Boundaries and dynamics of biomes

Daril Vilhena

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Carl Bergstrom, Chair

Chris Sidor

Tom Daniel

Peter Hoff

Program Authorized to Offer Degree:
UW Biology

University of Washington

Abstract

Boundaries and dynamics of biomes

Daril Vilhena

Chair of the Supervisory Committee:
Professor Carl Bergstrom
Department of Biology

Species are packed into biogeographic zones, where evolution can effectively operate independently to forge evolutionary novelty. Biomes are perhaps the most relevant unit of evolutionary progress, with the vast majority of evolutionary radiations being constrained within their walls. A fundamental question in Macroecology is how biomes historically assembled and why species are distributed in them as they are. First, quantitative methodology to delineate biomes are proposed here to identify where biomes are and have been. Second, extinction is studied as a process that contributes to biome turnover. Third, the Phanerozoic fossil record is assessed for biases that need to be overcome to delineate the shifting spatial boundaries biomes over 500 million years.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I am profoundly grateful to Carl Bergstrom for fostering an environment of interdisciplinary collaboration and for making science fun. There is no place in the world quite like Carl's lab, and for me that has been so influential in helping me shape my drive, values, and priorities. Thank you for patiently listening to all of my crazy ideas, each month in a different field, and for always being able to extract a simple problem from a seemingly complicated impossible problem. Jevin West, Jacob Foster, and Frazer Meacham: your curiosity about the world and drive to solve problems is inspirational. Thank you for listening to the endless ideas I have bounced off of you guys. Martin Rosvall - you have been greatly inspirational to me in the way you think about structure in the natural world and your work is behind everything in this thesis. The entirety of this work grew out of a few conversations with Elisha Harris, when we realized the potential for network methods to find spatial patterns in fossil datasets. We brought our approach to the UW Paleo faculty, where Christian Sidor, Gregory Wilson, and Caroline Strömberg helped us to develop the approach and create the initial foundation for which this work would be based. Alexandre Antonelli was instrumental in my realization that uniting modern biome science with paleontological biome science is the ultimate challenge. Andrew Smith, who has been writing about spatial biases in the fossil record for years, was a critical help for estimating the potential of the framework in this thesis to be applied to deep time. I am also deeply appreciative to Joshua Plotkin and Michael Levy, my undergraduate mentors, for fostering my love of theoretical biology.

Chapter 1

INTRODUCTION

1.1 Macroecology as an emerging discipline

Biologists have long recognized the macroecological organization of the natural world. From ecological communities to collections of communities (metacommunities) to collections of metacommunities (biomes), the planetary biota is structured hierarchically. The study of this inherent biogeographic structure in the natural world is a pillar of the natural sciences, and biogeographic units at any spatial scale often serve as units for analysis across myriad disciplines in the ecological sciences, from quadrats to localities to grid cells that span vast areas.

The term “biogeographic unit” therefore corresponds to real biological divisions that vary with spatial scale. In the oregon coastline, a 1x1 meter plot is an intertidal strip, a 5x5 meter plot is the rocky intertidal, a 100x100 meter plot incorporates a sandy beach, and 1x1 km plot adds in foredune forest and an estuary. Though these heterogeneous components are all ecologically connected, the scale we study determines our conclusions [172]. To study emergent biogeographic phenomena requires identification of biogeographic structures at the scale of interest, and assessment of the interconnectivity between those structures.

Given a collection of biogeographic units at any spatial scale, the challenge is to extract insight from those data to reveal spatial patterns in the biota. A variety of problems of could be tackled with an arsenal of methods, for example: the tendency for trees in a forest to cluster together, the correlation of climatic factors with species range limits, the inference of ecological interaction between species from overlapping species range data, the correlation of biological richness with increased plot size, and even determining what factors create the biogeographic units used for analysis. The study of these properties, in time and space, was coined “Macroecology” in 1989 by James Brown in a seminal paper in Science [23, 22, 82].

Since its inception, Macroecology has been recognized as one of the fastest growing fields in Ecology with perhaps the most broadly applicable relevance to disciplines throughout the life sciences. However, as a primarily quantitative, theoretical discipline, the greatest challenges remain in the discovery of quantitative methodology to enable the most interesting hypotheses to be rigorously scrutinized.

Perhaps the greatest challenge in Macroecology is why the biota are partitioned hierarchically as they are and what macroevolutionary and geological factors have forged those boundaries, and how those boundaries have been created, altered, and destroyed. For example, plant lineages that collectively create the Cerrado, the world's most species rich tropical savannah, coincides with a proliferation of C4 grasses and the global expansion of the savvanah biome [150]. The Cerrado is hypothesized to have formed in situ through rapid adapatation to fire from a multititude of lineages [150].

These processes have undeniably had a major role in forging macroevolutionary novelty -biogeography is at its heart the boundaries by which evolution operates and the local ecological ruleset by which evolutionary novelty can be achieved. In this dissertation, I tackle the practical methodological challenges that inhibit our understanding of the assembly of every biome, past and present.

1.2 Delineation of realms, biogeography, and biomes

Alfred Wallace was the first to demarcate the major biogeographic regions of the Earth, and his six biogeographical regions formed the basis for those used today (Fig. 1). Wallace based his divisions of the Earth on his intuitions about the regional fauna, and the limited geographic range data available to him. Since then, biologists have sought abiotic, evolutionary, and ecological explanations for the boundaries that divide the Earth into functional subunits.

Agglomerative clustering superseded intuition when Edwin Hagmeier released twin papers in 1964 and 1966 [54, 53], and this approach has formed the basis for the standard methodology in use today [89]. First, he recorded a species-locality matrix, where each entry in the matrix M_{ij} is 1 if species j occurs in locality i, and 0 otherwise. He built a distance matrix from the pairwise distances between each row vector in this matrix and used an agglomerative clustering method to produce a coarse hierarchical representation of the

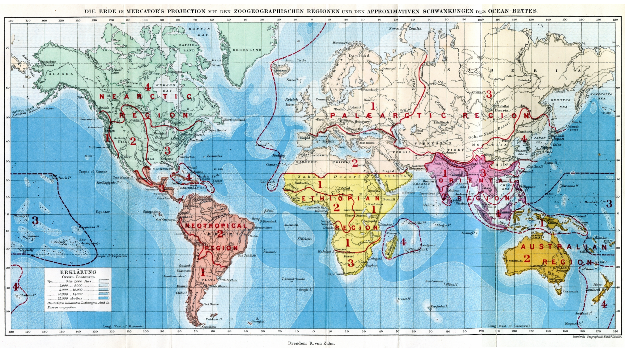

Figure 1.1: Alfred Wallace's six biogeographical regions. Technically, Wallace borrowed the six regions from Philip Sclater (who did it for birds), but Wallace gets the credit because he did it for groups of animals that were more suitable for biogeography.

distances between localities. He noticed that where he cut the tree produced biogeographic structures of different scale.

One might suspect from Hagmeier's pioneering work that quantitative divisions of the Earth would take over as the norm and enable the testing of fundamental hypotheses in macroecology. However, while the division of fauna and flora into biogeographic units, at any spatial scale, seems like a well-defined quantitative problem, the data are plagued by inconsistent taxonomy, coarse spatial resolution, and spatial autocorrelation. For paleontological studies, which one must tackle in order to deduce the dynamics that led to the present, fossilization processes can erase occurrences of taxa at huge spatial scales, systematically biasing estimates of biogeographic delineations.

These practical challenges have forced quantitative biogeography into relative obscurity. Quantitative methods are rarely used for large-scale studies due to limited data that are fraught with biases, though immense progress has been made for mammals, amphibians, and birds, where conservation effort has fueled an enormous drive to measure the distribution of those animals. Less preferable approaches that do not use geographic data from the species directly have been developed: large-scale biomes can be assembled from pixel color differences in satellite imagery data, and the best available database of biome structure in the world was assembled from the opinions of thousands of regional experts (Fig. 2).

While such approaches are certainly sufficient to provide a rough approximation, they

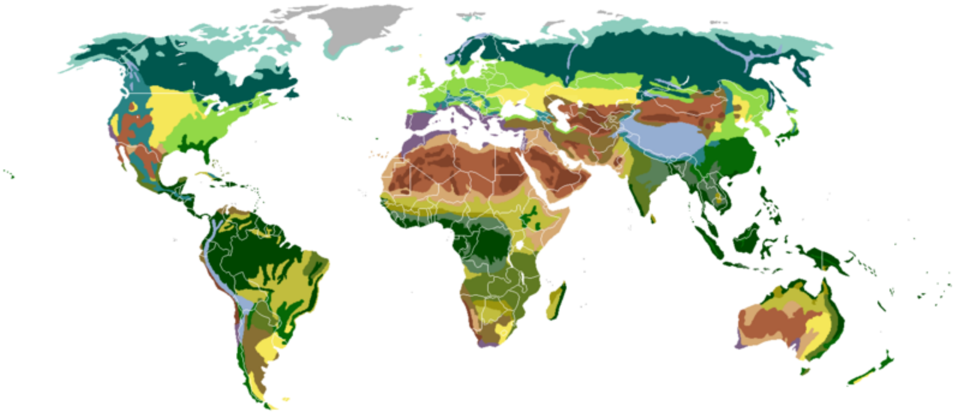

Figure 1.2: Delineation of plant biomes from thousands of expert opinions.

fail to provide the rigorous platform by which we can quantify what a biome is and deduce what forces have created what we have quantified.

1.3 To study biome dynamics requires new methodology

In this thesis I spend three chapters developing methodology and gauging the obstacles that, if surpassed, would enable the measurement of historical biome boundaries and then facilitate the creation of models to test what processes altered those boundaries in geological time.

In Chapter 2 the core machinery is developed: a network based approach that is intuitively simple but softens the effect of many of the biases that have plagued analysis of many biogeographic datasets. This is fundamentally based on the idea that the number of shared species between two regions is not necessarily always a good indicator of biological similarity. Ecologically, competition between species may prevent co-occurrence between species at small to medium spatial scales, despite that those species belong to the same biogeographical unit. Biomes also span enormous latitudes - it is rare for a specific plant species to be spread entirely throughout the North American Grasslands. This reveals the broader issue of biomes containing heterogeneous internal climates that affect individual species, but rarely lineages within the biome. In the fossil record, taphonomic processes can obscure co-occurrence relationships - for example, aragonitic and calcitic shells may not be preserved together in some lithological settings [80]. Here, global biomes are constructed for amphibians and the classical plant biomes are identified for the United States, which is a challenge that shared-species based clusterings have been unable to delineate.

In Chapter 3 the network methods developed in Chapter 2 are used to test a changing biome hypothesis in the fossil record. The Cretaceous-Paleogene (K/PG) was Earth's last great mass extinction, which caused enormous diversity loss and immense changes to the composition of the planetary biota in both the terrestrial and marine realm. Biomes change very slowly - so the mass extinction provides an excellent model system to test whether biomes can differentially buffer extinction - suggesting an emergent property of biomes rather than greater resistance to extinction of individual lineages. Extinction may be one of the largest macroevolutionary forces that shapes the composition of the biota.

In Chapter 4 the extent of spatial biases in the entire marine fossil record is quantified and studied. To develop theories of biome change grounded in historical biogeographic data requires enough data to distinguish biological signal from sampling artifact. New methods will be required to create fossil datasets that can be used to propose biogeographic boundaries in time and space, and the analysis in this chapter is intended to provide a guideline for where and when in the fossil record such methods will be successful.

Chapter 2

BEYOND SIMILARITY: DELIMITING BIOMES WITH NETWORKS

Daril Vilhena, Alexandre Antonelli

Species turnover measures are often used to quantify biodiversity patterns. However, when delimiting biomes from biogeographic data, clusterings of species turnover matrices appear to miss many biome-level regions. Biomes are challenging to identify from distributional data because they often contain climate heterogeneities and span massive latitudes, complicating the inference of their boundaries with measures based on shared species. To resolve this, we describe a network-based approach for biome delineation that incorporates complex, “higher-order” presence-absence patterns, and use it to detect amphibian biomes of the world with a dataset of c. 6,100 amphibian species, the classical plant biomes of North America with a dataset of c. 17,600 vascular plant species, and a published hypothetical dataset that contains a zone of faunal interchange. Our findings suggest that i) our approach can help resolve the debate over biome boundaries and their nomenclature, and ii) a network-driven approach for describing presence-absence relationships offers biogeography a new avenue for extracting biodiversity patterns from distributional data.

2.1 *Introduction*

Much attention has been devoted to develop methods that can confidently assign individuals to populations [15, 57], and then group those populations into phylogenetic entities that deserve the status of species or evolutionary units [95, 34]. How species then co-exist and co-interact to build biomes, of similar eco-physiological and climatic characteristics, and eventually realms, is considerably less understood [119, 118, 150, 32]. Most (macro-)ecological and biogeographic studies take for granted the identity of biomes as well known and accepted from African savannas to artic meadows or tropical rainforests. Yet, there is no consensus on how to best delimit biomes, nor which terminology should be used for referring to them collectively (including terms such as zoo/phyto-geographic regions, biomes, ecosystems, ecozones) or individually (see [65] for examples of different denominations of Neotropical biomes, such as the Cerrado vs the Brazilian savanna and the Pramo vs Puna vs high-altitude grasslands).

This lack of consensus and analytical development is surprising, given that biomes form the basis for the study of evolutionary, ecological, and geological processes that spatially structure the biota. A biome-based approach in macro-ecology can be used to assess the extent to which lineages are able to cross major ecophysiological barriers over evolutionary time, i.e. their degree of biome or niche conservatism in the sense of Wiens [174, 173, 88, 31]. This approach also offers important advantages in conservation biology as compared to the analysis of single lineages, not least in species rich tropical areas such as seasonally dry tropical forests [161] where efforts may be better directed towards protecting remaining patches of threatened biomes than focusing on particular species. Evidence is growing that different biomes will be affected differently by climate change [143], so understanding their origins and evolution may provide further indications of their expected resilience.

A data-driven approach for resolving the lack of established nomenclature and boundaries for regional biomes should draw on the rich history of establishing boundaries between arbitrary biogeographic zones. Since Wallace and his contemporaries [170], biologists have sought to delineate geography into biogeography. Deductive approaches developed early on were later made analytical [54]. However, the detection of BUs, and therefore biomes, is impacted by how we choose to quantify biogeographic structure, which for several decades has chiefly been a variety of species turnover measures based theoretically on beta diversity [171, 87, 54, 89]. Species turnover, as measured by set-based similarity measures such as the Jaccard [78], Sørenson [36], and β-similarity [87], quantifies the relationship of one region to another, typically by dividing the number of shared species between two regions by some measure of the total species in both regions.

Despite their widespread use, species turnover measures can miss intricacies of distributional data that are relevant for biome detection. First, turnover increases on average with distance from a source, potentially confusing a property of finite ranges with increased disparity [104]. Second, for small spatial scales turnover can overestimate disparity due to competitive exclusion, spatial clustering, and environmental gradients [167]. Though this problem can be reduced with large plot sizes, the problem can persist even for large spatial scales. Furthermore, competitive exclusion can create geographic boundaries between species that cohabit the same biome. Third, some biomes span many degrees of latitude, such as the North American Rocky Mountains and American Great Plains, and can contain climate and environmental heterogeneities that can cause endemics to occupy non-overlapping fractions of the biome. Fourth, differences in taxonomic sampling can inflate turnover. For example, taxonomic standards may differ within biomes for rare species. For deep time studies, marine fossil assemblages may for instance not co-preserve aragonitic and calcitic shells. These processes collectively bias turnover measures, because the number of shared species cannot always be trusted as good gauge of biogeographic membership.

Here we adapt a new approach that uses associational networks to minimize the problems described above and extract more community-level information from distributional data. We find that this representation of the data can be used to detect biome-level regions in two well-sampled and cleaned datasets: all amphibians at a global level and most plants in the United States of America. The datasets are aggregated at different scales (global versus country) and grain (two degree grid cells versus US counties) and contain vastly different sampling methodologies. Yet, all produce results that are strikingly comparable with opinion-generated biome delineations. Our findings suggest that distributional data can be mined for more than the number of shared species between regions to improve the delineation and nomenclature of the world's biomes.

2.2 Associational network of species distributions

A bipartite network has two disjoint sets of nodes with no links between nodes of the same set. Many biological systems have been abstracted as bipartite networks, such as plant-pollinator interactions inferred by visitation [14], sexual contact between heterosexual partners [43], and interactions between prey and bait proteins generated by yeast two hybrid screening [162].

The geographic relationships between species and localities can also be abstracted as a bipartite association network, where links are the occurrences of species within geographic

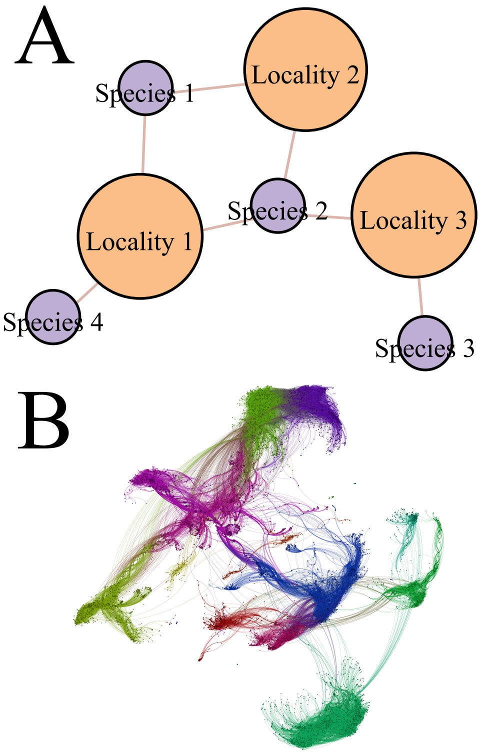

Figure 2.1: **Occurrence networks**. A) An example bipartite occurrence network. Species 1 and 2 jointly occur in Locality 1 and 2, which creates a 4-path that loops, while Species 3 and 4 share a 4-path that does not loop, revealing that an intermediary species "connects" the two. B) An abstract visualization of the amphibian network. Widespread species act as highways between biogeographic realms, creating links between clusters. Node positions determined by the Force Atlas algorithm in the Gephi package.

locations. Interpretations derived from analyses of presence-absence networks are comparable with plant-pollinator networks, because relationships between entities of the same set are associational, such as co-visitation versus co-occurrence. Second order relationships in presence-absence networks are paths of lengths two, or 2-paths. The number of 2-paths between species is the number of times those species co-occur, while the number of 2-paths between a pair of localities, regions, or grid cells is the number of species shared by both grid cells. A more complicated pattern is the number of joint occurrences, where two species occupy the same two localities. The number of joint occurrences can be measured as the number of 4-paths that complete a loop (Fig. 1A). These relationships can be combined to reveal properties of geographic ranges. For example, the number of 3-paths between a species A and locality B divided by the number of 2-paths exiting from species A is the fraction of co-occurrences of species A that also occupy locality B. By setting up the machinery to capture "higher-order" patterns, we can detect complex patterns of presence-absence.

The adjacency matrix A of this network formally expresses species occurrences, and is written

$$A_{ij} = \begin{cases} 1 & \text{if node } i \text{ is linked with node } j \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Because species cannot occur in species and localities cannot occur in localities, the upper left block and lower right block in this matrix are entirely zeroes, if the rows and columns of the matrix are ordered first by species ($1...n$) and second by grid cells ($n+1...n+m$), producing a square matrix with $n+m$ rows and $n+m$ columns. This is written

$$A = \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}, \tag{2.2}$$

where B is the binary presence-absence matrix, in which rows are taxa and columns are localities. The square of the adjacency matrix A gives the co-occurrence matrix C between taxa as the upper left square, or number of co-occurrences between pairs of species, and the matrix of shared species S as the bottom right square, or number of shared species between pairs of grid cells

$$A^2 = \begin{pmatrix} C & 0 \\ 0 & S \end{pmatrix}, \tag{2.3}$$

where elements in the upper right and lower left squares of the matrix are zeroes because 2-paths are exclusively between species and species or locality and locality. Total paths of length i between nodes can be expressed by raising the matrix to the ith power. By formulating the data in this way new measures can be derived and tools from network theory can be readily applied – perhaps most importantly motif (small-scale patterns) and community (large-scale patterns) detection algorithms.

2.3 Delimiting biomes with networks

By abstracting species distributions as a network, we can incorporate complex presence-absence relationships into our delineation, whereas with turnover measures only the percentage of shared species can be included. In the occurrence network, biomes appear as groups of localities and taxa that are highly interconnected. Figure 1B shows the network of native amphibian species and two degree grid cells. In this network, the broad spatial separations of clusters represent the realms, while the biomes are colored differently within each larger cluster. The links that cross between realms correspond to widespread species that inhabit multiple biomes on multiple realms and continents.

To classify biomes we must hierarchically classify groups of species and grid cells into realms and biomes. To do this, we borrow from the techniques developed in network science, where there has been enormous progress to develop fast and accurate clustering algorithms to hierarchically partition groups of nodes into clusters. Among candidate algorithms, the map equation is one approach that can be extended to bipartite networks [138, 139], making it an ideal first choice for biome delineation. The map equation is a general approach that, for our purposes, corresponds to an intuitive process. First, a scientist chooses a random grid cell. She then randomly chooses a species found in that grid cell. She then examines the geographic range of the species, and selects a grid cell at random from that species geographic range. She repeats this process iteratively and exhaustively. In biota with substantial biogeographic structure, she would spend long intervals of time within biomes, crossing only when she selected a cross-biome species.

If she would like to communicate a list of the grid cells and species she chose, it would save her time to simply list the biomes she visited. The map equation optimizes how much information is lost in the compression from a list of grid cells and species to a list of biomes versus the bandwidth saved by communicating a shorter list of visits. The map equation has been extended to deal with hierarchical partitions, which we used for the amphibian data to reveal realms and biomes [138, 139]. The software packages for the two-level and hierarchical approaches are available online (http://www.mapequation.org).

2.3.1 Global amphibian biomes

As a first empirical test case, we applied a network clustering algorithm to the IUCN amphibian range database, which contains range shape files for each amphibian species. Only native ranges were used for the analysis. We chose to analyze distributional data for amphibians [66] because we expected that the eco-physiological tolerance of the c. 6100 included species should be narrower than that for e.g. mammals or birds, and therefore more closely track biome-level regions. Moreover, this would allow a direct comparison with the recent study by [63], where both species distribution data alone and combined with phylogenetic information was used to infer zoogeographic regions and realms on a global level.

2.3.2 United States plant biomes

Our second empirical test was performed using the United States Department of Agriculture (USDA) county-level vascular plant database, which contains the presence or absence of each species in most US counties. These data are ideal as a benchmark because they contain several challenges for computational methods. First, United States county sizes are longitudinally biased, with larger counties in the west and smaller counties in the east. Second, plant distributions are aggregated differently across states, causing systematic compositional biases across state borders. Third, the focus of the dataset is also different between states, with clusters of counties that are poorly sampled. No quantitative delineation of these data are available for direct comparison, so we compared our results with the best recommended method from a recent survey [89].

2.3.3 Similarity approach

To compare with our network delineation, we applied the β_{sim} measure to the data, written $1 - \frac{a}{min(b,c)+a}$. Here a is the number of shared species between two species assemblages and b and c are the total unique species for each quadrat, locality, grid cell, etc. Note β_{sim} is 0 when the species assemblages are either identical or the smaller assemblage is a subset of the larger assemblage, and β_{sim} is 1 when the assemblages contain no shared species.

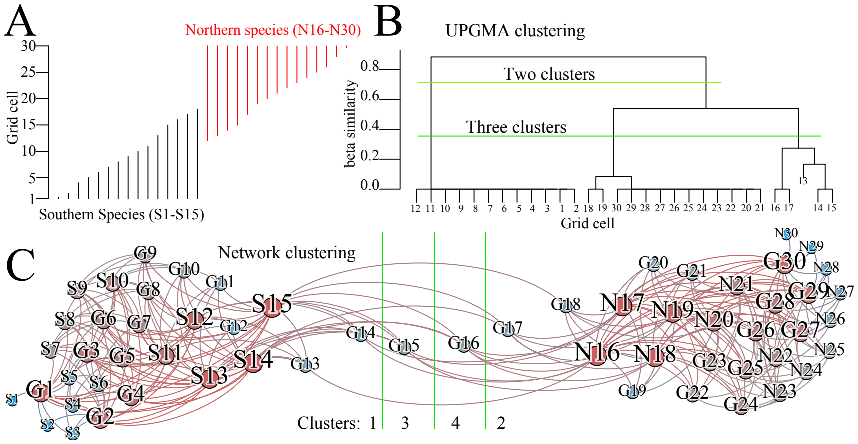

Figure 2.2: **Hypothetical transition zone**. A) Species range data across a line of grid cells. Evident in the data are two faunas which blend together in a transition zone. B) The best representation of the data is as two or three clusters, but three causes the transition zone to appear as a distinct biogeographic region. C) In the network clustering, the best representation is as two or four clusters, with four being optimal (shown). In the optimal partition (four clusters), the transition zone is clustered alone but without any species, revealing it as a zone that does not definitively belong to either fauna. In the two cluster solution, the grid cells are divided evenly between the faunas. The colors indicate the number links each node has - grid cells with higher richness and species with larger ranges are more red, while grid cells with less richness and species with smaller ranges are more blue. The sizes of the nodes are similarly proportional. “G” denotes grid cell, “N” denotes Northern species, and “S” denotes southern species. Node positions were determined with Force Atlas algorithm in the Gephi software package.

2.4 Hypothetical transition zone

We used a recent conceptual dataset to illustrate key differences between the network approach versus the species similarity approach. In a recent commentary [90], this dataset was used to showcase the potential pitfalls for selecting the wrong number of clusters. In their case, this choice either caused or did not cause a transition zone to appear as a distinct biogeographic region, illustrating the danger of classifying transition zones as distinct biogeographic regions. The data hypothetically illustrates a transition zone, where the most widespread species in a Northern and Southern fauna co-occur in a transition zone (Fig. 2A). Using beta-similarity and UPGMA, the transition zone is engulfed by the Northern realm for a choice of two clusters (the data are symmetric, so if the matrix rows are swapped the transition zone is engulfed by the Southern realm), and the transition zone is a distinct cluster if three clusters are chosen (Fig. 2B).

Applying network clustering algorithms (map equation) to the data results in an optimal partition of four clusters: one contains all of the Southern fauna and grid cells 1-14, one contains all of the Northern fauna and grid cells 17-30, while grid cells 15 and 16 each form their own cluster (Fig. 3C). This partition is slightly preferred over the two cluster solution, which evenly cuts the data into two biogeographic zones. This example reveals the benefit of clustering both species and grid cells together, as opposed to clustering grid cells with distances proportional to the number of shared species - grid cells 15 and 16 can easily be identified as transition zones because no species are clustered with them (Fig. 2C).

2.5 Global amphibian biomes

Our analysis identified 10 zoogeographic realms and 55 biome-level regions as an optimal representation of the full amphibian distribution data (Figure 3A). To illustrate how well range limits reflect biome structure, we colored geographic ranges by the region they were assigned (Figure 3B). This differs from the species turnover measure $p\beta$sim approach in Holt *et al.* [63], which identified 19 as optimal. Our analysis identifies both Wallace's and Weber's line, but Weber's line emerges as the boundary between the Oceanian and Oriental faunas. Our approach reveals Sulawesi and the islands between Wallace's and Weber's line as distinct subregions of the Oriental realm.

2.6 Classical United States plant biomes

The United States Department of Agriculture (USDA) plant database comprises X native vascular plants spread throughout Y United States counties. We delineated biome structure for all native plants, native trees, and non-tree native plants.

Species similarity approach. To compare our network-based clustering with a conventional species similarity approach, we opted for the methodology selected as best in a

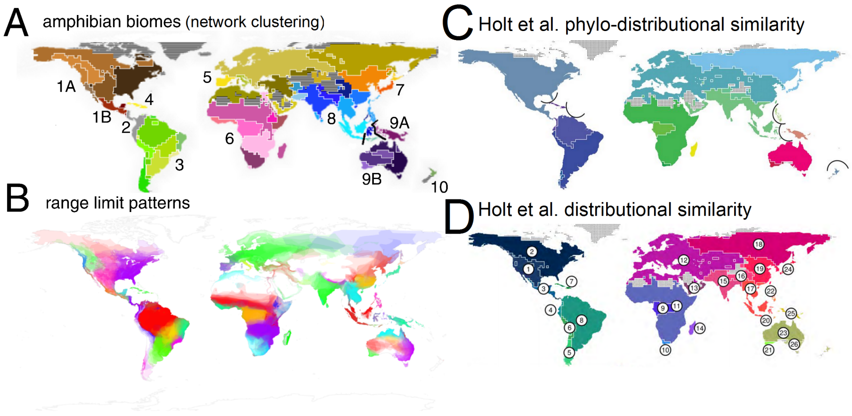

Figure 2.3: **Clustering biomes**. A) Amphibian biogeographic regions of the world determined from geographic range data. Similar colors indicate membership to a higher level realm. Our analysis used two degree grid cells and purged islands that Holt *et al.* [63] excluded, such as Jamaica. B) Species range limits colored by region. Close regions were given opposing colors to highlight boundaries and boundary mixing. Each geographic range polygon was plotted with a low opacity (0.1), from largest to smallest, so that regions with more species would appear brighter. C) Phylo-distributional similarity based approach to determine realms and regions. D) Distributional similarity based approach to determine realms and regions.

recent methods review by Kreft and Jetz [89]. To apply this approach to our data, we created a matrix of counties and computed the beta-similarity between each pair of US counties with species data. This measure is considered ideal over more conventional measures (such as the Jaccard) because it is less sensitive to differences in species richness [89]. We clustered this matrix with the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) approach to generate a hierarchical dendrogram that summarizes the distances between counties. From this dendrogram, we selected an optimum number of clusters by finding the "knee" in the evaluation curve [144], with average percent endemics as our evaluation measure [89]. Applied to our three USDA datasets, the number of clusters selected as optimal was 11 for all native plants, 22 for native trees, and 14 for non-tree native plants.

The resulting optimal partition of counties for all datasets reveals little biogeographic structure. For all native plants, the boundary between the two largest clusters approximates the boundary between the American Great Plains and Eastern Temperate Forests (Fig. 4), but is dominated by rigid state boundaries and fails to distinguish, for example, the Florida everglades, Pacific Coast, or Rocky Mountains. The tree dataset separates the Everglades from the rest of the United States (Fig. 4), and the non-tree dataset mimics the general boundaries in the all plant dataset but contains more clusters that are also US states (Fig. 4).

The identify deeper structure, we also chose to visualize the partitions with 40 clusters selected, though this delineation is not optimal (Fig. 4). Some biogeographic structure becomes apparent at this level - the American Great Plains is cleanly separated from the American West, though this biome unrealistically stretches into the American Southwest desert. The 40 cluster all plant partition is also plagued by US state clusters in the American midwest. In the tree-level data, the Great Plains division is also apparent, as well as a clean separation of the Southwest desert from the American West. In the non-tree dataset, a latitudinal boundary is evident in the Eastern Temperate Forests biome, but also contains ample state-level biases.

Network approach. We generated a network dataset from the USDA plant data, with county nodes connected to species nodes if the species was identified as (natively) present in that county. From these data, we clustered the map equation. Our analysis revealed little hiearchical structure in the data, so we opted to use a two-level implementation of the map equation, which produces k clusters instead of hierarchically nested groups of clusters [138]. The apparent lack of hierarchy in the data is likely an issue of large grain and low scale (counties within a single country) - higher resolution data, such as a database produced from geographic coordinates, might produce greater subdivision in the North American Great Plains, for example. Applied to our three USDA datasets, the number of clusters selected as optimal by the map equation was 25 for all native plants, 19 for native trees, and 16 for non-tree native plants. Because the algorithm that seeks the best partition is stochastic, we ran the algorithm many times and selected the partition that minimized the scoring function in the map equation.

Broad similarities are evident across the all native plants, native trees, and non-tree native datasets (Fig. 4). In all partitions, Southern Texas is determined as its own biome, while the Everglades is only evident from the tree-only dataset. The West Coast forms its own biome for the all plant data and the non-tree data, but the Pacific Northwest is omitted from this biome when only trees are considered. The Southern deserts of Arizona and surrounding area follow some rigid state boundaries, suggesting that large county sizes in the area obscure finer demarcation. In the American midwest, the American Great Plains appear much smaller when only trees are considered. State-level biases are evident in Lousiana for the native tree data, but not for the other two datasets.

2.7 Discussion

2.7.1 Global amphibian biomes

The differences in number of zoogeographic realms and biomes found in our study as compared to [63] do not arise from a lower cutoff threshold for our approach, because we followed their procedure for merging regions with less than 10 grid cells into the closest regions [63]. Rather, we interpret this difference as stemming from a fundamental difference in methodology – ours clusters patterns of presence-absence relationships while theirs identifies clusters of grid cells with low distributional and phylo-distributional turnover.

Our results suggest that, at least for amphibians, distributional turnover is sufficient to identify the realm boundaries. This conforms with the distribution-only approach undertaken by Holt *et al.* [63], which also distinguishes Weber's line for amphibians as the realm boundary but does not identify Wallace's line. Our analysis supports the idea that Weber's line is insensitive to methodology or use of strictly distribution data, whereas Wallace's line might be visible because of our network approach.

Our analysis was able to recover the majority of biome-scale regions around the world. Taking South America as an example, our analysis not only identified the 23 major regions found by [63], but also successfully recovered climatically and physiognomically distinct biomes such as the seasonally dry and fire-prone Brazilian Cerrado, the evergreen Atlantic forest of eastern Brazil, and the geologically old and genetically isolated Guianan highlands, among several others regions that were not unveiled by [63].

It is important to note that the method-dependent differences outlined here are not only in terms of resolution, i.e. the total number of regions identified, but also in the actual boundaries of regions. The western limits of the Amazonian region inferred by [63], for instance, spans across the Andean mountains, despite the enormous altitudinal and physiological differences between these two regions. Our delimitations better conform to the commonly recognised boundaries between the Andes and Amazonia, thus reflecting not only current topography and climate but also evolutionary history [64].

2.7.2 United States plant biomes

The failings of the species similarity approach to distinguish more than a few biogeographic features per partition is perhaps unsurprising given a few aspects of the task, which we chose to illustrate the vulnerabilities of species similarity clustering. First, we clustered the raw occurrence data (presence or absence of a species in a county), purposefully avoiding the construction of species geographic ranges given niche data. The method of aggregation in the database becomes aptly apparent when this is done - presence/absence data is often compiled at the state level rather than the county level, which produces apparently unique flora at the state level. Second, as previously noted, county sizes differ substantially, revealing a richness bias that is correlated with county size, a pitfall that could have been avoided by aggregating data by grid cell (but perhaps minimized by the beta similarity measure).

It is therefore prudent to question why we would avoid standard practice, when facilities are easily available to avoid these pitfalls. First, we note these pitfalls also existed for the network method, which managed to identify the majority of biome-level features despite them. Second, we argue that new methods for biogeographic delineation must be designed around these barriers. Ample reward is available for clustering raw presence-absence data with localities with unequal sizes:

1. *Projected geographic range circularity.* If we are to understand how boundaries between biomes evolve we cannot use predicted species distributions to delineate our biomes. Though precipitation and temperature predicts plant distributions very well, these factors need to be identified post-hoc rather than built in, or we risk a circularity: our biomes are determined by precipitation and temperature because we predicted

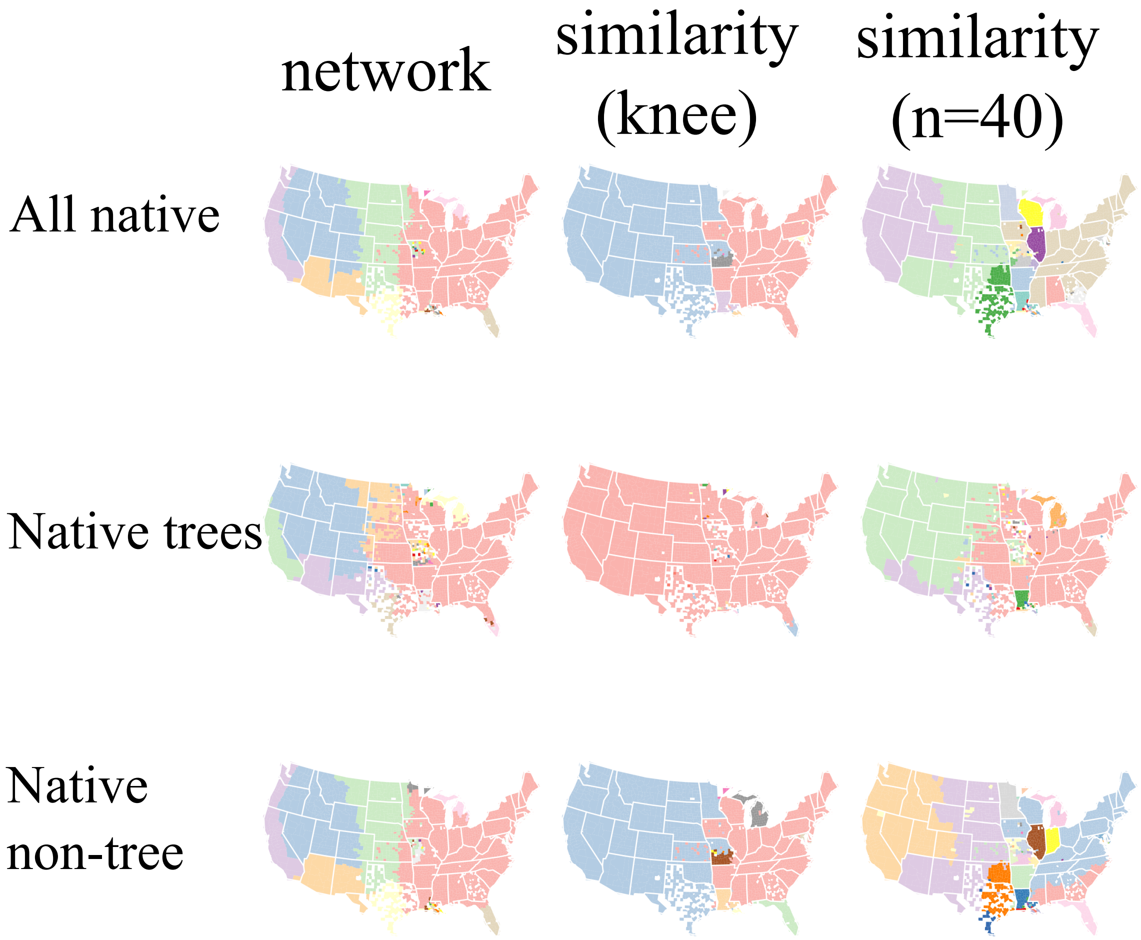

Figure 2.4: **United States plant biomes**. Demarcations of american biomes for three subsets of the USDA plant database: all native plants, native trees, and non-native trees. Column one in the data was determined by the map equation (optimal number of clusters shown), while column two and three in the data were determined by a similarity approach (optimal number of clusters shown as well as a finer scale delineation). Overall, the network approach captures with broad brushtrokes the general patterns of the United States biomes, while the similarity approach captures some of these patterns but is overcome by state-level biases.

plant ranges with precipitation and temperature.

2. *Many species cannot be modeled.* Nearly every terrestrial biogeogaphic delineation

published deals with mammals, amphibians, or birds, for which high resolution geographic range data is available. Yet flora creates the habitat for these animals to inhabit. Rare species in tropical plant databases often have just one or two occurrences, preventing the projection of their ranges despite that these rare species may be the most vulnerable to anthropogenic pressure.

3. *Collector's bias.* Global scale biogeographic analyses require an enormous amount of data aggregated together from thousands of sources. One critical challenge of biogeography for many clades (such as fungus, plants, and insects) is the fact that these data are collected by different scientists with potentially nonstandard nomenclature and differing areas of interest, causing the species assemblages to have different probabilities of presence/absence across species for different collectors even within the same locality. Global datasets, assembled from expert opinions, will therefore have built-in richness and assemblage biases, similar to the state-aggregation issue encountered with the USDA plant data.

4. *Biogeography over macroevolutionary time.* Mapping changing biogeography across the Phanerozoic is a union of biogeography and paleontology that holds great promise for our understanding of the forces that determine macroecological patterns and forge biogeographic boundaries over long timescales. Though our ability to do this is improving as the Paleobiology Database grows (c. 45,000 references, 275,000 taxa, 1.1 million occurrences, and 150,000 fossil collections), a primary limitation is the fossilization process; fossil basins with greater rock area will tend to produce species/genus assemblages with greater richness. Equal-area grid cells that are created from these fossil basin assemblages will then have an inbuilt area bias because the total fossiliferous area is not equal across grid cells.

The USDA plant database is therefore perhaps an unfair test to compare species similarity methods with network methods, but we argue that it is an essential test because newer methods must be able to cluster unaltered presence-absence data. We argue that the

USDA plant database has well-understood biases that make it an excellent benchmark for future work.

2.8 Conclusions

Biome conservatism has been suggested as a potentially crucial feature shaping the uneven distribution of the world's biota [37, 19, 174], including the establishment and maintenance of the tropical gradient in species richness. The evolution of entire biomes is also gaining focus in macroecological meta-analyses using phylogenetic and distribution data [31, 64, 63].

A critical component of this new research line is the robust classification and delineation of global biomes, which has been limited by our ability to quantify reproducible boundary lines between biomes. Our method demonstrates that distributional data holds the potential to achieve this goal. Phylogenies (especially when time-calibrated) can subsequentely be used to shed light on the temporal origin, evolution, and phylogenetic relatedness of biomes. To make future advances in poorly studied taxonomic groups and areas requires biogeographic methods that work for unequal locality sizes and do not require predicted range distributions.

More than a century after the first biogeographic regions were proposed [170], we may now have the data to delimit the world's realms and biomes in greater detail than Wallace could ever envision. Our study however illustrates that new methodology may play a vital role in this process, and that network methods offer biogeography a new set of exciting tools to classify and delimit biodiversity.

We thank J. Grummer for discussions on amphibian distributional patterns. We thank C.T. Bergstrom, M. Rosvall, and F. Meacham for discussions of network clustering. A.A. is supported by grants from the Swedish and European Research Councils and D.A.V. by a WRF-Hall fellowship.

Chapter 3

GLOBAL DISTURBANCES FACILITATE BIOME SHIFTS

Daril Vilhena, Elisha Harris, Carl Bergstrom, Max Maliska, Peter Ward, Christian Sidor, Caroline Strömberg, Gregory Wilson

Biogeographic patterns of survival help constrain causal factors of mass extinction. To test whether biogeography influenced end-Cretaceous (K-Pg) extinction patterns, we used a network approach to delimit biome-level units (modules) in a global Maastrichtian database of 329 bivalve genera. Geographic range is thought to buffer taxa from extinction, but the number of modules a taxon occurred in superseded geographic range as an extinction predictor. This suggests that more range is not necessarily better if it is within rather than between biomes. Contrary to former results that found no latitudinal pattern, we found a latitudinal selectivity gradient in the K-Pg, such that higher latitude modules had lower extinction than expected given the geographic ranges of the genera, implying that (i) high latitude biomes were more resistant to extinction, (ii) the intensity of the K-Pg kill mechanism declined with distance from the tropics, or (iii) both. Our results highlight the importance of macroecological structure in constraining causal mechanisms of extinction and estimating extinction risk of taxa.

3.1 Introduction

Mass extinctions have disproportionately shaped the evolutionary history of life [44]. During these rare, geologically rapid events, the rules of selectivity that prevail in background extinctions do not always clearly apply [72]. To delimit what may be independent selection processes and constrain possible causal factors, paleontologists have sought biologically meaningful patterns of survivorship in mass extinctions [125, 175]. In the marine realm, there are some taxon-specific cases where survivorship appears nonrandom and is linked to

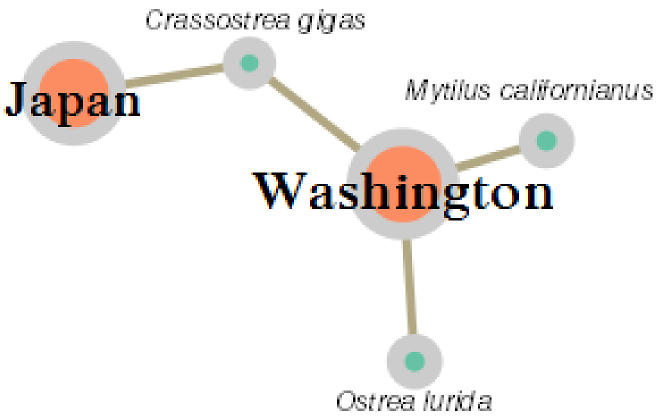

Figure 3.1: A bipartite occurrence network. *Ostrea lurida*, *Mytilus californianus*, and *Crassostrea gigas* each have a second order relationship with each other (co-occurrence). Japan and Washington have a single second order relationship (shared *Crassostrea gigas*). Both *Ostrea lurida* and *Mytilus californianus* have a single third order relationship with Japan. This does not imply that *Ostrea lurida*, for example, could occur in Japan, but more third order relationships than we would expect due to chance with Japan is evidence for occurrence potential, or depauperate fossilization.

ecological traits – for example, in the K-Pg mass extinction event, reliance on photosymbiosis among scleractinian corals severely reduced survivorship [86] and sea urchin feeding strategy correlates positively with survivorship [153]. However, overall, selectivity in mass extinctions often appears indifferent to ecology and has instead been described as “nonconstructive” [125]. The most prevalent pattern of this nonconstructive selectivity is the correlation of survivorship and geographic range above the species level [73].

Here we employ network methods to test whether biogeography and provincialism [58, 164, 131, 165], which relates to ecology, environment, and evolutionary history, affected survivorship among bivalve genera in the K-Pg mass extinction event. Network methods have been useful in a broad range of applications, for example, to model the transmission of disease in social networks [109], to describe the structure of scholarly communication [138], and to model the stability of ecosystems in response to extinction [135]. A network approach can also reveal spatial patterns of taxa from geographic range data [9]. Here, we introduce bipartite occurrence networks, which contain both localities and taxa as nodes (Fig. 1).

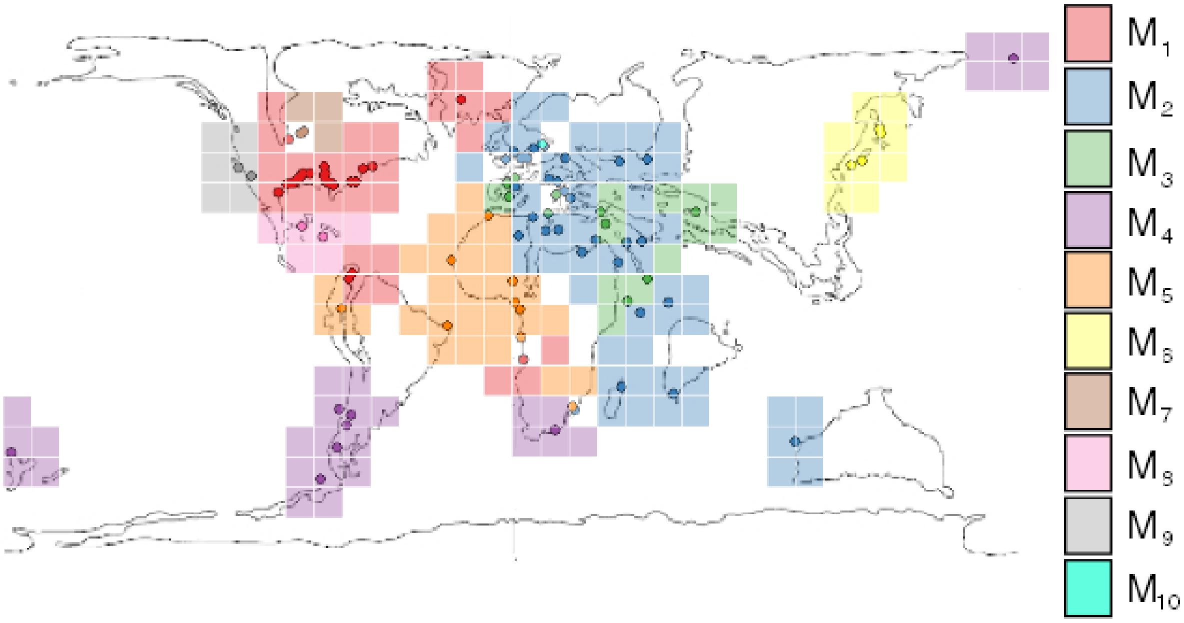

Figure 3.2: The modular structure of bivalves reveals spatial organization. Points correspond to fossil localities, and are colored by module. For visualization, ten-by-ten degree cells were colored by module only if the cell contained fossil localities from a single module, and cells without fossil localities were colored if they were less than 15 degrees from a locality. Although some cells may have been uninhabitable by marine bivalves, they were colored if they met the above criteria. Grids are only for visual aid and were not included in any way in subsequent analysis.

The links in this network (connections between nodes) are occurrences. This network has convenient higher-order biological properties. The set of localities a taxon links to is its geographic range, while the number of taxa a locality links to is its richness. A pair of nodes that are two links away from each other have a "second order relationship." For pairs of taxa, the number of second order relationships is the number of co-occurrences, while for localities, the number of second order relationships is the number of shared taxa (note that a second order relationship cannot exist between a locality and a taxon). Third order relationships in this network are between taxon and locality, and appear when a taxon is connected to a locality through an intermediary taxon and locality (Fig. 1).

Classical approaches for biogeographic analysis of occurrence data, such as ordination or agglomerative clustering of a locality-locality distance matrix, use only second order information [89]. Specifically, one cannot reconstruct the geographic range of a taxon from the distance matrix used for analysis. A network approach allows one to integrate geographic ranges, co-occurrence (taxon-taxon), shared taxa (locality-locality), and higher order relationships. The higher order relationships can help biogeographic analysis recover from competitive exclusion or taphonomy – for example, aragonitic and calcitic shells may not be preserved together in entire biogeographic regions [176, 28, 80]. For smaller spatial scales, such as in a Pacific Northwest intertidal ecosystem, goose barnacles may not occur in the same plot as california mussels, yet we would like them grouped within the same biogeographic structure (intertidal strip). A network approach can therefore applied to assemblage data collected at any spatial or taxonomic resolution, and has the added advantage that no dissimilarity measure is required for cluster analysis of the network [89].

The K-Pg event (ca. 66 Ma) is an ideal case to test whether biogeography influences survival in mass extinctions. It was geologically abrupt and was associated with significant changes to marine productivity and ocean chemistry, dramatic restructuring of marine and terrestrial communities, long-term effects on evolutionary rates and biogeography [146, 91], and the extinction of up to 76% of all species [69]. As the most recent of the "big five" mass extinctions, the quantity, quality, and spatial resolution of the geological and paleontological data for the K-Pg interval are also better than those available for more ancient mass extinctions [76]. Bivalves have emerged as a model system for examining macroevolutionary phenomena due to their excellent preservational record [76], deep evolutionary history, spatial ubiquity, and the significant effort dedicated to standardizing their systematics [128, 73]. The bivalve dataset that was used in this study was downloaded from the Paleobiology Database (PBDB, [33]) and consists of 3,445 occurrences of 329 bivalve genera from 105 Maastricthian faunal assemblages [128]. This taxonomically standardized and globally representative dataset is the same one that was previously used to show a geographically uniform pattern of extinction across the K-Pg boundary [128].

3.2 Biome structure in the late Cretaceous

We generated a Maastrichtian network from the bivalve dataset, and used PBDB stratigraphic ranges to determine which genera survived the K-Pg mass extinction event [33].

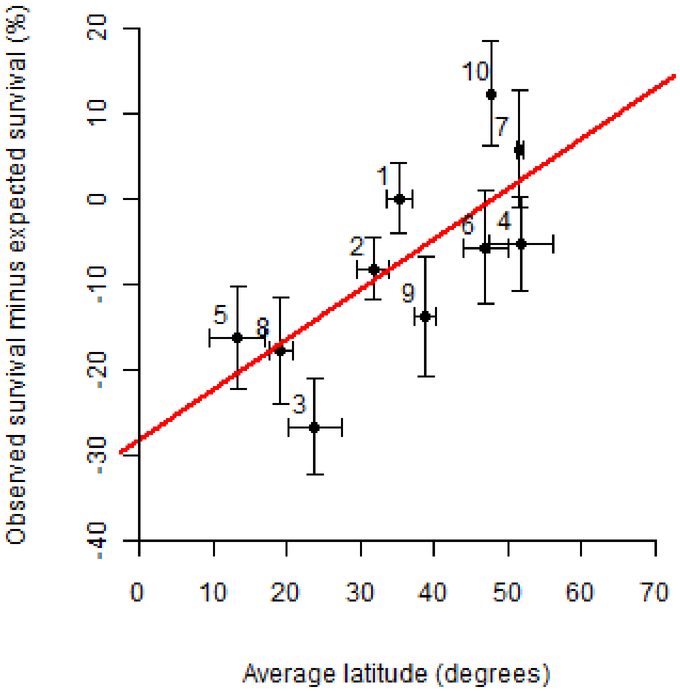

Figure 3.3: Observed survivorship minus expected survivorship, under a model that predicts survivorship solely based on geographic range, correlates with latitude (Regression with unequal variances, $P = 0.02$, horizontal and vertical standard error shown). Above zero indicates survivorship greater than expected given geographic range. Module numbers are shown, and rudist modules are included. Average latitude of each module was determined from the absolute latitudes of the localities in each module. This result is robust to changes of the model parameter, global extinction risk, use of median instead of average latitude, and choice of regression analysis (equal or unequal variance). Inoceramids are excluded here, but the result is not sensitive to their exclusion. If rudists are excluded the P-value increases slightly ($P = 0.03$).

We used network-based geographic delineation approach [138], which incorporates all orders of relationships to reveal biogeographic units. The ten biogeographic units (modules) we identified have boundaries that are naturally delimited by the major patterns of geographic ranges of taxa, and reflect sudden geographic transitions in the biota. The sampling intensity (number of localities per biogeographic region) do not affect module delineations unless poorly sampled regions contain little to no endemic taxa, which prevent our identification of a geographic transition. Australia is perhaps the best example of this – our analysis groups the single Australian locality with the mostly European non-rudist module (M_2). However, the Australian locality contains 17 taxa, 15 of which are found on average in five other modules. This suggests that, at least given the data genus-level data available, Australia had bivalve fauna that extended down from Europian shorelines.

Overall, while modules are biogeographic, they are not necessarily limited to localized continental shorelines (Fig. 2), at least at the genus level. Modules M_1 and M_2 are the best sampled and are distributed mainly in the North American Gulf and southern Europe. Modules M_3 and M_8 are characterized by tropical rudist bivalves (Order Hippuritoida) and are found in the Eastern and Western Hemispheres, respectively. Module M_4 is composed of several high latitude clam families, and is distributed across South Africa, South America, New Zealand, and the Asian-Alaska land bridge. Module M_5 is located in both central South America and West Africa. Module M_6, distributed along the east Asian coastline, is a provincial module with several Asian endemics. Module M_7 and M_9 are provincial and found along North American coastlines, but both contain several genera with large geographic ranges. Module M_{10} is small and tied to the European shoreline; it comprises a single locality with many endemics, and has the lowest extinction percentage.

3.3 Using biome structure to predict extinction vulnerability

The demarcation of module boundaries has predictive power for per-taxon extinction. Previous work has indicated that geographic range is the best predictor of genus-level survivorship in both background and mass extinction [117, 73], and presumably buffers taxa from extinction, perhaps because it is correlated with environmental breadth [60]. However, we found that geographic range recapitulates module range, or number of modules a taxon is distributed in. A logistic regression with module range alone was the best model to predict K-Pg survivorship (AIC=419.74, $P = 10^{-8}$), compared with a regression with both (AIC=421.67, $P = 0.78$ for geographic range, $P = 0.015$ for module range), and geographic range alone (AIC=425.8, $P = 10^{-7}$). This finding suggests that, at least for bivalves in the end-Cretaceous event, any amount of geographic range was equivalent within the demarcation of the module, while cross-module taxa were more buffered from extinction. These results are robust to the inclusion or removal (any combination) of inoceramid and rudist bivalves, the first whose extinction preceded the K-Pg boundary, and the second whose extinction might be tied to a physiological factor [[159], but see [73]].

To compare extinction percentage across modules requires an adjustment for the correlation of geographic range with extinction probability, because differences in distributions of geographic ranges will bias observed per-module extinction percentage. To correct this, we estimated expected per-module extinction percentage given the per-module distribution of geographic ranges [117], and analyzed the difference between expected per-module extinction percentage given geographic range and observed per-module extinction percentage. The entire assemblage of bivalves that occurred in each module was included in the calculation of per-module extinction unless otherwise noted.

After adjusting the per-module extinction percentage, we examined how extinction percentage varied by module geography. Whereas Raup and Jablonski [128] used nine geographic regions to infer that K-Pg extinction intensity was globally homogeneous, we used geographic regions determined by the modular structure of the data to show that modules had differential extinction across the Earth (Fig. 2B). This adjusted extinction percentage did not vary along a longitudinal gradient or with distance to either the bolide impact at Chixculub, Mexico or the Deccan flood basalt volcanism of peninsular India ($P > 0.05$). However, detected a latitudinal gradient of adjusted extinction percentage, such that modules with higher average latitudes had higher survival than expected given the geographic ranges of the genera that occurred in those modules (Fig. 3, $P = 0.02$, $R^2 = 0.495$). This result is robust to the inclusion or exclusion of both rudist and inoceramid bivalves, but the figure shown has inoceramids excluded.

3.4 Discussion

Our study refines the debate regarding the major causal hypotheses for the K-Pg mass extinction. Some researchers have suggested that the bolide impact alone triggered a throng of secondary effects, for example, a brief but intense global thermal pulse [132] and a dust cloud that inhibited photosynthesis [121], which together would have led to catastrophic extinction cascades. Others have contended that additional events, such as flood basalt volcanism in India that released massive amounts of sulfur and carbon dioxide and resulted in severe environmental perturbations, combined with the bolide impact to cause the K-Pg mass extinction. Our study does not reject these hypotheses, but further constrains the search space of the proximal causal agents. Our results imply that either the severity of the K-Pg kill mechanism declined with distance from the tropics, higher latitude modules were more resistant to extinction, or both. In turn, proposed causal agents and scenarios must either match the decline of latitudinal kill mechanism severity with decreased geographic range selectivity at the module level, or provide paleoecological evidence for decreased ecosystem extinction risk with latitude. Although neoecological evidence suggests that ecosystems at higher latitudes have less extinction risk [166], presumably because of wider abiotic tolerances, we cannot be certain that this relationship applies through geologic time given differences in latitudinal diversity gradients and configurations of continents. Analyses of biome-level extinction risk in background intervals are needed to calibrate the effect of latitude on ecosystem extinction risk throughout the Phanerozoic.

Previous studies that support geographically uniform extinction risk in the K-Pg extinction event have treated each taxon as statistically independent [128]. Our analysis, which uses the same data but takes the modular structure of taxon distribution into consideration, does not support such a scenario of uniform extinction risk. Rather, it indicates that taxa must be studied in the context of the biome in which they reside. In other words, the extinction vulnerability of organisms during the K-Pg was biogeographically coupled. Modular selectivity, or a difference between module vulnerability, in the K-Pg mass extinction has theoretical ramifications as well. If modules have differential extinction, then invasives from modules with lower extinction might displace modules with higher extinction as ecospace opens [20], effectively acting as colonizers for the module. Given the spatial complexity of the K-Pg recovery [73], the ability of modules to displace one another through biological invasions is a possibility. Our study underscores that the macroecological context of taxa should not be ignored because extinction vulnerability is inextricably coupled to biogeographic history.

We thank David Jablonski for guidance with the Paleobiology Database Maastrichtian invertebrates. We benefited from a methods discussion with Charles Marshall, and comments from Douglas Erwin. We are grateful to an anonymous reviewer for a critical insight. Additionally, we thank David Jablonski and David Raup for making their bivalve dataset publicly available on the Paleobiology Database.

3.5 Materials and Methods

3.5.1 Paleobiology Database download

We downloaded the Maastrichtian Raup and Jablonski dataset from the Paleobiology Database on March 21, 2012, which contains 3,445 occurrences of 329 bivalve genera in 105 assemblages [33]. These assemblages are basin-level resolution from the Maastrichtean epoch. We chose this dataset because it is spatially well sampled and it is taxonomically standardized. We could not use all Maastrichtian marine invertebrates (or even bivalves) from the Paleobiology Database because the majority of the taxa are from a USGS data dump. These data cannot be used because they are spatially uneven (Gulf bias) and taxonomically inconsistent with the rest of the database (causing duplicates of many genera). Additionally, we opted to use this dataset to make our results more comparable to those of Raup and Jablonski [128].

3.5.2 Determining survivors

Survivors were determined from the Paleobiology Database standard stratigraphic range intervals [33]. We could not use the Sepkoski compendium [148] because it lacked range interval data for 68 genera.

3.5.3 Data availability and software

Data not immediately accessible in the Paleobiology Database are available in the public repository Dryad (http://www.datadryad.org). This includes the bivalve network, module assignments, and the geographic ranges of the taxa. To infer geographic ranges, we used the province-counting approach outlined in Jablonski and Raup [75]. Their approach used the biogeographic provinces in the Atlas of Palaeobiogeography [55]. We created shape files for these provinces and used the R package **sp** to infer geographic range size. Though our analysis generated a newer map of the biogeographic provinces of Maastrichtian bivalves, we preferred to use the original methodology of Jablonski and Raup [75] to avoid bias in our results.

3.5.4 A network approach for biogeography

The taxa-locality matrix $\mathbf{M}$, or occurrence matrix, is a representation of the occurrence distributions of taxa (any level) across localities (any spatial scale, from plots to counties to provinces). Each entry in the matrix is either one (taxon present in locality) or zero (absent)

$$M_{ij} = \begin{cases} 1 & \text{if taxon } j \text{ is in locality } i \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

This matrix is the basic representation of a bipartite network, where there are two types of nodes: taxa and localities (Fig. 1). The links (occurrence relationships) in this bipartite network are exclusively between localities and taxa, taxa cannot be linked to taxa, and localities cannot be linked to localities.

The taxa in this bipartite network will be connected to one another in complex patterns through the localities they occupy, making it difficult to interpret the information in the network without first isolating the major patterns. To find biogeographic clusters in the network we wish to study, we must identify these patterns. As biologists, we would like to identify the boundaries between biogeographic units, which will be encoded as major topological features in the network. Moreover, we would like to capture most of the information

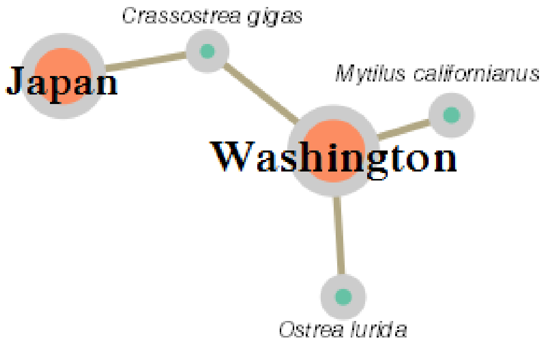

Figure 3.4: An example bipartite network. Large red nodes are localities and small green nodes are bivalve species from the Pacific Ocean. Here, Crassostrea gigas links the two localities together.

about specific relationships with broad brushstrokes.

Network community detection is a methodological process that does exactly this: identifies the major topological features of networks [46]. Though many algorithms have been proposed to do this task [93], the *map equation* is an excellent candidate for biogeography because of its accuracy [93].

The map equation is the theoretical limit of a description length of a random walk on the network, where nodes are aggregated into community structures. The following story provides some intuition about this process. A scientist looks at a random locality. She then randomly chooses a taxon found at that locality. Next, she randomly selects a locality that the taxa she chose is found in. She now chooses another taxon from the new locality, and repeats this process forever. In a network with emergent biogeographic features, she will likely spend long bouts of this process within biogeographic units, which we will refer to as modules. Specifically, she can only switch to a new biogeographic unit when she chooses a taxon that is not endemic to the biogeographic unit she is currently in. If she would like to communicate a list of the localities and taxa she chose, it would save her time to communicate a list of the major biogeographic units (which contain taxa and localities) she visited, rather than each individual locality and taxon.

Here we write the basic form of the map equation in the notation of the occurrence matrix. However, for more in-depth discussion and derivation, we refer the reader to a map equation tutorial [137], or the original paper [138].

This bipartite network is unweighted and undirected, all links are equal, and each link is symmetric. We refer to the total localities in the network is A, and the total number of taxa as T. In undirected networks, the probability that the scientist visits any node (locality or taxa) in her infinite random walk is the number of links that node has, divided by two times the number of links in the entire network. We refer to the number of links multiplied by 2 in the network as L, defined as

$$L = 2 \sum_{i=1}^{A} \sum_{j=1}^{T} M_{ij} \, . \tag{3.2}$$

We multiply the number of links by two because each link is symmetric, so we need to count each link twice. The probability that she visits locality i in her random walk is therefore

$$p_A^i = \frac{\sum_{j=1}^{T} M_{ij}}{L} \, , \tag{3.3}$$

while the probability she visits taxon j is

$$p_T^j = \frac{\sum_{i=1}^{A} M_{ij}}{L} \, . \tag{3.4}$$

As cartographers, we would like the scientist to convey her random walks as concisely as possible. This is an optimization problem - out of all possible partitions, we must choose the one that allows her to convey her random walk the most concisely. A partition is a division of the network into modules, such that each node is uniquely assigned to a module. The best partition will be the optimal compression of the major patterns in the bipartite network. To evaluate these partitions, we must express the probability that she switches between modules. The probability that she leaves module m is the number of links that lead from nodes in module m to nodes outside module m, written n_e^m and divided by the total number of links

$$p_e^m = \frac{n_e^m}{L} \, , \tag{3.5}$$

while the probability that she stays inside module m is

$$p_s^m = \frac{2n_s^m}{L}\,, \tag{3.6}$$

where n_s^m is the total number of links between nodes in module m, multiplied by 2 because each link must be counted twice. These probabilities are sufficient to express the extended form of the map equation [137]

$$\begin{aligned} L(M) = \Bigg(& \left(\sum_m p_e^m\right) \log \left(\sum_m p_e^m\right) \\ & -2\sum_m \left(p_e^m\right) \log \left(p_e^m\right) \\ & -\sum_m \left(p_e^m + p_s^m\right) \log \left(p_e^m + p_s^m\right) \\ & -\left(\sum_{j=1}^{T} p_T^j \mathrm{log} p_T^j + \sum_{i=1}^{A} p_A^i \mathrm{log} p_A^i\right)\Bigg) \end{aligned} \tag{3.7}$$

where $L(M)$ is the amount of information required to convey an infinite random walk on partition M. All logarithms listed are in base 2, because information is measured in units of bits [149]. The fourth term is independent of the partition M, while the first three terms change based on the proposed partition. To seek the best partition among many, we used an algorithm that is freely available online at (http://www.tp.umu.se/ rosvall/code.html). An applet that illustrates the concept behind the map equation is available online (http://www.mapequation.org).

3.5.5 Partition robustness

The algorithm that minimizes the map equation returned largely equivalent partitions across random seeds (7.8-7.92 bits, 1.3-1.33% compression). Choice of partition did not change our results.

3.5.6 Geographic range null model

We use a modification of the approach outlined in Payne and Finnegan [117] to test the relative magnitude of geographic range selectivity between modules. What follows is a "fields of bullets" null model, where each genus is exposed to a globally homogeneous but province-level kill probability, Pr(die). We use this approach because it generates a correlation between survivorship and geographic range. In this model, the genera that are in more provinces have more chances to survive an indiscriminate culling. Such an assumption generates a correlation between geographic range and survivorship. The probability that a genus goes extinct is the probability that it dies in every province

$$\mathrm{Pr}(\mathrm{extinction}|n_i) = \mathrm{Pr}(\mathrm{die})^{n_i}, \tag{3.8}$$

where n_i is the number of provinces genus i is in. The probability that a genus survives is binomially distributed, because it can survive if it evades the kill probability once or more in any combination of regions

$$\begin{aligned} \mathrm{Pr}(\mathrm{survive}|n_i) &= \sum_{k=0}^{n_i-1} \binom{n_i}{k} \mathrm{Pr}(\mathrm{live})^{n_i-k}\mathrm{Pr}(\mathrm{die})^{k} & (3.9)\\ &= 1 - \mathrm{Pr}(\mathrm{extinction}|n_i). & (3.10) \end{aligned}$$

Here, the sum over k is iterated to include the probability of all combinations of local extinction in k geographic regions. The null model is parametrized by Pr(die), therefore the value must be inferred from data. We use the maximum likelihood estimate of $\hat{\mathrm{Pr}}(\mathrm{die})$ given the observed survivorship data

$$\hat{\mathrm{Pr}}(\mathrm{die}) = \underset{\mathrm{Pr(die)}}{\mathrm{argmin}} \left(\sum_i (X_i - (1 - \mathrm{Pr}(\mathrm{die})^{n_i}))^2 \right), \tag{3.11}$$

where X_i is the survival outcome of a genus, one for survival, zero for extinction. Here, argmin is used to indicate a minimization procedure across Pr(die), such that the Pr(die) with the least squared error is chosen. Caution should be used when applying this maximum likelihood estimate to data that do not have many broadly distributed taxa, because although it is the maximum likelihood, if too few broadly distributed taxa are in the data the outcome may be overfit to those few broadly distributed taxa. However, the Raup and Jablonski dataset has taxa from all geographic range levels, so the model is not overfit. The maximum likelihood estimate $\hat{\Pr}(\text{die})$ for the Raup and Jablonski dataset was 0.805. While changes in this estimate could change the expectation of the model, our latitudinal gradient result is robust to changes in this probability. The observed survival percentage for a group of taxa is therefore

$$Y = \frac{1}{m}\sum_i X_i, \tag{3.12}$$

where m is the total number of taxa under consideration. The survival percentage Y is a random variable, and its expected value can be calculated from the geographic ranges of the taxa in the group under consideration (for us, a module). The random variable Y will be normally distributed when the number of taxa is large enough, allowing us to use the standard normal distribution function to calculate the probability that the observed extinction percentage was generated by the null model – which assumes a province-level kill probability that is the same across provinces. The expected value of Y is

$$\begin{aligned} \mathrm{E}[Y] &= \frac{1}{m}\sum_i \mathrm{E}[X_i] && (3.13)\\ &= \frac{1}{m}\sum_i (1 - \hat{\Pr}(\text{die})^{n_i}), && (3.14) \end{aligned}$$

while the variance is

$$\begin{aligned} \mathrm{Var}(Y) &= \sum_i \mathrm{Var}\left(\frac{1}{m}X_i\right) && (3.15)\\ &= \frac{1}{m^2}\sum_i (1 - \hat{\Pr}(\text{die})^{n_i})(\hat{\Pr}(\text{die})^{n_i}). && (3.16) \end{aligned}$$

In the main text, we used this geographic range null model to compare the observed module extinction percentage to the expected module extinction percentage. The error bars for Figure 3 from the main text were calculated assuming that the observed module extinction percentage was binomially distributed (the variance above plus the observed binomial variance).

Chapter 4

BIOMES OVER TIME: BIASES, DATA, AND CHALLENGES

Daril Vilhena, Andrew Smith

Inference of past and present global biodiversity requires enough global data to distinguish biological pattern from sampling artifact. Pertinently, many studies have exposed correlated relationships between richness and sampling in the fossil record, and methods to circumvent these biases have been proposed. Yet, these studies often ignore paleobiogeography, which is undeniably a critical component of ancient global diversity. Alarmingly, our global analysis of 481,613 marine fossils spread throughout the Phanerozoic reveals that where localities are and how intensively they have been sampled almost completely determines empirical spatial patterns of richness, suggesting no separation of biological pattern from sampling pattern. To overcome this, we analyze diversity using occurrence records drawn from two discrete paleolatitudinal bands which cover the bulk of the fossil data. After correcting the data for sampling bias, we find that these two bands have similar patterns of richness despite markedly different spatial coverage. Our findings suggest that i) long-term diversity trends result from large-scale tectonic evolution of the planet, ii) short-term diversity trends are region-specific, and iii) paleodiversity studies must constrain their analyses to well-sampled regions to uncover patterns not driven by sampling.

4.1 Introduction

Patterns of biogeography, latitudinal diversity gradients, macroecology, and macroevolution result from biological processes constrained by the configurations of continents and earth processes [23]. Paleobiologists interested in the detection of these patterns and the processes that gave rise to them must first overcome unevenness and inconsistency of spatiotemporal sampling in the fossil record [163, 155]. This problem is manifested in various, often correlated ways: entire biogeographical regions without aragonitic shells [28, 27], uneven sampling of the latitudinal diversity gradient [81], inter and intra-regional variation in rock amount and quality [152, 79, 106], spatiotemporal differences in sampling effort and taxonomic identification [111, 84, 11, 3], and cross-regional differences of preserved sedimentary environments and habitats [120, 30]. These biases can lead to erroneous results without proper precautions [17].

One important factor, largely overlooked to date, is the geographical distribution of sampling effort. As first noted by Allison & Briggs [2], if the spatial distribution of fossil taxa changes between time bins, how can we reject the hypothesis that observed changes in recorded diversity are due to shifts in spatial sampling patterns? A quantitative assessment of the spatial pattern of paleontological sampling and sampled taxonomic richness is therefore an essential first step for any large-scale paleobiodiversity analysis. Here we first document the biogeographical distribution of paleontological sampling and sampled taxonomic richness through the Phanerozoic and establish their strong covariance. We do this using the best available database of fossil occurrences, the Paleobiology Database (www.paleodb.org), which has been the preferred source of data for global Phanerozoic biodiversity analyses for over ten years [6, 3, 155]. Next, we demonstrate that the paleolatitudinal distribution of paleontological sampling and taxonomic richness does shift significantly over geological time, revealing time intervals of the rock record where geographical bias needs to be accounted for. Finally we apply modeling to see if we can establish how marine invertebrate diversity has changed over the Phanerozoic within fixed paleolatitudinal strips, controlling for sampling biases.

Overall, our study underscores the need to develop new subsampling approaches that can generate fossil datasets without spatiotemporal bias, which has ramifications for paleobiogeography, Phanerozoic diversity, and biostratigraphy.

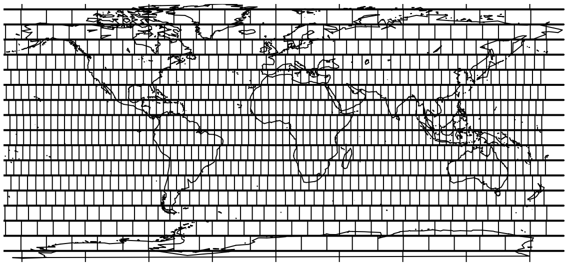

Figure 4.1: **500,000 meter equal area gridding scheme for geographic coverage measure superimposed on a geographic map of the present day.** This measure is equal to the number of equal area cells in a paleolatitude strip with fossil occurrences of our target taxa.

4.2 Methods

4.2.1 Data

The marine invertebrate groups with the best fossil records and which consequently are the most likely to have stable spatiotemporal sampling (anthozoans, brachiopods, echinoids, molluscs, and trilobites) were downloaded from the Paleobiology Database (PaleoDB). Subgenera were elevated to the genus level. From 481,613 occurrences across all but the earliest Cambrian (PaleoDB standardized 10 myr bins), we deduced the presence or absence of each genus for ten-degree paleolatitude strips from the paleolatitudes listed in the PaleoDB. To measure richness for each of the paleolatitude strips, we counted the number of these presences. Occurrences that were not constrained to one 10 myr bin were excluded. We chose paleolatitude because paleolongitude is less certain in deep time [145], and differential sampling of paleolatitudes is more likely to confound paleobiogeographical analyses [2]. Polar paleolatitude bins (90°S-80°S) and (80°N-90°N) were excluded from analyses to avoid edge effects.

Two measures of sampling bias were chosen, one measure for geographical breadth of sampling within a paleolatitude strip (to discriminate, for example, between paleolatitude strips where one marine shelf on one continent is sampled versus paleolatitude strips where many marine shelves on many continents are sampled), and one measure for collecting effort. We note that our intention is not to choose measures that correlate best with sampled richness, but rather to choose easily available proxies that capture the chief biases of sampled richness. These measures can be used in turn to produce estimates of richness that correct for these biases.

Measure of geographical breadth. Geographical breadth of sampling within each paleolatitude strip (extent of sampling) was established by summing the number of equal area cells with fossil occurrences. Figure 1 shows our demarcation of equal area cells (500,000 meter wide cells). We note that this quantity is not normalized by the maximum number of equal area cells with epicontinental seas that could conceivably preserve fossils. These data would improve our measure but are not easily available at a global scale. This is unlikely to be a problem, however, because studies of the distribution of molluscan fauna have continually shown that provincial area does not correlate with the total richness of that area [140, 58], suggesting that latitude (associated with many key ecological factors such as temperatue and primary productivity) rather than occupiable area in a province (or area within a given latitude band) is a bigger driver of species richness.

Measure of collection effort. For each paleolatitude strip, we also counted the total number of collections with an identifier in the PaleoDB. Only collections that contained our target taxa were considered. We used this quantity as a measure of collecting effort. While collection sizes vary and cover different scales, these factors are unlikely to affect our global study of ten degree paleolatitude bands.

In principle, as sampling improves the spatial distribution of taxonomic richness should diverge from the spatial distributions of sampling proxies. Consider, for example, the astounding biodiversity of many taxonomic groups in the modern neotropics despite much more sampling in northern high latitudes.

4.2.2 Quantifying spatial shifts

In the results we demonstrate the strong covariance between our sampling proxies and sampled richness. Yet this is only a problem if paleolatitude strips that are well sampled in one time interval become poorly sampled in the next time interval. It is therefore critical to quantify spatial shifts in the recorded distribution of taxa in the fossil record, such as a shift from high to low latitude sampling. We measured the median paleolatitude of fossil occurrences for our target taxa. A Mann-Whitney U test was used to gauge whether the paleolatitudinal distribution of fossil occurrences for adjacent time intervals were statistically different.

4.2.3 Null model of richness

For comparison with empirical patterns, we model what time series of sampled diversity would result if we assumed true generic richness within paleolatitudinal strips is invariant over time and driven purely by sampling pattern following the approach of Smith & McGowan [156] and Lloyd [98]. To create a single sampling proxy, we multiplied the number of five by five degree grid cells within fixed paleolatitude strips sampled by the number of collections in those strips. Because our analysis focused only on single paleolatitude strips, demarcations based on meters were unnecessary. Note this sampling proxy captures the multiplicative relationship between geographic coverage and sampling effort (a ratio such as collections per grid cell would fail to predict richness – twice the grid cells and twice the collections would equal the same ratio despite producing more richness). The time series for the sampling proxy and richness within a particular paleolatitude band were first logged to remove the effects of outliers. Next the sampling proxy and taxon richness counts (drawn from the PaleoDB) were independently ordered from smallest to largest. The equation of the best-fitting model (a linear regression) to these data can then be applied to the rock record in its original time series. The difference between the expectation of this null model and the actual empirical count of sampled generic richness provides an estimate of whether diversity in a paleolatitudinal band is greater, less, or equal to what we would expect given the sampling proxy.

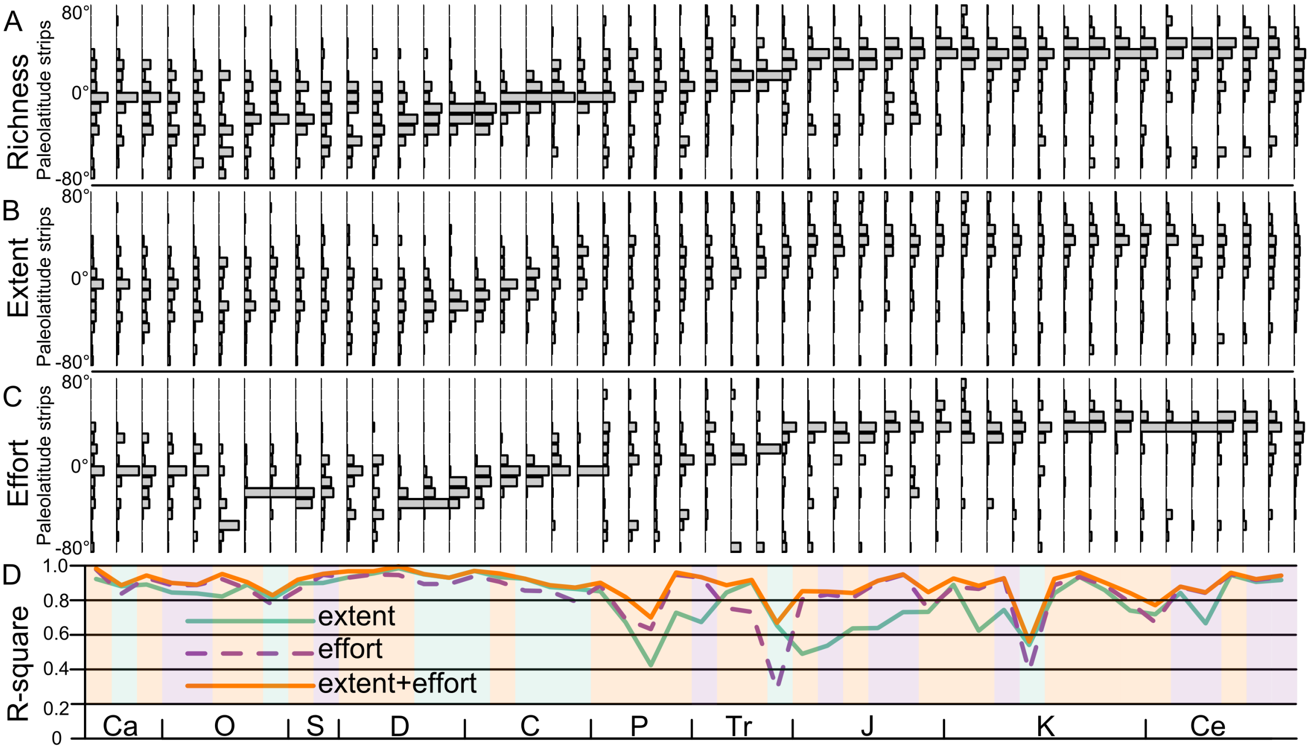

Figure 4.2: **Patterns of richness and sampling proxies through the Phanerozoic.** A) Distribution of genus richness across paleolatitude strips. B) Distribution of total equal area grid cells with at least one fossil locality recorded in the PaleoDB across paleolatitude strips. C) Distribution of faunal lists with collection IDs in the PaleoDB across paleolatitude strips. D) The percentage variation of richness in each paleolatitude strips explained by geographic coverage and sampling intensity in each of those paleolatitude strips. Note that the sampling proxies are not rendundant; parts of the Phanerozoic lack geographic coverage but have high sampling intensity and vice versa. Each interval is shaded by the color of the model with the lowest AIC score.

4.3 Determinants of the geographical distribution of richness

To visualize geographical richness patterns across the Phanerozoic, we plotted the distribution of generic richness next to the distribution of total sampled grid cells (extent) and collection effort (Figure 2A-C). To gauge how much of the spatial variation in richness (Figure 2A) is determined by the extent and intensity of geographical sampling (Figure 2B-C), we calculated the R^2 from a multiple regression analysis with richness as the response variable and the sampling proxies as the predictors. We logged each covariate to account for nonlinear relationships between richness, effort, and extent. We performed this analysis for each time interval. Figure 2D shows the resulting R^2 values for three models, sampling extent (number of sampled grid cells) and sampling effort alone (number of collections), and sampling extent and sampling effort combined.

All regressions were significant ($P < 0.05$). Extent alone has a mean R^2 of 0.80. Extent alone predicts richness the worst in Permian 3 and Jurassic 1, dropping to R^2 values of 0.42 and 0.49, respectively. Effort alone has a mean R^2 of 0.84 and predicts richness the worst in Triassic 4 and Cretaceous 4, dropping to R^2 values of 0.28 and 0.39. However, this apparent inability of the sampling proxies to explain richness patterns for a few intervals is lost when the sampling proxies are combined as predictors. Extent and effort together never drop below an R^2 of 0.55, and the average R^2 for this combined model is 0.89. AIC scores from these models reveal no systematic advantage of one model over another, with extent alone being the preferred model in 21% of the time intervals, effort alone being the preferred model in 25% of the time intervals, and the combined model being preferred for 54% of the time intervals.

Taken alone, extent and effort have systematically lower R^2 values through several intervals. However, the combined model is able to predict richness through these intervals, which suggests that intervals are differentially affected by extent and effort in the database. For example, intervals may be intensively sampled but lack geographic breadth, or contain geographic breadth but lack intensive sampling. This result underscores the need for more intensive sampling effort and geographic breadth in order to uncover global biological pattern.

4.4 Shifts in geographic coverage

Our analysis reveals (i) a marked heterogeneity in paleolatitudinal coverage over time and (ii) marked shifts in paleolatitudinal coverage between time intervals. Figure 3A shows the median paleolatitude of fossil occurrences over the Phanerozoic, with the error bars reflecting the 25th and 75th percentile. To compare the long-term signal in the data with the median paleolatitude in each time interval, a red moving average line (5 points) is shown in Figure 3A. We tested for a difference between successive time intervals with a Mann-Whitney U test. All tests were significant ($P < 0.05$) and Figure 3B shows the log value of the U statistic across the Phanerozoic. Repeated contractions and expansions of the error bars are evident, revealing repeated areas of the record where geographic coverage is expanded but then lost in the next time interval.

Among the most drastic shifts includes the Middle Ordovician (Ordovician 2 to 3 and 3 to 4), with a shift in sampling from high to low latitudes (47° drop) and subsequent reversion (30° increase). The record remains comparatively stable until Devonian 1, with a 15° median increase. The record is volatile until the late Devonian, where the record appears to stably increase in median latitude, likely reflecting geologic signal rather than sampling bias. The PermoCarb boundary is exceptionally violent, with a vast expansion of geographical coverage in Permian 1. Geographic coverage is volatile until Jurassic 1, where the record remains mostly consistent. The remainder of the Mesozoic and Cenozoic is characteristic of well-sampled high latitude European and North American formations [152], though several intervals have sudden expansions of coverage that is subsequently lost (see Cretaceous 4). The last large shift is in the late Cenozoic (Cenozoic 6), with the recovery of Southern high latitude regions (Figure 4).

Some intervals have severe alterations of biogeographical representation. At the Permo-Carboniferous boundary, high latitudes (above 30 degrees) collectively increase richness by over 12-fold, while low latitude richness increases by just 2-fold.

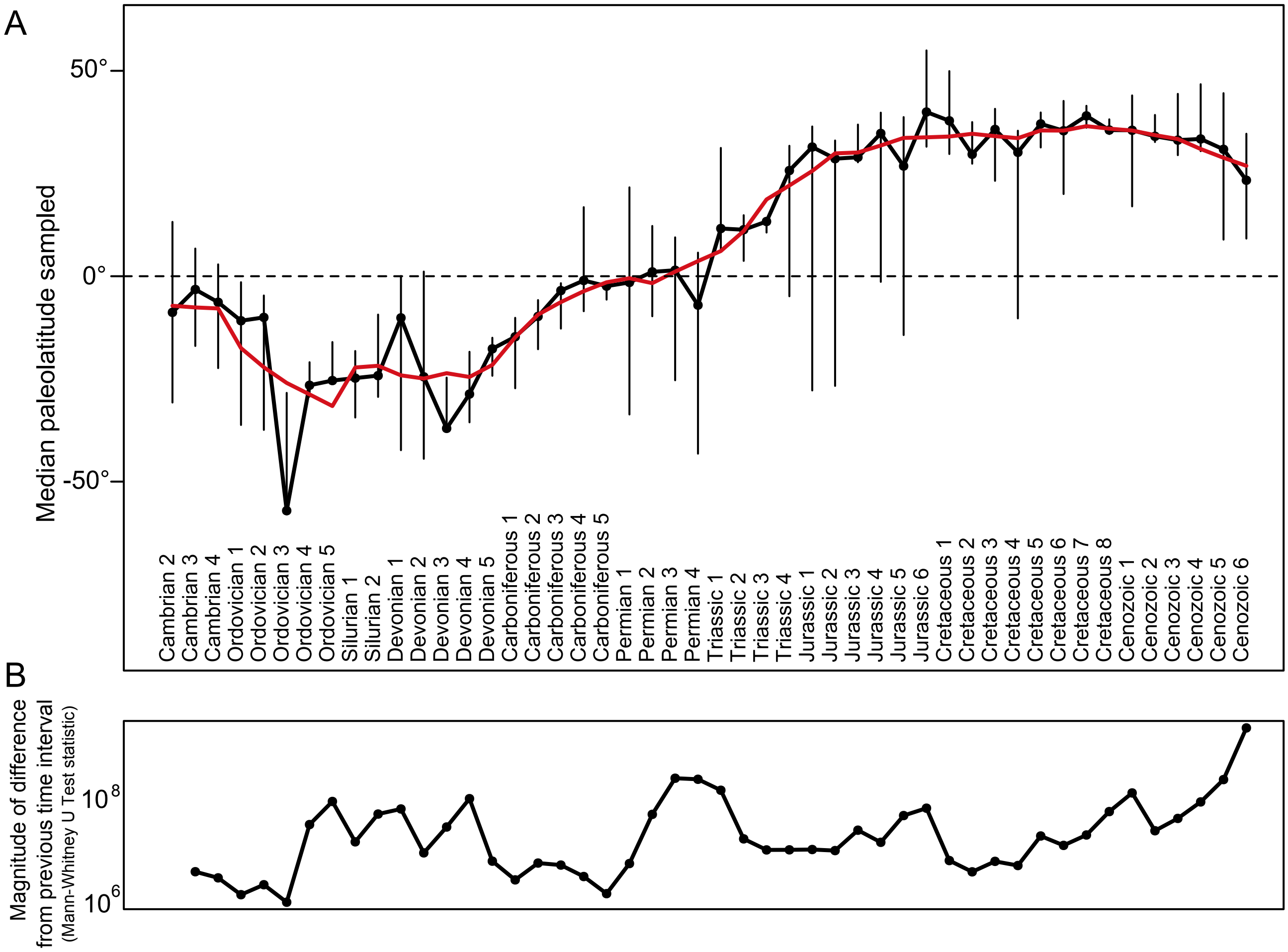

Figure 4.3: **Spatial shifts in fossil occurrences through the Phanerozoic.** A) The median latitude of fossil occurrences steadily rises through the Phanerozoic, but is punctuated by short-term noise and contractions and expansions of geographic coverage. Error bars indicate 25th and 75th quantiles, while the red line is a moving average across five points. B) The Mann-Whitney U test statistic plotted for each interval. A higher test-statistic corresponds to a more severe change in latitude. All transitions are statistically significant but vary in their effect size.

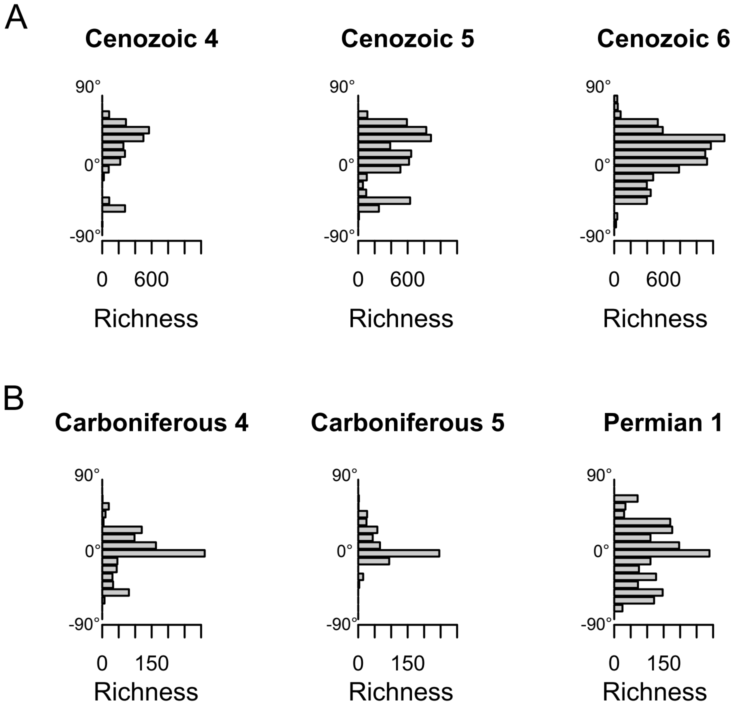

Figure 4.4: **The distribution of genus richness across latitudes plotted for key intervals.** A) For the last three Cenozoic time intervals (Cenozoic 4-6), Cenozoic 6 has more equitable sampling across latitudes than its predecessors. B) The Permo-Carboniferous boundary reflects a weakness in geographic coverage that biases estimates of global diversity inferred by subsampling.

4.5 Paleodiversity through time within fixed latitudinal strips

We examined spatial sampling and recorded genus richness through the Phanerozoic within two fixed paleolatitudinal bands, a paleotropical band from 10°S to 10°N and a paleotemperate band from 30°N to 50°N. Fig. 5A, B shows the relative spatial coverage attained in the two paleolatitudes. The paleotemperate latitude band shows a general rise in area sampled since the late Paleozoic, in marked contrast to the marked drop in spatial coverage in tropical latitudes from the Jurassic to the Neogene. The two time series are not correlated in the raw data: (Spearman rank correlation $\rho = -0.18$, $P = 0.21$) but after first differenced data show a marginally significant level ($\rho = 0.35$, $P = 0.02$). The differences between the two paleolatitudinal bands is largely explained by continental drift of well-studied areas from equatorial into temperate paleolatitudes over time.

To remove the effects of continental drift and spatial heterogeneity on Phanerozoic diversity estimates dominated by North American and European data we plotted the mean genus richness (number of genera divided by the number of equal area grids with records) over time in paleotropical and paleotemperate latitudes (Fig. 5C, D). While richness per unit area shows a progressive increase over time since the Carboniferous in temperate latitudes no such trend is apparent in tropical latitudes and the two time series show no correlation (Spearman rank correlation: raw data $\rho = 0.09$, $P = 0.55$; first differenced $\rho = 0.07$, $P = 0.67$). Furthermore the rise in temperate latitude richness is matched by a rise in the number of collections recorded in the database at these paleolatitudes (data not shown), as well studied continental blocks drift northwards over geological time. Changes in richness per unit area are strongly correlated with changes in sampling intensity per unit area. The lack of a common trend in richness per unit area over time between equatorial and temperate regions suggests sampling pattern is dominating the signal.

To discover what residual signal resides in these data that cannot be attributed to variation in sampling, we constructed models of Phanerozoic richness assuming true richness to be uniform over time and entirely driven by variation in sampling effort. The match between model and empirical data for both paleolatitude datasets is good but not perfect (Fig. 5E, F), with a significant proportion of the data left unexplained (Fig. 5G, H).

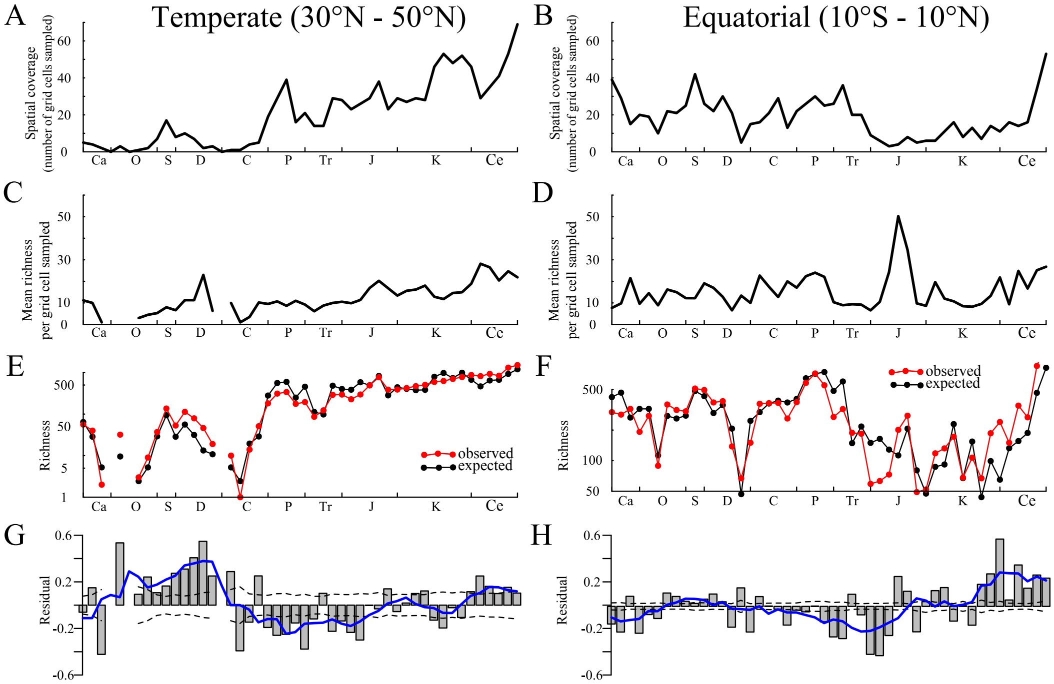

Figure 4.5: **Diversity and sampling bias in latitude strips**. A-B) Number of grid cells with sampled fossils for each time bin within two fixed paleolatiutudinal belts (temperate and equatorial). A gradual increase in sampled grid cells is evident in the temperate strip (A), while no such pattern is evident in the equatorial strip (B). C-D) Mean richness per sampled grid cell reveals no obvious pattern for faunas in the two paleolatitudinal belts. E-F) Null model that assumes biodiversity is driven purely by sampling (black) compared with observed genus richness (red). The null model explains the overall signal in the data, but select portions of the Phanerozoic deviate from the expectation. G-H) Plots show the difference between empirical richness and the expectation of the null model. Dashed bars indicate 99% confidence intervals for the null model. Overall, we find that the temperate and tropical faunas have similar trajectories despite markedly different trends in spatial sampling pattern over time.

Despite the very different spatial sampling records in these two bands (Fig. 5A, B), the residuals show similar long term trends: rising diversity from the Cambrian to mid Devonian, by a steady decline to a nadir in the Triassic (equatorial) to early Jurassic (temperate), followed by a rise to the Recent with short term downturns in the mid Cretaceous and Neogene (Fig 5G, H). The two time series of residuals are positively correlated for raw data (Spearman rank correlation $\rho = 0.36$, $P = 0.01$), but not for first differenced data ($\rho = -0.11$, $P = 0.49$).

4.6 Discussion

Heterogeneity of the geological record makes it impossible to sample the fossil record uniformly over time. If we are to understand patterns of evolution then we need to first understand the nature of the problem and then develop methods that can compensate for such problems. Good progress has been made now to ensure that the fossil record that is preserved can be sampled fairly [3, 5]. But, before even the first fossil is collected, there is already an inbuilt bias to that record that needs to be taken into account if we are to interpret patterns of diversity correctly. So far consideration has only been given to how the macroarchitecture of the geological record affects our ability to sample fairly, either through controlling the amount of rock surviving from each time interval [152] or changing the proportional representation of environments that can be sampled [154]. What we show here is that there is also an inbuilt unfair sampling in terms of geographical coverage over time. Current methods that employ subsampling or rarefaction to correct for sampling irregularities assume that the area from which the data are drawn is uniform over time, which we now know is not true.

Two biases are chief candidates for driving spatial patterns in sampled richness: the distribution of sampling effort across latitudes, and the distribution of fossil localities across latitudes. In principle, the pattern of sampled richness should converge on the true, biological pattern of richness when sampling is equitable and sufficient across latitudes. However, before that point, the distribution of richness should mirror the geography of localities and effort. Limiting our analyses to the best-sampled groups within the largest macrofossil database available, we found that these biases explain the majority of the variation in richness for all time intervals. This suggests that the observed latitudinal richness for major marine animal groups, at least on a global scale, is a pattern that is majorly determined by sampling for the entire Phanerozoic.

Yet richness determined by sampling is not in itself a major problem so long as the geographic coverage of the data remains stable, because differences in sampling effort and rock amount between time bins could in principle be corrected for. However, spatiotemporal shifts in richness, driven by sampling, have the potential to profoundly influence analyses of biogeographical change and global biodiversity. For example, a change in relative sampling of distinct oceanic regions and climate zones [16], each with different levels of biodiversity and unique biotic compositions, could falsely give the impression of altered global diversity. We found a series of severe sampling shifts, with some time bins bearing little resemblance to the time bins before them (Figure 3A). The Paleobiology Database remains dominated by entries from Northern Hemisphere countries, particularly from Europe and North America (for example, 80% of Silurian records come from these two regions). This is reflected in our data, with long-term shifts in the relative dominance of sampling within different latitudinal bands created by the paleogeographical movement of well-studied regions over time. It is noteworthy that the median latitude in the sampling distribution steadily rises from the Cambrian to the latest Cenozoic (Figure 3A), and this has already been identified as a source of bias for diversity inference [2, 169].

Changes in sampled diversity that occur within blocks of time during which spatial shifts in richness are minimal clearly cannot be a result of differences in latitudinal representation in the database. However, where a change in diversity also coincides with a change in latitudinal sampling it is critical to test whether the latter could be driving the former. For example, we note that the last two periods of the Cenozoic, where diversity is perhaps rising, coincide with a rise in equitable sampling across latitudinal strips, with an overall high latitude richness increase of 3% and a low latitude richness increase of 118% (Figure 4A; above versus below 30 degrees). Our approach shows that it is impossible to reject the idea that the Neogene rise in global marine diversity is due to more equable latitudinal recording in the database, specifically in the tropics.

Another example is the major change in latitudinal sampling that occurs between the

Late Carboniferous and Permian (Figure 4B). This coincides with anomalously low reported diversity for time interval Carboniferous 5 and high diversity for Permian 1 in sample standardized estimates of Phanerozoic diversity [5, 155], suggesting that uneven biogeographical sampling plays a large part in creating this. This is evident from our analysis of this boundary (Figure 3A), with a drastic increase in geographic coverage in Permian 1 (Figure 4B), and a 12-fold increase in high latitude richness.

Changes in the extent of epicontinental seas, driven by large-scale sea-level cycles or major tectonic events such as continental rifting or collision, can also affect the spatial distribution of fossiliferous marine deposits. So a common cause explanation [120], where the geographical distribution of marine fossiliferous rocks and marine diversity are both affected by the same driver, needs to be considered. Indeed, a recent study of genus richness and geographic area during the late Cretaceous [92] has clearly demonstrated a positive relationship between genus richness and geographic area in both epicontinental seas and ocean-facing coastlines. By looking at genus richness from fixed paleolatitudinal strips, our analysis draws data from multiple cratonic blocks, each with its own unique tectonic history, rock record [106], genus-area relationships [92], and idiosyncratic response to sea-level change [61]. Despite showing very different patterns of spatial sampling (Fig. 5A, B) and recorded genus richness through time (Fig. 5E, F, red line), the fossil records from equatorial and temperate paleolatitudinal strips reveal the same long-term trend in diversity after spatial sampling differences are accounted for (Fig. 5G, H). Diversity rises to the mid Devonian, falls gradually through the late Paleozoic and into the early Mesozoic before rising again to the Recent. While a similar long-term trend has been noted before [156, 5, 155, 56, 155], our analysis provides the strongest evidence yet for it being a truly global biodiversity signal by showing it is replicated at different paleolatitudes. This strongly suggests that the long-term diversity trend is indeed the result of the large-scale tectonic evolution of the planet, following the first order Wilson cycles of continental accretion and dispersal as suggested previously [156, 56, 155].

On the other hand, small scale (stage-to-stage) shifts in genus richness (after accounting for spatial sampling differences) show no correlation between the two paleolatitudinal strips, suggesting that they were not responding to a common global driver but rather result from a sequence of region-specific events. This highlights that while a common cause explanation fits large-scale (100-200 myr cycle) patterns in the fossil record, the shorter (ca. 50-60 myr) cycles [108] are much more likely to be a reflection of region-specific changes in the original marine area [120, 110, 92] and the extent of any subsequent degradation of that rock record at outcrop [154].

In this paper we have shown that there is a strong geographical bias to the distribution of paleontological records and this needs to be taken into account in any assessment of biodiversity trends over geological time. Even in the best available database of fossil occurrences the fact that some time intervals are better sampled than others introduces a strong confounding effect on how we perceive global biodiversity to have changed over Phanerozoic time. In any biodiversity survey, sampling effort and recorded diversity will track each other initially but then start to diverge as sufficient records accumulate, until additional sampling adds no new records and they are effectively independent. We have shown that sampling effort in the fossil record still closely tracks recorded diversity and thus poses a severe problem for paleobiodiversity analysis. Yet, with the right techniques and approaches it is possible to tease out a global signal that is independent of spatial sampling biases. Spatial bias needs to be considered seriously in all future analyses of paleobiodiversity.

We thank the Paleobiology Database for providing the Phanerozoic invertebrate data. This is Paleobiology Database publication number is 188. We thank Carl Bergstrom for helpful discussions.

BIBLIOGRAPHY

[1] R. Albert and A.L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[2] Briggs DEG Allison, PA. Paleolatitudinal sampling bias, phanerozoic species diversity,, the end-permian extinction. *Geology*, 21(1):65–68, 1993.

[3] Aberhan M Bottjer DJ Foote M Fürsich FT et al Alroy, J. Phanerozoic trends in the global diversity of marine invertebrates. *Science*, 321(5885):97–100, 2008.

[4] J. Alroy. Appearance event ordination: a new biochronologic method. *Paleobiology*, pages 191–207, 1994.

[5] J Alroy. The shifting balance of diversity among major marine animal groups. *Science*, 329(5996):1191–1194, 2010.

[6] Marshall CR Bambach RK Bezusko K Foote M et al Alroy, J. Effects of sampling standardization on estimates of phanerozoic marine diversification. *Proceedings of the National Academy of Sciences*, 98(11):6261–6266, 2001.

[7] L.W. Alvarez, W. Alvarez, F. Asaro, and H.V. Michel. Extraterrestrial cause for the cretaceous-tertiary extinction. *Science*, 208(4448):1095, 1980.

[8] L.A.N. Amaral, A. Scala, M. Barthélémy, and H.E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21):11149, 2000.

[9] M.B. Araújo, A. Rozenfeld, C. Rahbek, and P.A. Marquet. Using species co-occurrence networks to assess the impacts of climate change. *Ecography*, 34:897–908, 2011.

[10] J.D. Archibald, WA Clemens, K. Padian, T. Rowe, N. Macleod, P.M. Barrett, A. Gale, P. Holroyd, H.D. Sues, N.C. Arens, et al. Cretaceous extinctions: multiple causes. *Science*, 328(5981):973–973, 2010.

[11] Peters SE Ausich, WI. A revised macroevolutionary history for ordovician early silurian crinoids. *Paleobiology*, 31:538–551, 2005.

[12] R.K. Bambach, A.M. Bush, and D.H. Erwin. Autecology and the filling of ecospace: key metazoan radiations. *Palaeontology*, 50(1):1–22, 2007.

[13] A.L. Barabási. Emergence of scaling in complex networks. *Handbook of graphs and networks*, pages 69–84, 1999.

[14] J. Bascompte, P. Jordano, C.J. Melián, and J.M. Olesen. The nested assembly of plant–animal mutualistic networks. *Proceedings of the National Academy of Sciences*, 100(16):9383–9387, 2003.

[15] L Baudouin, S Piry, and JM Cornuet. Analytical bayesian approach for assigning individuals to populations. *Journal of Heredity*, 95(3):217–224, 2004.

[16] Christina L Belanger, David Jablonski, Kaustuv Roy, Sarah K Berke, Andrew Z Krug, and James W Valentine. Global environmental predictors of benthic marine biogeographic structure. *Proceedings of the National Academy of Sciences*, 109(35):14046–14051, 2012.

[17] Upchurch P Benson, RBJ. Diversity trends in the establishment of terrestrial vertebrate ecosystems: Interactions between spatial, temporal sampling biases. *Geology*, 41(1):43–46, 2013.

[18] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008:P10008, 2008.

[19] DD Bloom and NR Lovejoy. Molecular phylogenetics reveals a pattern of biome conservatism in new world anchovies (family engraulidae). *Journal of Evolutionary Biology*, 25(4):701–715, 2012.

[20] AJ Boucot. Does evolution take place in an ecological vacuum? ii. *Journal of Paleontology*, 57:1–30, 1983.

[21] C.E. Brett. Sequence stratigraphy, biostratigraphy, and taphonomy in shallow marine environments. *Palaios*, pages 597–616, 1995.

[22] James H Brown. *Macroecology.* University of Chicago Press, 1995.

[23] James H Brown and Brian A Maurer. Macroecology: the division of food and space among species on continents. *Science*, 243(4895):1145–1150, 1989.

[24] L.B. Buckley and W. Jetz. Linking global turnover of species and environments. *Proceedings of the National Academy of Sciences*, 105(46):17836–17841, 2008.

[25] Milena Camardelli and Marcelo F Napoli. Amphibian conservation in the caatinga biome and semiarid region of brazil. *Herpetologica*, 68(1):31–47, 2012.

[26] Cathryn A Campbell and James W Valentine. Comparability of modern and ancient marine faunal provinces. *Paleobiology*, pages 49–57, 1977.

[27] Wheeley JR Wright VP Cherns, L. Taphonomic bias in shelly faunas through time: early aragonitic dissolution, its implications for the fossil record. In *Taphonomy: Process, Bias Through Time*, pages 79–105. Springer, 2011.

[28] Wright VP Cherns, L. Missing molluscs as evidence of large-scale, early skeletal aragonite dissolution in a silurian sea. *Geology*, 28(9):791–794, 2000.

[29] V. Courtillot and F. Fluteau. Cretaceous extinctions: the volcanic hypothesis. *Science*, 328(5981):973, 2010.

[30] Foote M Beu AG Cooper RA Matcham I et al Crampton, JS. Second-order sequence stratigraphic controls on the quality of the fossil record at an active margin: New zealand eocene to recent shelf molluscs. *Palaios*, 21(1):86–105, 2006.

[31] Michael D Crisp, Mary TK Arroyo, Lyn G Cook, Maria A Gandolfo, Gregory J Jordan, Matt S McGlone, Peter H Weston, Mark Westoby, Peter Wilf, and H Peter Linder. Phylogenetic biome conservatism on a global scale. *Nature*, 458(7239):754–756, 2009.

[32] Mike Crisp. Biome assembly: what we know and what we need to know. *Journal of Biogeography*, 33(8):1332–1333, 2006.

[33] Paleobiology Database. http://www.paleodb.org.

[34] Kevin De Queiroz. Species concepts and species delimitation. *Systematic Biology*, 56(6):879–886, 2007.

[35] S. D'Hondt. Consequences of the cretaceous/paleogene mass extinction for marine ecosystems. *Annual review of ecology, evolution, and Systematics*, pages 295–317, 2005.

[36] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[37] Michael J Donoghue. A phylogenetic perspective on the distribution of plant diversity. *Proceedings of the National Academy of Sciences*, 105(Supplement 1):11549–11555, 2008.

[38] S.N. Dorogovtsev. *Lectures on complex networks.* Oxford University Press, 2010.

[39] M.L. Droser, D.J. Bottjer, P.M. Sheehan, and G.R. McGhee Jr. Decoupling of taxonomic and ecologic severity of phanerozoic marine mass extinctions. *Geology*, 28(8):675–678, 2000.

[40] J.A. Dunne, R.J. Williams, and N.D. Martinez. Food-web structure and network theory: the role of connectance and size. *Proceedings of the National Academy of Sciences*, 99(20):12917, 2002.

[41] J.A. Dunne, R.J. Williams, N.D. Martinez, R.A. Wood, and D.H. Erwin. Compilation and network analyses of cambrian food webs. *PLoS biology*, 6(4):e102, 2008.

[42] S. Ekman. Zoogeography of the sea: 1-417. *Sidgwick & Jackson, London*, 1953.

[43] G. Ergün. Human sexual contact network as a bipartite graph. *Physica A: Statistical Mechanics and its Applications*, 308(1):483–488, 2002.

[44] D.H. Erwin. Lessons from the past: Biotic recoveries from mass extinctions. *Proceedings of the National Academy of Sciences*, 98(10):5399–5403, 2001.

[45] D.H. Erwin. Extinction as the loss of evolutionary history. *Proceedings of the National Academy of Sciences*, 105(Supplement 1):11520, 2008.

[46] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[47] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821, 2002.

[48] S.J. Gould. The paradox of the first tier: an agenda for paleobiology. *Paleobiology*, pages 2–12, 1985.

[49] J. Guex and E. Davaud. Unitary associations method: use of graph theory and computer algorithm. *Computers & Geosciences*, 10(1):69–96, 1984.

[50] R. Guimera and L.A.N. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895, 2005.

[51] R. Guimerà, M. Sales-Pardo, and L.A.N. Amaral. Classes of complex networks defined by role-to-role connectivity profiles. *Nature physics*, 3(1):63–69, 2006.

[52] R. Guimerà, M. Sales-Pardo, and L.A.N. Amaral. Classes of complex networks defined by role-to-role connectivity profiles. *Nature physics*, 3(1):63–69, 2006.

[53] E.M. Hagmeier. A numerical analysis of the distributional patterns of north american mammals. ii. re-evaluation of the provinces. *Systematic Biology*, 15(4):279–299, 1966.

[54] E.M. Hagmeier and C.D. Stults. A numerical analysis of the distributional patterns of north american mammals. *Systematic Zoology*, pages 125–155, 1964.

[55] A. Hallam. *Atlas of palaeobiogeography*. Elsevier Science & Technology, 1973.

[56] Peters SE Hannisdal, B. Phanerozoic earth system evolution, marine biodiversity. *Science*, 334(6059):1121–1124, 2011.

[57] Michael M Hansen, Ellen Kenchington, and Einar E Nielsen. Assigning individual fish to populations using microsatellite dna markers. *Fish and Fisheries*, 2(2):93–112, 2001.

[58] Jablonski D Krug AZ Valentine JW Harnik, PG. Genus age, provincial area, the taxonomic structure of marine faunas. *Proceedings of the Royal Society B: Biological Sciences*, 277(1699):3427–3435, 2010.

[59] Hannes Heikinheimo, Mikael Fortelius, Jussi Eronen, and Heikki Mannila. Biogeography of european land mammals shows environmentally distinct and spatially coherent clusters. *Journal of Biogeography*, 34(6):1053–1064, 2007.

[60] N.A. Heim and S.E. Peters. Regional environmental breadth predicts geographic range and longevity in fossil marine genera. *PloS One*, 6(5):e18946, 2011.

[61] SM Holland. Sea level change, the area of shallow-marine habitat: implications for marine biodiversity. *Paleobiology*, 38(2):205–217, 2012.

[62] N. Holstein and S.S. Renner. A dated phylogeny and collection records reveal repeated biome shifts in the african genus coccinia (cucurbitaceae). *BMC evolutionary biology*, 11(1):28, 2011.

[63] B.G. Holt, J.P. Lessard, M.K. Borregaard, S.A. Fritz, M.B. Araújo, D. Dimitrov, P.H. Fabre, C.H. Graham, G.R. Graves, K.A. Jønsson, et al. An update of wallaces zoogeographic regions of the world. *Science*, 339(6115):74–78, 2013.

[64] Carina Hoorn, FP Wesselingh, H Ter Steege, MA Bermudez, A Mora, J Sevink, I Sanmartín, A Sanchez-Meseguer, CL Anderson, JP Figueiredo, et al. Amazonia through time: Andean uplift, climate change, landscape evolution, and biodiversity. *Science*, 330(6006):927–931, 2010.

[65] Colin E Hughes, R Toby Pennington, and Alexandre Antonelli. Neotropical plant evolution: assembling the big picture. *Botanical Journal of the Linnean Society*, 171(1):1–18, 2013.

[66] IUCN. The iucn red list of threatened species, 2012.

[67] A.R. Ives and S.R. Carpenter. Stability and diversity of ecosystems. *Science*, 317(5834):58–62, 2007.

[68] D. Jablonski. Background and mass extinctions: the alternation of macroevolutionary regimes. *Science*, 231(4734):129, 1986.

[69] D. Jablonski. *Extinction in the fossil record. In: R.M. May and J.H. Lawton, eds., Extinction rates.* Oxford, Oxford University Press, 1995.

[70] D. Jablonski. Geographic variation in the molluscan recovery from the end-cretaceous extinction. *Science*, 279(5355):1327–1330, 1998.

[71] D. Jablonski. Survival without recovery after mass extinctions. *Proceedings of the National Academy of Sciences*, 99(12):8139, 2002.

[72] D. Jablonski. Mass extinctions and macroevolution. *Paleobiology*, 31(2_Suppl):192–210, 2005.

[73] D. Jablonski. Extinction and the spatial dynamics of biodiversity. *Proceedings of the National Academy of Sciences*, 105(Supplement 1):11528, 2008.

[74] D. Jablonski et al. Extinction: past and present. *Nature*, 427(6975):589–589, 2004.

[75] D. Jablonski and D.M. Raup. Selectivity of end-cretaceous marine bivalve extinctions. *Science*, 268(5209):389–391, 1995.

[76] D. Jablonski, K. Roy, J.W. Valentine, R.M. Price, and P.S. Anderson. The impact of the pull of the recent on the history of marine diversity. *Science*, 300(5622):1133–1135, 2003.

[77] David Jablonski and J John Sepkoski Jr. Paleobiology, community ecology, and scales of ecological pattern. *Ecology*, pages 1367–1378, 1996.

[78] Paul Jaccard. *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz, 1901.

[79] Johnson KB Jackson, JB. Measuring past biodiversity. *Science*, 293:2401–2404, 2001.

[80] N.P. James, Y. Bone, and T.K. Kyser. Where has all the aragonite gone? *Journal of Sedimentary Research*, 75(3):454–463, 2005.

[81] KG Johnson. New data for old questions. *Paleobiology*, 29(1):19–21, 2003.

[82] Sally A Keith, Tom J Webb, Katrin Böhning-Gaese, Sean R Connolly, Nicholas K Dulvy, Felix Eigenbrod, Kate E Jones, Trevor Price, David W Redding, Ian PF Owens, et al. What is macroecology? *Biology letters*, 8(6):904–906, 2012.

[83] G. Keller, T. Adatte, A. Pardo, S. Bajpai, A. Khosla, and B. Samant. Cretaceous extinctions: evidence overlooked. *Science*, 328(5981):974, 2010.

[84] Holland SM Kidwell, SM. The quality of the fossil record: implications for evolutionary analyses. *Annual Review of Ecology, Systematics*, 33:561–588, 2002.

[85] W. Kiessling. Long-term relationships between ecological stability and biodiversity in phanerozoic reefs. *Nature*, 433(7024):410–413, 2005.

[86] W. Kiessling and R.C. Baron-Szabo. Extinction and recovery patterns of scleractinian corals at the cretaceous-tertiary boundary. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 214(3):195–223, 2004.

[87] Patricia Koleff, Kevin J Gaston, and Jack J Lennon. Measuring beta diversity for presence–absence data. *Journal of Animal Ecology*, 72(3):367–382, 2003.

[88] Kenneth H Kozak and JohnJ Wiens. Does niche conservatism promote speciation? a case study in north american salamanders. *Evolution*, 60(12):2604–2621, 2006.

[89] H. Kreft and W. Jetz. A framework for delineating biogeographical regions based on species distributions. *Journal of Biogeography*, 37(11):2029–2053, 2010.

[90] Holger Kreft and Walter Jetz. Comment on “an update of wallaces zoogeographic regions of the world”. *Science*, 341(6144):10.1126/science.1237471, 2013.

[91] A.Z. Krug, D. Jablonski, and J.W. Valentine. Signature of the end-cretaceous mass extinction in the modern biota. *Science*, 323(5915):767–771, 2009.

[92] Miller AI Lagomarcino, AJ. The relationship between genus richness, geographic area in late cretaceous marine biotas: Epicontinental sea versus open-ocean-facing settings. *PLoS One*, 7(8):e40472, 2012.

[93] A. Lancichinetti and S. Fortunato. Community detection algorithms: A comparative analysis. *Physical Review E*, 80(5):056117, 2009.

[94] D.R. LAWRENCE. Taphonomy and information losses in fossil communities. *Bulletin of the Geological Society of America*, 79(10):1315, 1968.

[95] Adam D Leaché, Michelle S Koo, Carol L Spencer, Theodore J Papenfuss, Robert N Fisher, and Jimmy A McGuire. Quantifying ecological, morphological, and genetic variation to delimit species in the coast horned lizard species complex (phrynosoma). *Proceedings of the National Academy of Sciences*, 106(30):12418–12423, 2009.

[96] A. Lee McAlester and D.C. Rhoads. Bivalves as bathymetric indicators. *Marine Geology*, 5(5-6):383–388, 1967.

[97] L. Li and G. Keller. Abrupt deep-sea warming at the end of the cretaceous. *Geology*, 26(11):995, 1998.

[98] GT Lloyd. A refined modelling approach to assess the influence of sampling on palaeobiodiversity curves: new support for declining cretaceous dinosaur richness. *Biology Letters*, 8(1):123–126, 2012.

[99] R. Lockwood. Abundance not linked to survival across the end-cretaceous mass extinction: patterns in north american bivalves. *Proceedings of the National Academy of Sciences of the United States of America*, 100(5):2478, 2003.

[100] R. Lockwood. The k/t event and infaunality: morphological and ecological patterns of extinction and recovery in veneroid bivalves. *Paleobiology*, 30(4):507, 2004.

[101] J.S. Madin, J. Alroy, M. Aberhan, F.T. F
"ursich, W. Kiessling, M.A. Kosnik, and P.J. Wagner. Statistical independence of escalatory ecological trends in phanerozoic marine invertebrates. *Science*, 312(5775):897, 2006.

[102] C.R. Marshall and P.D. Ward. Sudden and gradual molluscan extinctions in the latest cretaceous of western european tethys. *Science*, 274(5291):1360–1363, 1996.

[103] K.S. McCann. The diversity–stability debate. *Nature*, 405(6783):228–233, 2000.

[104] Earl D McCoy and Kenneth L Heck Jr. Some observations on the use of taxonomic similarity in large-scale biogeography. *Journal of Biogeography*, pages 79–87, 1987.

[105] G.R. McGhee, P.M. Sheehan, D.J. Bottjer, and M.L. Droser. Ecological ranking of phanerozoic biodiversity crises: ecological and taxonomic severities are decoupled. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 211(3):289–297, 2004.

[106] Smith AB McGowan, AJ. Are global phanerozoic marine diversity curves truly global? a study of the relationship between regional rock records, global phanerozoic marine diversity. *Paleobiology*, 34:80–103, 2008.

[107] Smith AB McGowan, AJ. Comparing the geological, fossil records: implications for biodiversity studies. Geological Society of London, 2011.

[108] Bambach RK Melott, AL. A ubiquitous 62-myr periodic fluctuation superimposed on general trends in fossil biodiversity. ii. evolutionary dynamics associated with periodic fluctuation in marine diversity. *Paleobiology*, 37(3):383–408, 2011.

[109] L.A. Meyers, M. Newman, and B. Pourbohloul. Predicting epidemics on directed contact networks. *Journal of Theoretical Biology*, 240(3):400–418, 2006.

[110] Peters SE Meyers, SR. A 56 million year rhythm in north american sedimentation during the phanerozoic. *Earth, Planetary Science Letters*, 303:174–180, 2011.

[111] Foote M Miller, AI. Calibrating the ordovician radiation of marine life: implications for phanerozoic diversity trends. *Paleobiology*, 31:533–661, 1996.

[112] Norman Myers, Russell A Mittermeier, Cristina G Mittermeier, Gustavo AB Da Fonseca, Jennifer Kent, et al. Biodiversity hotspots for conservation priorities. *Nature*, 403(6772):853–858, 2000.

[113] S. Nee and R.M. May. Extinction and the loss of evolutionary history. *Science*, 278(5338):692, 1997.

[114] ND Newell. Periodicity in invertebrate evolution. *Journal of Paleontology*, pages 371–385, 1952.

[115] M.E.J. Newman. The structure and function of complex networks. *SIAM review*, pages 167–256, 2003.

[116] M.E.J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577, 2006.

[117] J.L. Payne and S. Finnegan. The effect of geographic range on extinction risk during background and mass extinction. *Proceedings of the National Academy of Sciences*, 104(25):10506–10511, 2007.

[118] R Toby Pennington, Quentin CB Cronk, and James A Richardson. Introduction and synthesis: plant phylogeny and the origin of major biomes. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1450):1455–1464, 2004.

[119] R Toby Pennington, James E Richardson, and Matt Lavin. Insights into the historical construction of species-rich biomes from dated plant phylogenies, neutral ecological theory and phylogenetic community structure. *New Phytologist*, 172(4):605–616, 2006.

[120] SE Peters. Geologic constraints on the macroevolutionary history of marine animals. *Proceedings of the National Academy of Sciences of the United States of America*, 102(35):12326–12331, 2005.

[121] E. Pierazzo and N. Artemieva. Local and global environmental effects of impacts on earth. *Elements*, 8(1):55–60, 2012.

[122] DM Raup. Species diversity in the phanerozoic: an interpretation. *Paleobiology*, pages 289–297, 1976.

[123] D.M. Raup. Extintion: bad genes or bad luck? *Acta geológica hispánica*, 16(1):25–33, 1981.

[124] DM Raup. Evolutionary radiations and extinctions. *Patterns of change in Earth evolution. Springer, Berlin*, pages 5–14, 1984.

[125] D.M. Raup. Biological extinction in earth history. *Science*, 231(4745):1528, 1986.

[126] D.M. Raup. *Extinction: bad genes or bad luck?* WW Norton & Company, 1992.

[127] D.M. Raup and R.E. Crick. Measurement of faunal similarity in paleontology. *Journal of Paleontology*, pages 1213–1227, 1979.

[128] D.M. Raup and D. Jablonski. Geography of end-cretaceous marine bivalve extinctions. *Science*, 260(5110):971–973, 1993.

[129] D.M. Raup and J. Sepkoski. Mass extinctions in the marine fossil record. *Science*, 215(4539):1501, 1982.

[130] Paul R. Renne, Roland Mundil, Greg Balco, Kyoungwon Min, and Kenneth R. Ludwig. Joint determination of 40k decay constants and 40ar*/40k for the fish canyon sanidine standard, and improved accuracy for 40ar/39ar geochronology. *Geochimica et Cosmochimica Acta*, 74(18):5349 – 5367, 2010.

[131] M.A. Rex, J.A. Crame, C.T. Stuart, and A. Clarke. Large-scale biogeographic patterns in marine mollusks: a confluence of history and productivity? *Ecology*, 86(9):2288–2297, 2005.

[132] D.S. Robertson, M.C. McKenna, O.B. Toon, S. Hope, and J.A. Lillegraven. Survival in the first hours of the cenozoic. *Geological Society of America Bulletin*, 116(5-6):760–768, 2004.

[133] P. Roopnarine. Graphs, networks, extinction and paleocommunity food webs. 2010.

[134] P.D. ROOPNARINE. Ecological modeling of paleocommunity food webs. *PALEONTOLOGICAL SOCIETY PAPERS*, 15:196.

[135] P.D. Roopnarine. Extinction cascades and catastrophe in ancient food webs. *Paleobiology*, 32(1):1–19, 2006.

[136] P.D. Roopnarine, K.D. Angielczyk, S.C. Wang, and R. Hertog. Trophic network models explain instability of early triassic terrestrial communities. *Proceedings of the Royal Society B: Biological Sciences*, 274(1622):2077–2086, 2007.

[137] M. Rosvall, D. Axelsson, and C.T. Bergstrom. The map equation. *European Journal of Physics*, 178(1):13–23, 2009.

[138] M. Rosvall and C.T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[139] M. Rosvall and C.T. Bergstrom. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one*, 6(4):e18209, 2011.

[140] Jablonski D Valentine JW Rosenberg G Roy, K. Marine latitudinal diversity gradients: tests of causal hypotheses. *Proceedings of the National Academy of Sciences*, 95(7):3699–3702, 1998.

[141] K. Roy, D. Jablonsk, and J.W. Valentine. Thermally anomalous assemblages revisited: patterns in the extraprovincial latitudinal range shifts of pleistocene marine mollusks. *Geology*, 23(12):1071, 1995.

[142] P.M. Sadler. Quantitative biostratigraphy-achieving finer resolution in global correlation. *Annu. Rev. Earth Planet. Sci.*, 32:187–213, 2004.

[143] Luis F Salazar, Carlos A Nobre, and Marcos D Oyama. Climate change consequences on the biome distribution in tropical south america. *Geophysical Research Letters*, 34(9):L09708, 2007.

[144] Stan Salvador and Philip Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, pages 576–584. IEEE, 2004.

[145] WF Schmachtenberg. Paleolongitudinal estimates for paleocontinents derived from interplate distances based on late ordovician bivalves. *Paleobiology*, 37(3):438–444, 2011.

[146] P. Schulte, L. Alegret, I. Arenillas, J.A. Arz, P.J. Barton, P.R. Bown, T.J. Bralower, G.L. Christeson, P. Claeys, C.S. Cockell, et al. The chicxulub asteroid impact and mass extinction at the cretaceous-paleogene boundary. *Science*, 327(5970):1214–1218, 2010.

[147] RW Scott, PA Fernández-Mendiola, E. Gili, and A. Simo. Persistence of coral-rudist reefs into the late cretaceous. *Palaios*, pages 98–110, 1990.

[148] J.J. Sepkoski. A compendium of fossil marine animal genera. *Bulletins of American Paleontology*, 363:1–560, 2002.

[149] C.E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[150] Marcelo F Simon, Rosaura Grether, Luciano P de Queiroz, Cynthia Skema, R Toby Pennington, and Colin E Hughes. Recent assembly of the cerrado, a neotropical plant diversity hotspot, by in situ evolution of adaptations to fire. *Proceedings of the National Academy of Sciences*, 106(48):20359–20364, 2009.

[151] C. Simpson and W. Kiessling. The role of extinction in large-scale diversity–stability relationships. *Proceedings of the Royal Society B: Biological Sciences*, 277(1686):1451, 2010.

[152] AB Smith. Large–scale heterogeneity of the fossil record: implications for phanerozoic biodiversity studies. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1407):351–367, 2001.

[153] A.B. Smith and C.H. Jeffery. Selectivity of extinction among sea urchins at the end of the cretaceous period. *Nature*, 392(6671):69–71, 1998.

[154] Benson RBJ Smith, AB. Marine diversity in the geological record, its relationship to surviving bedrock area, lithofacies diversity,, original marine shelf area. *Geology*, 41(2):171–174, 2013.

[155] Lloyd GT McGowan AJ Smith, AB. Phanerozoic marine diversity: rock record modelling provides an independent test of large-scale trends. *Proceedings of the Royal Society B: Biological Sciences*, 279(1746):4489–4495, 2012.

[156] McGowan AJ Smith, AB. The shape of the phanerozoic marine palaeodiversity curve: how much can be predicted from the sedimentary rock record of western europe? *Palaeontology*, 50(4):765–774, 2007.

[157] R.R. Sokal, P.H.A. Sneath, et al. Principles of numerical taxonomy. *Principles of numerical taxonomy.*, 1963.

[158] Thorvald Sørenson. *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons.* 1948.

[159] T. Steuber, S.F. Mitchell, D. Buhl, G. Gunter, and H.U. Kasper. Catastrophic extinction of caribbean rudist bivalves at the cretaceous-tertiary boundary. *Geology*, 30:999–1002, 2002.

[160] J Terborgh. On the notion of favorableness in plant ecology. *American Naturalist*, pages 481–501, 1973.

[161] R Toby Pennington, Darién E Prado, and Colin A Pendry. Neotropical seasonally dry forests and quaternary vegetation changes. *Journal of Biogeography*, 27(2):261–273, 2000.

[162] P. Uetz, L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, et al. A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 2000.

[163] Jablonski D Kidwell S Roy K Valentine, JW. Assessing the fidelity of the fossil record by using marine bivalves. *Proceedings of the National Academy of Sciences*, 103(17):6599–6604, 2006.

[164] J.W. Valentine, T.C. Foin, and D. Peart. A provincial model of phanerozoic marine diversity. *Paleobiology*, pages 55–66, 1978.

[165] J.W. Valentine and D. Jablonski. Origins of marine patterns of biodiversity: some correlates and applications. *Palaeontology*, 53(6):1203–1210, 2010.

[166] J.C. Vamosi and S.M. Vamosi. Extinction risk escalates in the tropics. *PloS One*, 3(12):e3886, 2008.

[167] Mark Vellend. Do commonly used indices of β-diversity measure species turnover? *Journal of Vegetation Science*, 12(4):545–552, 2001.

[168] P.J. Wagner, M.A. Kosnik, and S. Lidgard. Abundance distributions imply elevated complexity of post-paleozoic marine ecosystems. *Science*, 314(5803):1289, 2006.

[169] Wilkinson BH Ivany LC Walker, LJ. Continental drift, phanerozoic carbonate accumulation in shallow-shelf, deep-marine settings. *Journal of Geology*, 110(1):75–87, 2002.

[170] A.R. Wallace. *The geographical distribution of animals.* 1876.

[171] Robert Harding Whittaker. Vegetation of the siskiyou mountains, oregon and california. *Ecological Monographs*, 30(3):279–338, 1960.

[172] John A Wiens. Spatial scaling in ecology. *Functional ecology*, 3(4):385–397, 1989.

[173] John J Wiens. The niche, biogeography and species interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1576):2336–2350, 2011.

[174] John J Wiens, David D Ackerly, Andrew P Allen, Brian L Anacker, Lauren B Buckley, Howard V Cornell, Ellen I Damschen, T Jonathan Davies, John-Arvid Grytnes, Susan P Harrison, et al. Niche conservatism as an emerging principle in ecology and conservation biology. *Ecology letters*, 13(10):1310–1324, 2010.

[175] G.P. Wilson, A.R. Evans, I.J. Corfe, P.D. Smits, M. Fortelius, and J. Jernvall. Adaptive radiation of multituberculate mammals before the extinction of dinosaurs. *Nature*, 483:457–460, 2012.

[176] P. Wright, L. Cherns, and P. Hodges. Missing molluscs: field testing taphonomic loss in the mesozoic through early large-scale aragonite dissolution. *Geology*, 31(3):211–214, 2003.