenetic diversity, functional diversity and level of invasion. Tests of theory along these gradients can illuminate how the mechanisms associated with those gradients drive communities away from the statistical idealizations of METE. Such an understanding will help ecologists better predict when METE represents a sufficient model for their system, what ecological processes might be needed to improve predictions, or theoretical scaling relationships among state variables.

Only further comparison to data can address these open questions surrounding METE. To detect generalities, tests must be performed in a wide range of systems, which means that these tests must be performed by a large collection of researchers with different expertise. Achieve this goal motivated out development of an `R` package [19], `meteR`, which makes the theory accessible and provides an efficient workflow for evaluating METE with empirical data.

3.2 `meteR`

Our `R` package, `meteR`, helps to address two key challenges with using METE: it reduces the need for practitioners to understand many of the mathematical details of the theory, and allows users to readily explore predictions without re-deriving maximum entropy solutions by hand. The prerequisite for using `meteR` is an understanding of its predicted macroecological distributions and relationships (Table 3.1), while details of exact solutions, numerical methods and mathematical approximations used to derive them are relegated to (user-accessible) functions "under the hood." Importantly, `meteR` automatically checks that the conditions for all approximations used in METE predictions are met, and if not, implements more computationally costly exact solutions. Thus `meteR` provides not only the quickest approach to testing METE because of its mathematical abstraction, but also because it is operationally optimized. We contrast `meteR` with other software resources for METE, which either include fewer predictive features [20] or are exclusively low-level (e.g. code available at `github.com/weecology/METE`) and can only be used by those already well-versed in scientific computation. Additionally `meteR` is the only `R` resource available for METE, notable due to the accessibility and popularity of `R`, as opposed to other programming languages, among ecologists. Furthermore we have unit tested `meteR` using alredy published and verified solutions [4, 15] (results from these tests can be found at `github.com/cmerow/meteR/tree/master/tests/testthat`), meaning that users can confidently proceed with the results produced by our package. Additionally, bugs can be reported via GitHub's issue tracking feature (`github.com/cmerow/meteR/issues`).

`meteR` also simplifies model comparison and visualization. Fitted macroecological distributions can be readily used to predict, simulate, or compare to data for use in further analyses. These operations are performed by familiar `R` functions; for example each fitted distribution contains a `d` element which gives the function of the probability distribution for

use in visualization and an `r` element for random number generation (in analogy to `dnorm` and `rnorm` for a Gaussian distribution). Likelihoods can be computed using the standard `logLik` function and plotting carried out using `plot`. By reducing the effort spent on these tasks, `meteR` makes it practical to more fully explore METE models, e.g., by rapidly testing models built from data aggregated at different spatial, temporal or biological scales. Furthermore, owing to its object-oriented construction `meteR` allows users to take full advantage of other packages in `R`, for example model comparison procedures achieved via likelihood and AIC.

`meteR` provides a workflow to facilitate empirical tests of METE (Fig. 3.1). `meteR` calculates state variables directly from a variety of data formats. Next, it calculates the core probability distributions—the *Ecosystem Structure Function* (ESF) and *Spatial Structure Function* (SSF)—from which METE's macroecological predictions arise. These core functions are next used to construct two types of macroecological predictions: probabilistic distributions (e.g. species abundance distribution) and deterministic relationships (e.g. the species-area relationship). The predictions are readily assessed via both plotting functions and statistical tests.

3.3 Package Features

3.3.1 Inputs

`meteR` accepts data in several formats, each of which is represented in the *Input* column in Figure 3.1:

1. One row per individual: This is useful when individual metabolic rate measurements are available

2. Multiple individuals of the same species per row: This is useful when multiple individuals are observed with the same values; e.g., if no metabolic rates are observed or if an average species metabolic rate is used. This format is also helpful if species are only located at the plot level, and individual-level coordinates are not needed/available.

3. Spatial information: This can be provided for individuals or counts of species either by the x, y coordinates of those observations or the *row* and *column* IDs of those observations from a gridded landscape.

4. State variables: One can directly specify the values of N_0, S_0, and E_0, to derive only the theoretical predictions, which is useful to compare how predictions change with different state variables from hypothetical datasets.

Note that models can be fit even when some components of the dataset are missing; e.g., if metabolic rate data are unavailable, `meteR` can still provide predictions that do not rely on these data (e.g., SAD or SAR). If location data are missing, all non-spatial predictions are

available. The data are subsequently included in all model objects (see below) so that plots comparing predictions to observations can be be automatically produced. This flexibility in data formats, along with the fact that data are included with model objects, allows users to aggregate data at different spatial, temporal and taxonomic scales and fit models rapidly to test and organize predictions under a variety of assumptions.

3.3.2 Core Functions

At the heart of METE are two probability distributions that are fit using the principle of maximum information entropy, the ESF and SSF [4]. The ESF is fit with `meteESF(...)` using a nonlinear equation solver [package `nleqslv`; 21] to find the Lagrange multipliers. `meteESF` returns an object of class `meteESF`, for which a variety of $S3$ methods are available. The ESF is typically not compared directly to data, but rather is used to obtain a number of more familiar macroecological patterns such as the SAD, IPD and SPD [Table 3.1; 2, 4, note that we use power—the "P" in IPD and SPD to refer to metabolic rate]. In `meteR`, all of METE's predicted distributions and relationships (cf. Table 3.1) can be obtained simply by passing a `meteESF` object to the appropriate function (Fig. 3.1).

The SSF ($\Pi(n|n_0, A, A_0)$) describes spatial structure via the probability that n individuals of a species are observed in a plot of area A given that n_0 individuals of that species were observed in the entire study area of size A_0. For the case where $A = A_0/2$ METE predicts a uniform distribution over n [p. 159 4]. The more general form of the SSF is obtained similarly to the ESF using `nleqslv` in `meteSSF(...)`, which returns an object of class `meteSSF` inheriting from `meteESF`. `meteSSF` objects can be used directly to predict patterns of spatial aggregation such as the SAR or the endemic-area relationship (Fig. 3.1).

3.3.3 Macroecological Predictions

From the ESF and SSF, all of METE's macroecological predictions can be obtained (Table 3.1). These predictions are divided into probability distributions (class `meteDist`) and deterministic relationships (class `meteRelat`), as denoted in Table 3.1.

We use $S3$ methods for object classes `meteESF`, `meteDist` and `meteRelat` to allow users to query models using familiar tools in `R`. For example, we include `print` and `plot` methods to explore model output. Similarly, for likelihood-based inference, `logLik` and `AIC` can be used to compare METE predictions to other distributions (e.g. comparing log-normal and log-series distributions for the SAD). Measures of model fit comparing METE predictions to data are readily performed using `residuals` or `mse` (for mean squared error). In a hypothesis testing framework, `meteR` also provides tools to estimate the z-score of model fit using either the `logLikZ` or `mseZ` functions. If the z-score is smaller than 1.96 this typically corresponds to failing to reject the hypothesis that the observed data came from a METE distribution.

3.3.4 Directly Working with Probability Distributions

A useful feature of `meteR` is its conceptual treatment of distributions using the `distr` package [22]. For each of the macroecological probability distributions, `meteR` provides functions for the density function (d), cumulative distribution (p), quantile function (q) and simulation (r). This functionality for fitted METE distributions is analogous, e.g., to `dnorm, pnorm, qnorm, rnorm` for the normal distribution in the `stats` package [19]. These functions are used extensively internally in `meteR` functions; e.g. `r` functions are used simulate data from fitted distributions to evaluate model fit via z-scores in `mseZ` and `loglikZ` (see below), while `p` is used for plotting cumulative distributions in `plot`. These distribution functions enable users to extend analyses to include more nuanced hypothesis testing, simulation for theoretical studies and custom plotting.

3.4 Sample Code

We use two datasets, both made freely available and distributed with `meteR`, to illustrate its functionality. The first dataset comes from a 16x16m survey of herbaceous plants in the Anza Borrego Reserve in southern California [4] which we will use to study patterns of abundance and spatial distribution (available as `anbo` in `meteR`). Briefly, the data set consists of 16 $1m^2$ contiguous subplots. In each subplot, the abundance of each species was recorded, however individual metabolic rates were not recorded. Consequently, we also study a complementary dataset which contains individual body size measurements for one plot. This dataset comes from an extensive sample of arthropods collected by canopy fogging in native Hawaiian montane forest [23], which we will use to study patterns of abundance and power (metabolic rate) distribution (available as `arth` in `meteR`).

With the `anbo` data set, we illustrate how to predict the species abundance distribution ($\Phi(n)$), the Species-Area relationship ($S(A)$), and the *Spatial Structure Function* ($\Pi(n)$). The code below generates the predicted distributions and produces Figure 3.2a.

```
data(anbo)
esf1 <- meteESF(spp=anbo$spp, abund=anbo$count)
sad1 <- sad(esf1)
sad1 # print function returns useful summary
```

```
Species abundance distribution predicted using raw data
with parameters:
  S0   N0   E0
  24 2445   NA
    la1     la2
0.00037 0.00109
```

From the SAD object we can extract familiar probability distribution functionality and plot it, e.g.,

```
sad1$r(20) # simulate 20 values
sad1$q(seq(0,1,length=10)) # 10% quantiles
plot(sad1, ptype='rad', log='y') # plotting can be customized by passing
                                 # arguments to generic 'plot' via '...'
```

The SAR and EAR follow similarly; we can either compute these relationships from the ESF or produce an object bundling the theoretical SAR with the observed data (plotting this bundled object produces Fig. 3.2b:

```
sar1 <- downscaleSAR(esf1, A=2^(seq(-2, 4, length=7)), A0=16)
ear1 <- downscaleSAR(esf1, A=2^(seq(-2, 4, length=7)), A0=16, EAR=TRUE)
## sar2 bundles both the predicted SAR and the observed SAR similarly to
## the output of 'sad(...)'
sar2 <- meteSAR(anbo$spp, anbo$count, anbo$row, anbo$col, Amin=1, A0=16)
plot(sar2, xlim=c(1, 2^8), ylim=c(1, 45), log='xy')
```

The upscaled SAR uses a more involved, recursive optimization routine [see eqns (7.70) and (7.71) in 4] but is equally accessible in `meteR`:

```
sarUP <- upscaleSAR(esf1, A0=16, Aup=2^8)
plot(sarUP, add=TRUE, col='blue') # add to previous SAR plot
```

With the `arth` data set, we illustrate the individual power distribution ($\Psi(\epsilon|S_0, N_0, E_0)$ as well as assessment of model fit using likelihood z-scores. Generating power distributions follows the same routine as species abundance distributions, starting with the ESF:

```
data(arth)
esf2 <- meteESF(spp=arth$spp, abund=arth$count, power=arth$mass^(3/4))
ipd2 <- ipd(esf2)

plot(ipd2, ptype='rad') # showing two plotting options: rank and cumulative
plot(ipd2, ptype='cdf')
```

Note that we assume the relationship $power = bodymass^{3/4}$ based on metabolic theory [13]. The two different plotting options display the rank plot and cumulative density plots, respectively, which are shown in panels (a) and (b) of Figure 3.3.

To evaluate model fit, we calculated the z-score of the likelihood of the data under the "null" model that the data truly did come from the METE distribution. Thus a z-score of less than ≈ 2 fails to reject the hypothesis that the data came from a METE distribution.

```
 ipd2.z <- logLikZ(ipd2, nrep=999, return.sim=TRUE) # return simulated values
                                                    # for plotting
 ipd2.z$z # the z-value itself
'log Lik.' -2.713953 (df=2)
```

The z-value less than ≈ 2 indicates we cannot reject METE. Figure 3.3 (panel c) visually confirms this:

```
 plot(density(ipd2.z$sim)) # density of simulated likelihoods
 abline(v=ipd2.z$obs) # observed likelihood
```

More detailed examples are availible as a vignette included in `meteR` and accessible with the command

3.5 Conclusions

By relegating mathematical details and conveniently organizing analyses, `meteR` allows users to readily address new macroecological questions. Until now, tests of METE have primarily been conducted by only a handful of experts, even though many more datasets surely exist that could help us to better understand METE's strengths and weaknesses and macroecological patterns more generally. `meteR` lowers the activation energy for new users to begin to study both macroecology and METE, with the hope that more researchers will become engaged and invest the necessary time to learn the theory. As such, `meteR` is valuable as both a teaching and research tool.

Acknowledgements

We thank J. Harte, Y. Zhang, E. Newman, and C. Lewis for constructive input throughout the development process. We are grateful to D. Gruner and E. Newman for providing data, and the Rocky Mountain Biological Lab where data were collected. AJR acknowledges funding from NSF grant DEB-1241253, the NSF Graduate Research Fellowship and the Gordon and Betty Moore Foundation. CM acknowledges funding from USDA-NRI grant 2008-35615-19014 and NSF grants CHN-1414108 and DEB-1137366.

References

1. Brown, J. *Macroecology* The University of Chicago Press. Chicago, Illinois, USA, 1995.
2. Harte, J., Zillio, T., Conlisk, E. & Smith, A. Maximum entropy and the state-variable approach to macroecology. *Ecology* **89,** 2700–2711 (2008).
3. Harte, J., Smith, A. B. & Storch, D. Biodiversity scales from plots to biomes with a universal species-area curve. *Ecology Letters* **12,** 789–797 (Aug. 2009).

4. Harte, J. *Maximum entropy and ecology: a theory of abundance, distribution, and energetics* (Oxford University Press, Oxford, UK, 2011).

5. Phillips, S. J., Anderson, R. P. & Schapire, R. E. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* **190,** 231–259 (2006).

6. Shipley, B., Vile, D. & Garnier, E. From Plant Traits to Plant Communities: A Statistical Mechanistic Approach to Biodiversity. *Science* (2006).

7. Williams, R. Simple MaxEnt models explain food web degree distributions. *Theoretical Ecology* **3,** 45–52 (2010).

8. Skilling, J. *Maximum Entropy and Bayesian Methods in Science and Engineering, Volume 1, ed* (GJ Erickson and CR Smith, 1988).

9. Jaynes, E. Probability Theory: The Logic of Science. *Cambridge University Press* (2003).

10. Banavar, J., Maritan, A. & Volkov, I. Applications of the principle of maximum entropy: from physics to ecology. *Journal of Physics: Condensed Matter* **22,** 063101 (2010).

11. Jaynes, E. Information theory and statistical mechanics. II. *Physical Review* **108,** 171–190 (1957).

12. Jaynes, E. T. On the rationale of maximum-entropy methods. *Proceedings of the IEEE* **70,** 939–952 (1982).

13. Brown, J., Gillooly, J., Allen, A., Savage, V. & West, G. Toward a metabolic theory of ecology. *Ecology* **85,** 1771–1789 (2004).

14. White, E. P., Thibault, K. M. & Xiao, X. Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. *Ecology* **93,** 1772–1778 (2012).

15. Newman, E. A., Harte, M. E., Lowell, N., Wilber, M. & Harte, J. Empirical tests of within-and across-species energetics in a diverse plant community. *Ecology* **95,** 2815–2825 (2014).

16. Xiao, X., McGlinn, D. J. & White, E. P. A strong test of the maximum entropy theory of ecology. *The American Naturalist* **185,** E70–E80 (2015).

17. Damuth, J. Population density and body size in mammals. *Nature* (1981).

18. White, E. P., Ernest, S. M., Kerkhoff, A. J. & Enquist, B. J. Relationships between body size and abundance in ecology. *Trends in ecology & evolution* **22,** 323–330 (2007).

19. R Development Core, T. R: A language and environment for statistical computing. *Vienna, Austria* (2013).

20. Kitzes, J. & Wilber, M. macroeco: Reproducible ecological pattern analysis in Python. *Ecography* (2016).

21. Hasselman, B. nleqslv: Solve systems of non linear equations. R package version 2.1. 1 (2015).

22. Ruckdeschel, P., Kohl, M., Stabla, T. & Camphausen, F. S4 classes for distributions. *R News* (2006).

23. Gruner, D. S. Geological age, ecosystem development, and local resource constraints on arthropod community structure in the Hawaiian Islands. *Biological Journal of the Linnean Society* **90,** 551–570 (2007).

Figures

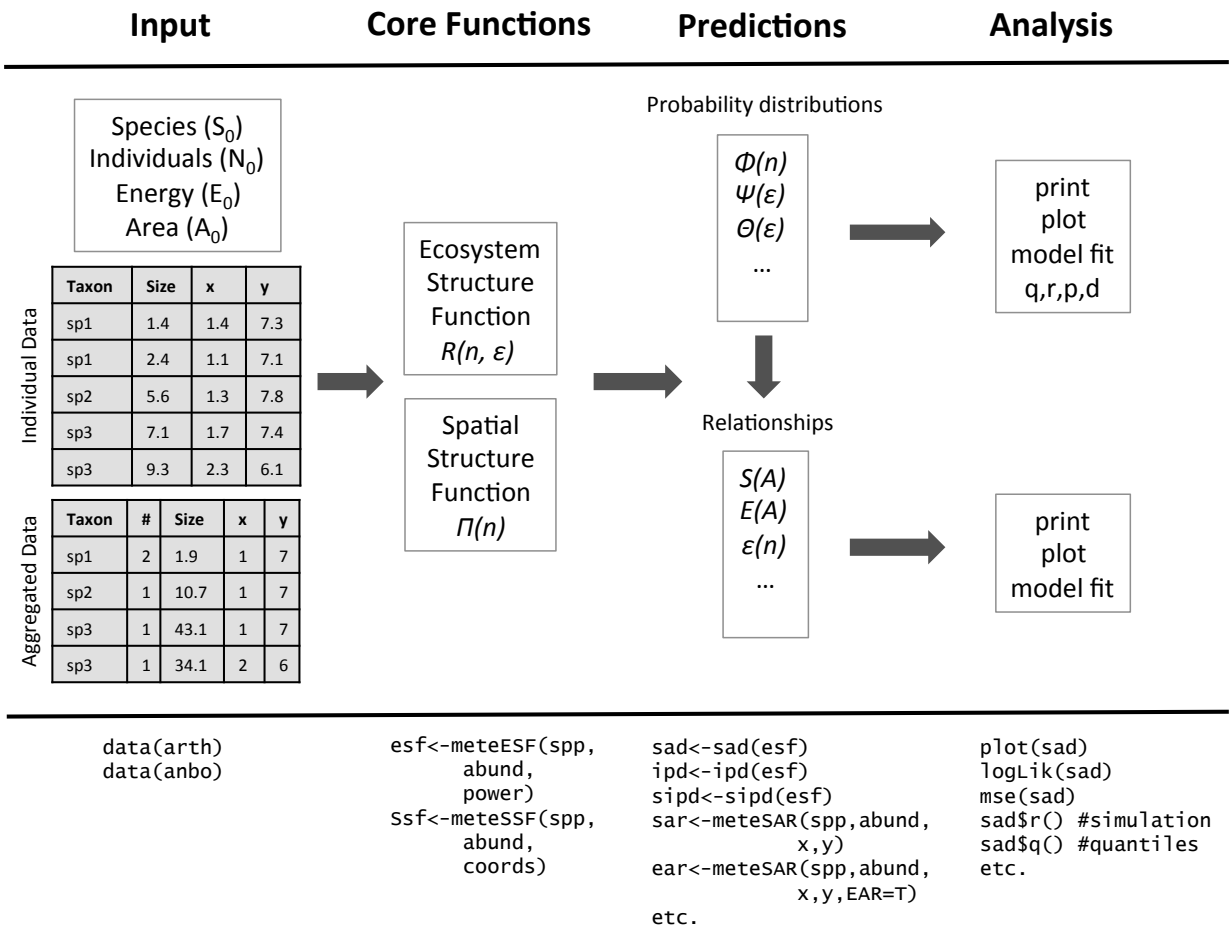

Figure 3.1: `meteR`'s workflow. `meteR` accepts multiple data types, which are used to calculate the core probability distributions from which all predictions arise: the *Ecosystem Structure Function* (ESF) and the *Spatial Structure Function* (SSF). From the ESF and SSF, a variety of macroecological distributions can be calculated (see Table 3.1 for definitions). Each of these can be plotted or summarized in various ways and model fit is readily assessed. Furthermore, density, distribution function, quantile function and random generation is available for each function, allowing for custom plotting, simulation and model evaluation.

Tables

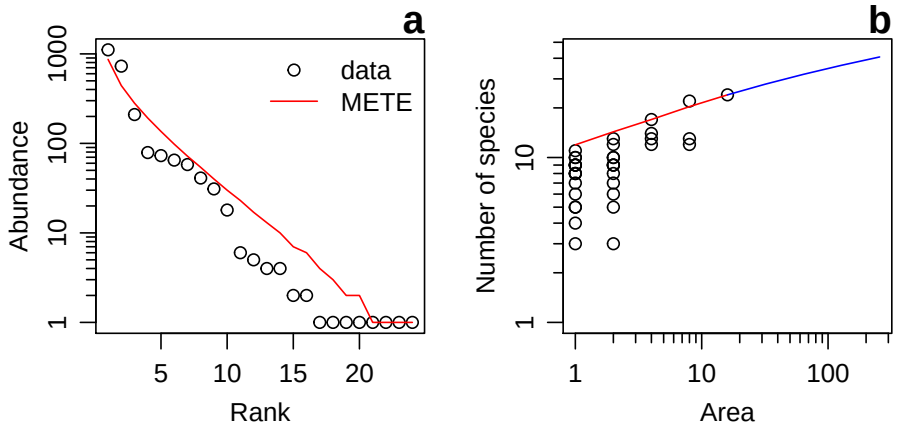

Figure 3.2: Species abundance distribution (a) and species area relationship (b) for the `anbo` dataset. The blue line in (b) depicts the upscaled SAR predicted by METE. METE fits the Anza Borrego data poorly at increasingly small scales. Understanding such systematic deviations from theory could be an opportune area of exploration, and initial work suggests that such deviations could be an indicator of anthropogenic disturbance (Newman *et al.*, in prep.) such as invasive species, which are present in the Anza Borrego dataset. Note that we have stylized the log axes using custom code not included in `meteR`.

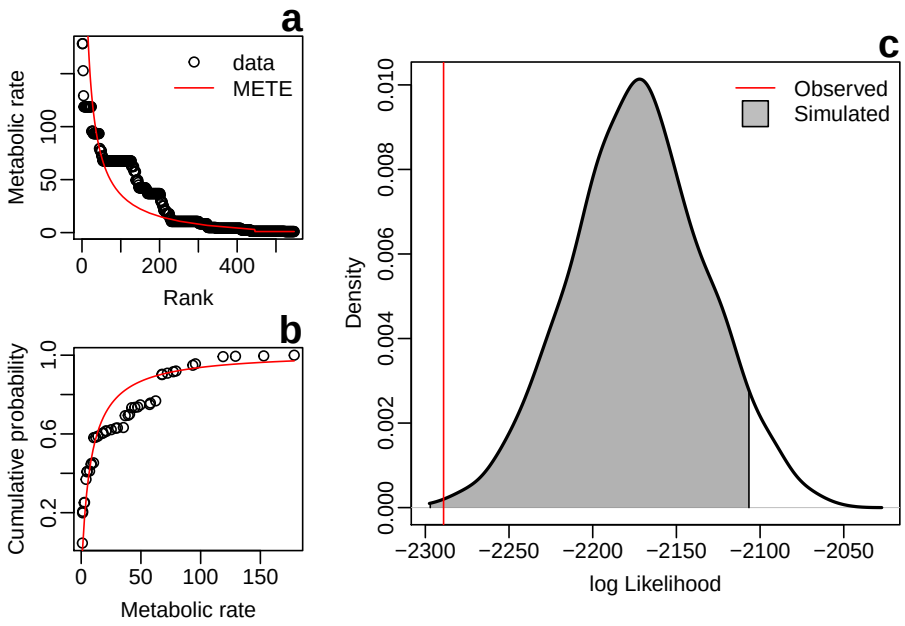

Figure 3.3: Individual power distributions shown as a rank plot (a) and cumulative density plot (b). To evaluate model fit we compare the observed likelihood of the data given the METE model (red line in panel c) to simulated likelihoods produced by simulated `nrep` communities from the fitted METE object and calculating their likelihoods. Thus the simulated distribution represents hypothetical likelihoods for datasets obeying METE.

Table 3.1: METE predictions. Following notational conventions in Harte [4], Z and Z_Π are partition functions (i.e. to ensure their respective probability distributions to sum to 1), λ_1, λ_2, and λ_Π are Lagrange multipliers, and commonly used combinations of them include $\beta = \lambda_1 + \lambda_2$, $\sigma = \lambda_1 + E_0\lambda_2$, and $\gamma = \lambda_1 + \epsilon\lambda_2$.

Predictions	Acronym	Functional Form	Type	`meteR` function
Ecosystem Structure Function	ESF	$R(n, \epsilon \vert N_0, S_0, E_0) = \frac{1}{Z} e^{-\lambda_1 n} e^{-\lambda_2 n \epsilon}$	Core Distribution	`meteESF`
Spatial Structure Function	SSF	$\Pi(n, \vert n_0, A, A_0) = \frac{1}{Z_\Pi} e^{-\lambda_\Pi n}$	Core Distribution	`meteSSF`
Distribution of Species Abundance	SAD	$\Phi(n \vert N_0, S_0) \approx \frac{1}{log\frac{1}{\beta}} \frac{e^{\beta n}}{n}$	Distribution	`sad`
Distribution of individual Power across community	IPD	$\Psi(\epsilon \vert N_0, S_0, E_0) \approx \lambda_2 \cdot \beta \frac{e^{-\gamma}}{(1-e^{\gamma})^2}$	Distribution	`ipd`
Distribution of individual Power for species with n individuals	SIPD	$\Theta(\epsilon \vert n, N_0, S_0, E_0) \approx \lambda_2 n e^{-\lambda_2 n(\epsilon - 1)}$	Distribution	`sipd`
Distribution of average (over individuals) metabolic rates over species	SPD	$\nu(\overline{\epsilon} \vert N_0, S_0, E_0) \approx \frac{1}{log\frac{1}{\beta}} \frac{e^{\beta/\lambda_2(\overline{\epsilon}-1)}}{\overline{\epsilon}-1}$	Distribution	`spd`
Average individual metabolic rate for species with n individuals	ebar	$\overline{\epsilon}(n \vert N_0, S_0, E_0) \approx 1 + \frac{1}{n\lambda_2}$	Relationship	`ebar`
Species Area Relationship	SAR	$\bar{S}(A) = S_0 \sum_{n=1}^{N_0} (1 - \Pi(0 \vert n)) \Phi(n)$	Relationship	`meteSAR`, `downscaleSAR`, `upscaleSAR`
Endemics Area Relationship	EAR	$\bar{E}(A) = S_0 \sum_{n=1}^{N_0} \Pi(n \vert n) \Phi(n)$	Relationship	`meteSAR`, `downscaleSAR`

Conclusion

stub

Appendix A

Supplement to "Punctuated non-equilibrium and niche conservatism explain biodiversity fluctuations through the Phanerozoic"

A.1 Limit distribution of a time-averaged homogeneous origination-extinction process

Fossil taxa gain and lose taxa according to an origination-extinction process. We assume that most fossil occurrences of a taxon come from the period of its history when it is dominant and in steady state. In a time slice of duration τ during such a period of steady state the latent per capita rates of origination and extinction would be equal (i.e. $\lambda = \mu \equiv \rho$) and the number of origination or extinctions events (call such events Y) each follow an inhomogeneous Poisson process with rate ρN_t where N_t is the number of species or genera in the taxon of interest at time t. Allowing N_t to vary smoothly with time, and invoking the comunicative property of the Poisson distribution, we arrive at the number Y of extinction *or* origination events in τ being distributed

$$Y \sim \text{Pois}(\rho \int_{t=0}^{\tau} N(t)dt). \tag{A.1}$$

Under the steady state assumption we can approximate $N(t)$ by $\bar{N}$, the steady state diversity, leading to

$$Y \sim \text{Pois}(\rho \bar{N} \tau). \tag{A.2}$$

Assuming the τ of each time period in the Paleobiology Database or Sepkoski's compendium to be approximately equal (i.e. equal durations of major stratographic units) then the distribution of fluctuations within taxa will be asymptotically Gaussian.

A.2 Additional super-statistical analyses

To evaluate the sensitivity of our super-statistical analysis on the particular data used and we tested our predictions on different data sets (see below). The fact that it works in all different applications indicates that it is robust to vagaries of different recording strategies and bias corrections in paleobiology. This could mean that much of the raw signal in massive fossil datasets, at least signals regarding fluctuations, are not artifacts of sampling, as has been proposed before [1].

A.2.1 Raw PBDB data

We calculated the super-statistical prediction at the order level from raw genus diversity recorded in the PBDB without correcting for taphonomic or sampling bias (Fig. A.3). The super-statistical calculation also closely fits the raw data as in the case of sampling and publication bias-corrected data.

A.2.2 Different taxonomic ranks in PBDB data

As noted in the main text, the super-statistical prediction predictably breaks down at higher taxonomic scales. In Figure A.4 we present this worsening fit graphically using class level data with three-timer and publication corrected PBDB data

A.2.3 Sepkoski's compendium

Sepkoski's compendium [2] provided the first hypothesis of Phanerozoic diversification. As such, it has served as a benchmark for further investigation into large-scale paleobiological patterns[3]. We conducted the same super-statistical analysis as in the main text and find comparable results. Specifically, the super-statistical prediction far out preforms the null Gaussian model (Fig. A.5) and worsens with increasing taxonomic scale (Fig. A.5), again implying the uniqueness of orders.

References

1. Hannisdal, B. & Peters, S. E. Phanerozoic Earth System Evolution and Marine Biodiversity. *Science* **334,** 1121–1124 (2011).
2. Sepkoski, J. J. *A compendium of fossil marine animal families* (Milwaukee Public Museum, Milwaukee, WI, 1992).
3. Alroy, J. *et al.* Phanerozoic Trends in the Global Diversity of Marine Invertebrates. *Science* **321,** 97–100 (2008).
4. Alroy, J. The Shifting Balance of Diversity Among Major Marine Animal Groups. *Science* **329,** 1191–1194 (2010).

Supplemental Figures

Figure A.1: Relationship between number of publications and genus diversity as recorded by the PBDB.

Figure A.2: Comparison of SQS method [4] (solid black line) with the raw data (dashed black) and our three-timer and publication bias correction method (red). The time-series of all marine invertebrate genera shows general agreement with the only major deviations toward the modern (A). Despite these differences the distribution of fluctuations in genus diversity across all marine invertebrates show good agreement (B).

Figure A.3: Super-statistical prediction of raw (i.e. not bias corrected) order-level fluctuations in genus diversity recorded in the PBDB. Grey dots are the full data of orders, while black ones are orders with more than 15 points. The red line is our theoretical prediction and the blue line the best Gaussian fit to the data.

Figure A.4: Super-statistical prediction of bias corrected class-level fluctuations in genus diversity recorded in the PBDB. Grey dots are the full data of orders, while black ones are orders with more than 15 points. The red line is our theoretical prediction and the blue line the best Gaussian fit to the data. Note at the class level the fit is predictably worse, see main text for discussion.

Figure A.5: Super-statistical prediction (red line) of fluctuations in genus diversity recorded in Sepkoski's compendium of marine invertebrates compared to maximum likelihood normal distribution (blue line). Super-statistical theory explains order level fluctuations well (A) with increasingly poorer fits at the class (B) and phylum (C) levels.

Appendix B

Supplement to "Community assembly on isolated islands: Macroecology meets evolution"

B.1 Compilation of networks and metric validation

Researchers have put forward a set of "network metrics," including nestedness [1, 2] and modularity [3, 4], to understand the complex structure of ecological networks. Null models are used to evaluate the statistical significance of these metrics and to compare between networks of different size [2]. We compare the results derived from two common null models: the "probabilistic null" of [1] using the relative degree distributions of plants and herbivores as weights while randomizing links and suffers from high Type II error [2]; the "fixed-fixed null" [2] maintains the exact number of links assigned to each species while randomizing which interactors fill the requisite set of links and suffers from high Type I error [2]. We find that using these different null models does not change any trends in our network statistics across the Hawaiian chronosequence but different null models do influence the sign and significance of the network metrics (Fig. B.1). We therefore do not interpret the sing or significance of the metrics but only their relative trends across site age.

Because these networks are based on opportunistic data associated with species descriptions, and not based on standardized ecological surveys, we cannot interpret patterns in network metrics without evaluating possible sampling biases [5–7]. To do so we rarify networks by the number of Hemiptera species included and, for each subsampled network, re-calculate nestedness and modularity z-scores. This rarefaction procedure shows that nestedness is very sensitive to network size (Fig. B.3), a known property of nestedness [5–7]. However the relative nestedness z-scores across networks remain qualitatively similar to those observed for the complete networks (Fig. B.3). Modularity depends on network size in a more variable way (Fig. B.3). Modularity is expected to decrease with network size [7] and so the marked increase in modularity with network size on Haleakala is unexpected. However in light of

the large number of highly specialized taxa this pattern is more reasonable—if most species only have within module links then removing these species through subsampling will only reduce overall modularity. Thus this pattern speaks to the high level of specialization on Haleakala, and to a lesser extent at Kokee which also shows a slight increase in modularity with network size (Fig. B.3).

B.2 R Scripts for Maximum Entorpy Analysis and Monte Carlo Methods

B.2.1 Needed Functions

```
require(distr)

## d, p and r functions for the maxent link distribution
## also a funciton to calculate the MLE and log likelihood

dmelink <- function(x, la) {
    exp(-la*(x-1)) - exp(-la*x)
}

pmelink <- function(x, la, lower.tail=TRUE) {
    if(length(x) > 1) {
        cp <- sapply(x, function(q) sum(dmelink(1:q, la)))
    } else {
        cp <- sum(dmelink(1:x, la))
    }

    if(lower.tail) {
        return(cp)
    } else {
        return(1 - cp)
    }
}

rmelink <- function(n, la, X0) {
    sample(X0, n, rep=TRUE, prob=dmelink(1:X0, la))
}


qmelink <- function(x, la) {
    fun <- DiscreteDistribution(1:3000, dmelink(1:3000, la))
```

```
    fun@q(x)
}

## likelihood functions (for simple cases maxent solution is equivilant
## to maximum likelihood solution so we use MLE for computational ease)
melink.mle <- function(x) {
    log(1 + 1 / mean(x - 1))
}

melink.logLik <- function(x) {
    la <- melink.mle(x)

    length(x) * log(1 - exp(-la)) - la * sum(x - 1)
}


## function to make rank function (for plotting) of maxent link distrib
melink.rankFun <- function(x) {
    qmelink(seq(1-1/(length(x)+1), 1/(length(x)+1),
                by=-1/(length(x)+1)), melink.mle(x))
}

## mean squared error function for maxent link distrib
melink.mse <- function(x) {
    mean((sort(x, TRUE) - melink.rankFun(x))^2)
}

## monte carlo method for calculating distribution of mse values
## under model where maxent truely generated the link distribution
sim.melink.z <- function(x, X0, nsim=999) {
    la <- melink.mle(x)

    res <- replicate(nsim, {
        # browser()
        sim <- rmelink(length(x), la, X0)
        return(melink.mse(sim))
    })

    obs.mse <- melink.mse(x)

    return(list(obs=obs.mse, z=(obs.mse - mean(res)) / sd(res), sim=res))
}
```

B.2.2 Example Use

```
## set a Lagrange multiplier value and simulate data
la <- 0.01
x <- rmelink(2000, la, 2000)

## fit the maxEnt model to simulated data
x <- rmelink(2000, la, 2000)

## evalute if fitted maxEnt model matches the data
melink.mle(x) # the MLE should be near 0.01
plot(sort(x, TRUE)) # the plotted data should look like the theory
lines(pmelink(1:200, melink.mle(x)), col='red')

## test that the logLike function is returning the correct value
melink.logLik(x) - sum(log(dmelink(x, melink.mle(x)))) < 1e-12

## test that the log likelihood simulation is working
sim.melink.z(x, 200) # should be a small z-value
sim.melink.z(rbinom(100, 200, 0.1), 200) # should be a large z-value
```

References

1. Bascompte, J., Jordano, P., Melian, C. J. & Olesen, J. M. The nested assembly of plant-animal mutualistic networks. *Proc. Natl Acad. Sci. USA* **100,** 9383–9387 (2003).
2. Ulrich, W., Almeida-Neto, M. & Gotelli, N. J. A consumer's guide to nestedness analysis. *Oikos* **118,** 3–17 (2009).
3. Newman, M. & Girvan, M. Finding and evaluating community structure in networks. *Physical Review E* **69,** 026113 (2004).
4. Olesen, J., Bascompte, J., Dupont, Y. & Jordano, P. The modularity of pollination networks. *Proc. Natl Acad. Sci. USA* **104,** 19891–19896 (2007).
5. Nielsen, A. & Bascompte, J. Ecological networks, nestedness and sampling effort. *Journal of Ecology* **95,** 1134–1141 (2007).
6. Gibson, R. H., Knott, B., Eberlein, T. & Memmott, J. Sampling method influences the structure of plant-pollinator networks. *Oikos* **120,** 822–831 (2011).
7. Rivera-Hutinel, A., Bustamante, R., Marín, V. & Medel, R. Effects of sampling completeness on the structure of plant-pollinator networks. *Ecology* **93,** 1593–1603 (2012).

Supplemental Figures

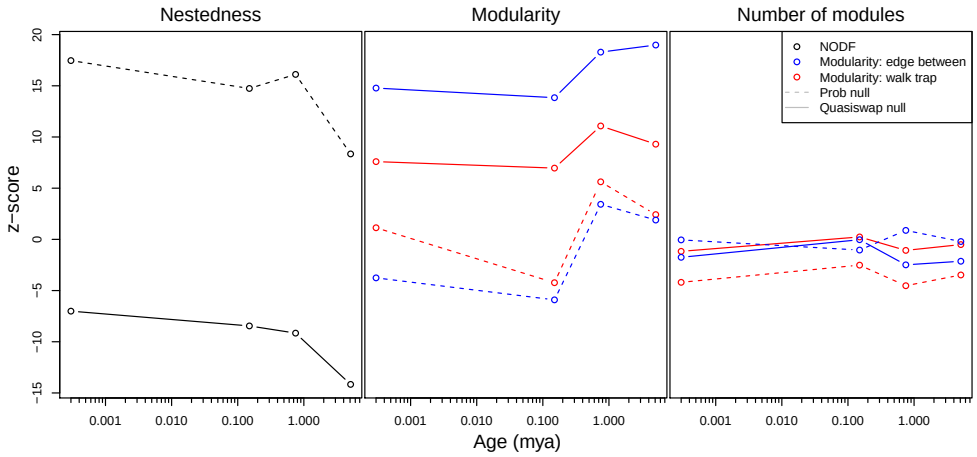

Figure B.1: Comparison of different null models ("Prob" and "Quasiswap") used to standardize network metrics and comparison of different algorithms for assessing modularity ("edge between" and "walk drap"). Choice of null model has a strong influence on the sign and magnitude of metrics but not on their relative trends. The different modularity algorithms lead to largely similar qualitative patterns.

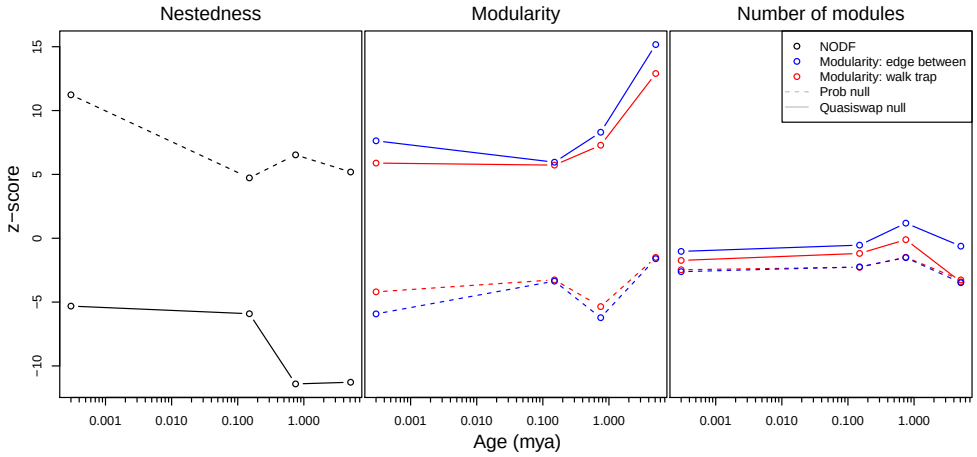

Figure B.2: Metrics NODF and modularity calculated for networks based on more biogeographically conservative assignment of Hemiptera to localities. Colors and metric specifics as in Figure B.1.

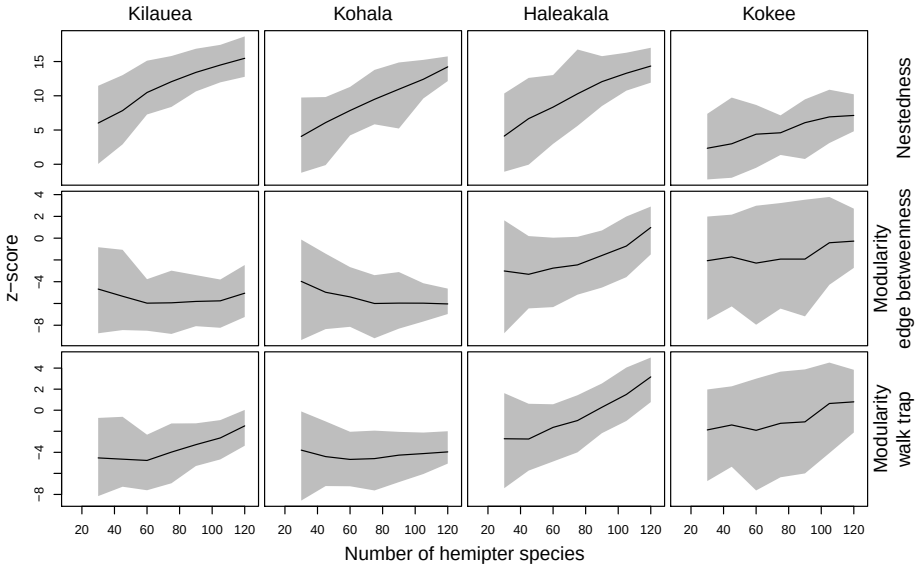

Figure B.3: Result from rarification analysis showing sensitivity of network metrics to number of Hemiptera sampled.