Title

C-PHAST: Community Phylogenetic Analysis at Speedy Time Characterizing the Evolutionary History of California Across Taxonomic Groups

Author

Chari, Maya Rain

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

C-PHAST: Community Phylogenetic Analysis at Speedy Time

Characterizing the Evolutionary History of California Across Taxonomic Groups

A thesis submitted in partial satisfaction of the requirements

for the degree Master of Science in Bioinformatics

by

Maya Rain Chari

2024

ABSTRACT OF THE THESIS

C-PHAST: Community Phylogenetic Analysis at Speedy Time

Characterizing the Evolutionary History of California Across Taxonomic Groups

by

Maya Rain Chari

Masters of Science in Bioinformatics

University of California, Los Angeles, 2024

Professor Michael Edward Alfaro, Chair

The emerging field of spatial phylogenetics allows scientists to understand biodiversity and community ecology with the added lens of deep time. Biodiversity metrics that account for the shared evolutionary history in an ecosystem can give insight into the environmental filters that drive patterns of community assembly, and can help guide conservation efforts toward preservation of functional diversity. Here I present a new method: Community Phylogenetic Analysis at Speedy Time (C-PHAST), which builds on existing frameworks to produce a comprehensive pipeline for analysis of large-scale phylogenetic dispersion and community assembly. I apply this new method to characterize the phylodiversity of birds, plants

squamates, mammals and butterflies that are endemic to California. I create high resolution visualizations of the distribution of phylogenetic diversity across these clades within California, and find that patterns of over- and under- dispersion have low relative cross-clade correlation. Cumulative evolutionary history across clades reveals patches of overdispersion in northeastern California, in the southern Palm Springs desert regions, and in select coastal areas. I use the C-PHAST findings to pinpoint unmanaged regions harboring exceptionally high evolutionary history across taxonomic groups as recommendations for future conservation efforts.

The thesis of Maya Rain Chari is approved

James O. Lloyd-Smith

Felipe Zapata

Michael Edward Alfaro, Committee Chair

University of California, Los Angeles

2024

To my dear family and my loving friends

TABLE OF CONTENTS

INTRODUCTION 1

Introduction to Phylogenetic Diversity 1

Ecological Importance of Phylogenetic Diversity 2

Common Metrics of Community Phylogenetic Structure 4

Spatial Phylogenetics 5

Previous Spatial Phylogenetics Methods 7

C-PHAST: A New Tool for Spatial Phylogenetics 9

METHODS 11

Overview of Community Phylogenetic Analysis at Speedy Time (C-PHAST) 11

Partitioning California 14

Preparing California Phylogenies and Spatial Data 15

Fitting Contours of Evolutionary History 17

C-PHAST on California 17

Comparing C-PHAST to Biodiverse 20

Cross-Clade Correlations 21

Analysis of Protected Areas 21

RESULTS 22

California Phylogenies and Spatial Data 22

Contours of Evolutionary History 23

C-PHAST California 26

Comparing C-PHAST to Biodiverse 29

Cross-Clade Correlations 31

Phylodiversity of California's Protected Areas 33

DISCUSSION 37

Overview of the Project Goals 37

Plants 38

Birds 39

Mammals 40

Butterflies 40

Squamates 41

Patterns of Cross-Clade Correlation 42

Toward Biodiversity Conservation 43

Study Limitations 46

Conclusion 47

APPENDIX 48

Supplementary I 48

Supplementary II 50

Supplementary III 52

Supplementary IV 59

Supplementary V 60

Supplementary Datasets 61

Supplementary Code 61

LIST OF FIGURES

1. C-PHAST schematic illustrated with a four-community example of plants across California 13
2. Baro5 model fits of upper and lower 95% confidence intervals of PD, MPD and MNTD varying with richness across clades 25
3. Maps of cumulative significance of PD, MPD and MNTD across clades when compared to both California and Ecoregions 28
4. Model output of PD significance across California compared to a previous study with regions of similarity between the two outputs highlighted 30
5. Pairwise Spearman's Rank Correlation Coefficient matrices of PD, MPD and MNTD between clades for California and ecoregion regional models 32
6. California Reserve coverage of phylogenetically over, under, and neutrally dispersed regions across management agencies 35
7. Regions recommended as future protected areas across California 36

S1. California geographic partitions into 11,000 hexagons and 13 unique ecoregions 48
S2. A California plant species range before and after filtering 49
S3. Histograms of hexagonal incomplete sampling at the species level 50
S4. Histograms of hexagonal incomplete sampling at the genus level for select clades 51
S5. Baroflex five parameter log logistic model fits of phylodiversity against richness 52
S5. Akaike weights for the seven tested functions for PD, MPD and MNTD 53
S6. Pairwise Spearman's Rank Test large matrices 59

LIST OF TABLES

1.	Data sources of range and phylogeny for the five clades of interest	23
S1.	Raw Akaike weight values for all 7 tested models across all five clades	54
S2.	Mean Absolute Percent Error across all clades and metrics	56
S3.	Reserve area phylodiversity coverage	60

ACKNOWLEDGEMENTS

This work would not be possible without the continued help and support of my Principal Investigator, Michael Alfaro, and the rest of the Alfaro lab. I particularly want to thank Jonathan Chang for guidance across all aspects of this project; from scripting, to writing, to parsing the meaning of results. I thank Pascal Title for my introduction to the field of spatial phylogenetics. I would also like to thank all contributing studies from which I was able to gather valuable range and phylogenetic data. I thank Mishler et al. 2017 as Figure 4 is reproduced from Mishler et al 2017, used with permission under the Creative Commons Attribution License 4.0. I thank Vaughn Shirey for providing the North American butterfly range data.

A special appreciation for my committee, Dr. Michael Alfaro, Dr. Felipe Zapata and Dr. Jamie Lloyd-Smith for their feedback and support. Thank you to Dr. Zapata for the direction with plant community phylogenetics and for continued excitement about this project.

A final thanks to all my professors and mentors at UCLA, and to all my prior external research mentors who equipped me with the skills and knowledge to write this thesis. Thank you to Dr. Tonya Kane, Dr. Anthony Baniaga, Dr. Van Savage, Dr. Andrew Chang, and to the many more impactful research mentors and graduate students without whom this could not have happened. A final thanks to my wonderful colleague Mia Taylor for their phenomenal work on the C-PHAST website.

INTRODUCTION

Introduction to Phylogenetic Diversity

Biodiversity– described as the "vast diversity of life on earth" (Wilson 1988) is a central concept to modern conservation biology. Preservation of biodiversity is prioritized largely because it is correlated with a diverse representation of functional traits and consequently, with long-term sustainability of ecosystem services (Cadotte et al. 2011). As such, quantitative metrics of biodiversity have become part of the common toolkit of park managers and conservation scientists. Species richness, which is simply a count of the number of species within a region, remains the most widely used and readily accessible measure of biodiversity (Gotelli et al. 2001). Traditional diversity metrics that build from species richness include alpha measures of species per area, rank-abundance calculations, and beta diversity metrics comparing assemblages between communities (Moreno et al 2017). However, contemporary research has revealed that richness metrics applied in isolation fail to account for community evolutionary distinctiveness, and can result in extremely misleading decisions about how to best protect and preserve threatened areas (Gotelli et al. 2001, Buerki et al 2014, Miller et al. 2018)

To address these shortcomings, conservation scientists have developed a range of additional statistics to partially improve on species richness-based values. The introduction of phylogenetic diversity metrics to measure biodiversity marked a revolution in conservation studies (Faith 1994), as these measures frequently outperform evolutionarily naive methods in predicting key ecological function in a conservation context (Miller et al. 2018, Winter et al. 2013). Phylogenetic diversity considers not only abundance and richness, but also the relatedness among species, capturing the extent of evolutionary history represented in a community (Srivastava et al. 2012). While traditional alpha and beta metrics are useful in directing land

management practices and conservation techniques, an increase in availability of phylogenetic data allows a more direct assessment and evaluation of species-specific conservation value (Cadotte et al. 2010).

Ecological Importance of Phylogenetic Diversity

Phylogenetic diversity metrics are useful foremost in their approximation of functional diversity: we expect more closely related species to have more closely related traits (Winter et al. 2013, Lean et al. 2016). Functional diversity examines the range of biological functions performed by species and serves as an estimator of ecosystem resilience in the face of pulsating or rapid ecological change (Ceulemans et al. 2019). However, it can be difficult to quantify empirically for large scale communities without sufficient availability of trait data and with sporadic sampling accessibility (Swenson 2013). Quantifications of phylogeny-based evolutionary history can provide a reasonable approximation of functional diversity and capture the evolutionary potential of communities (Winter 2013, Cavender-Bares et al. 2009). As the scale of phylogenies increase to span large clades or multiple large taxonomic groups, traits follow increasing conservatism, and analysis of phylogenies becomes stronger in estimation of community functional diversity (Cavender-Bares et al. 2009). Under the assumption that traits on large phylogenetic scales are relatively conserved, metrics of phylogenetic diversity can, at the very least, help to approximate the underlying drivers of community assembly. Following this theory, abiotic and biotic forces acting on traits across clades can predictably drive patterns of phylogenetic signal that are captured in the communities we analyze.

When studied in the context of functional diversity, phylogenetic diversity provides valuable insight into community assembly processes. We see phylogenetic clustering, or

underdispersion, in a community that captures less evolutionary history than expected under neutral assembly given its size. This tends to occur when conserved traits dictate habitat suitability by influencing tolerance to abiotic conditions (Vamosi et al. 2009, Webb et al. 2002). Conversely, overdispersion occurs when a community captures more evolutionary history than expected under neutral assembly given its size, and arises when convergent traits are filtered, or when competitive exclusion occurs among species inhabiting similar niches (Vamosi et al. 2009, Webb et al. 2002). Measuring phylogenetic dispersion allows us to directly connect inferred trait evolution to underlying ecological filters. For example, a study by Montaño‐Centellas et al. 2019 details the latitudinal variation of both evolutionary and functional diversity in avian species, and highlights abiotic features like elevation and precipitation as candidate forcers on phylogenetic clustering. The evolutionary information gleaned from analyses like this is crucial in understanding and managing extant communities (Lean et al. 2016).

Linking phylogenetics, functional diversity and community assembly is vital when informing conservation priorities, as the incorporation of phylogenetic diversity metrics in policy has proven to increase biodiversity outcomes tenfold (Hartmann et al. 2013). Selective expansion of targeted protected areas, particularly in the context of diminishing land availability, has the potential to increase captured phylodiversity across clades in excess of 30% (Pollock et al. 2015, Rosauer et al. 2017). Unfortunately, current positioning of protected areas without phylogenetic planning often fails to capture evolutionarily distinct communities and high-risk species (Isaac et al. 2007).

Common Metrics of Community Phylogenetic Structure

Several methods and packages for numerical quantification of phylodiversity exist and can be applied to help us understand drivers of community assembly (Kembel et al 2010, Revell 2012). The first and most common metric is Faith's Phylogenetic diversity (PD) (Faith 1992), which is formally defined as:

$$\mathrm{PD} = \sum_{i=1}^{m} l_i$$

where m is the number of branches in the minimum spanning path of the tree that connects all the species in the set and l_i is the length of the ith branch. Here, branch length corresponds to the amount of evolutionary time passed. Essentially, PD answers the question: How much evolutionary history is captured in a species set?

Another common metric, Mean Pairwise Distance (MPD) is defined as:

$$\mathrm{MPD} = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} C_{ij}$$

where n is the number of taxa in the species set and C_{ij} represents the cophenetic distance between taxa i and j. MPD is sensitive to recent divergence in a community, and is a measure of the mean phylogenetic distance separating all pairs of taxa in a sample. MPD answers the question: How recent is the divergence present in a community?

Finally, Mean Nearest Taxon Distance (MNTD) is defined as:

$$\mathrm{MNTD} = \frac{1}{n} \sum_{i=1}^{n} \min_{j \neq i} C_{ij}$$

where n is the number of taxa in the community and $\min(C_{ij})$ is the cophenetic distance value between taxon i and its closest relative in the community, taxon j. MNTD answers the question: How closely related are sister taxa in a community?

While several other metrics for quantifying evolutionary history exist, when evaluated for ecological significance, many known metrics tend to cluster both in behavior across species richness and in core questions they address (Miller et al. 2017, Tucker et al. 2016). Of these clusters, MPD best represents mean relatedness, MNTD captures nearest-relative measures of community relatedness, and PD is indicative of total community diversity, thus tracking closely with species richness (Miller et al. 2017, Tucker 2017). Together, these metrics can accurately detect habitat filtering and competitive exclusion when assessed with richness-controlling null models (Miller et al. 2017). In deciphering community assembly patterns, Faith's Phylogenetic Diversity (PD) and Mean Nearest Taxon Distance (MNTD) consistently demonstrate robustness in detecting phylogenetic clustering driven by habitat filtering, exhibiting low type I error rates and high power across various null models (Miller et al. 2017, Kraft et al. 2007). Mean Pairwise Distance (MPD) focuses on mean relatedness and proves effective in identifying phylogenetic overdispersion resulting from competitive exclusion (Miller et al. 2017). PD and MPD represent "anchor" alpha phylodiversity metrics, with the former an encompassing statistic to explore richness-based questions and the latter a strong indicator of degree of divergence (Tucker et al. 2016). Ultimately, these metrics can be particularly useful in quantifying the underlying forces in community ecology that lead to variable community assemblages, and are thus vital in determining areas of highest conservation priority.

Spatial Phylogenetics

One important application of phylodiversity quantification is in spatial phylogenetics–defined by Mishler et al. 2023 as the transformation of phylo-geographic data into a GIS layer that can be visualized and analyzed in conjunction with abiotic geospatial features.

Using this method, biogeographic regions of interest can be studied comparatively with regionally-specific quantifications of evolutionary history (Scherson et al. 2017). Spatial phylogenetics has a wide array of applications, but can, for instance, be used to quantify the relative diversity of protected and non-protected areas, informing conservation tactics and determining how well reserves capture areas of phylogenetic concern (Thornhill et al. 2016, Zhang et al. 2022, Nitta et al. 2022). Many such studies have revealed interesting patterns of phylodiversity and endemism and highlight regions of high conservation value. However, they are often limited in scope, with an overrepresentation of plants and a lack of representation of large animal clades (Emerson et al. 2008, Vamosi et al. 2009). Relying on patterns of over- and under-dispersion of a single lineage to inform spatial phylogenetic-driven conservation efforts may fail to capture areas of significant evolutionary history in orthogonal clades (Emerson et al. 2008, Laity et al. 2015, Thornill et al. 2016).

In conservation management, the traditional focus on single species or specific clades often neglects many supporting or auxiliary species that are essential for sustaining ecosystem functions (Toffelmier et al. 2022). Failing to identify regions of significant phylogenetic diversity at various geographical scales results in poorly targeted conservation strategies and an irresponsible management of valuable ecosystem services (Veron et al. 2019). Initiatives exist to formalize and map global biodiversity across clades using traditional richness-based metrics (Ferrier et al. 2004). However, the few studies that have explored cross-clade phylogenetic diversity and the relative impact on conservation often find incongruences between species richness and evolutionary distinctiveness, highlighting major differences in the behavior of phylodiversity statistics across clades (González-Orozco et al. 2015, Laity et al. 2015).

Previous Spatial Phylogenetics Methods

The current commonly employed method for large spatial phylogenetic analyses is the Biodiverse (Laffan et al 2010) pipeline, which has been employed to analyze the expected spatial phylodiversity of many regions across the world (Mishler et al. 2014, Earl et al. 2021, González-Orozco et al. 2015). In this randomization method, a geographic area is divided into a rectangular grid, and true phylodiversity metric values are calculated for each cell. Bootstrapping randomizations then re-assign taxa randomly to each grid cell from the regional taxa pool, such that the new cell-specific community is equal in richness to the original community, and such that the total number of occurrences of each species in the regional pool is held constant. This method is repeated 1,000 times to generate 1,000 distinct communities for each cell. From these communities, phylodiversity statistics can be calculated to form bootstrapped null distributions, which can be evaluated against the true statistic in non-parametric tests.

Reliance on past methods requires an exponentially increasing amount of time for repeated randomizations with increasing species richness due to the combinatorial complexity and combinatorial explosion. As the number and size of communities observed in a given region increases, bootstrapping of the 95% confidence intervals will become expensive and redundant. For example, if we had 10,000 communities covering a geographic area, we would require 10 million randomizations to conduct our analyses.

The commonly used Biodiverse method also applies unnecessary constraints in analyzing range data, as it holds the number of occurrences of each species in a region constant across all cells through the randomizations. However, this tactic is not explicitly necessary when the community species pool does not contain count data (Miller et al. 2017). Using modeled species ranges can avoid sampling bias in occurrence data that causes error propagation in downstream

analysis, and provides a more comprehensive baseline for modeling changes in biodiversity with a changing climate (Howard et al. 2014). It is imperative that we are consistent in evaluating large cross-clade geospatial patterns of evolutionary history, and that the framework for spatial analyses is compatible with range data.

Current methods generally assume a fixed null species pool for metric comparison, and cannot accommodate variable spatial scales of analyses. Many spatial phylogeographic studies are restricted to a political boundary such as a state or country. Ecoregions and bioregions are geographic areas that partition areas based on ecology (largely floral communities), geography, and climate (Province 2006). Ecoregions are shown to strongly delineate biotic communities and conservation planning at the ecoregion level supports persistence at both the species and community level (Smith et al. 2018). The ability to focus on high conservation priority within both the larger political boundary and natural ecoregions using phylogenetic dispersion permits conservation prioritization via socially and ecologically informed spatial scales.

Finally, methods of spatial randomization separate areas of interest into grids of equally sized square pixels. Although rectangular lattices are common in ecological systems, hexagons are a better choice for spatial sampling and data visualization. Hexagons effectively tessellate a geospatial area without leaving gaps, while achieving a low perimeter-to-area ratio. This helps minimize sampling bias and edge effects compared to squares (Birch et al. 2007). Thus, to increase precision of geospatial patterns in phylodiversity, a hexagonal approach should be favored over a rasterized approach.

A New Tool for Spatial Phylogenetics

I introduce a new algorithm for enabling the rapid assessment of community evolutionary history called Community Phylogenetic Analysis at Speedy Time (C-PHAST). C-PHAST replaces expensive repeated randomizations with lookup tables, accommodates for the presence/absence form of range data, compares communities against both ecoregions and encompassing governing boundaries, and provides users with a high-resolution visualization of spatial phylogenetic dispersion.

Using C-PHAST we investigate the evolutionary history of California and test whether existing reserve structures capture exceptional diversity across birds, plants, mammals, squamates and butterflies. California is an excellent study system as it hosts more species of plants and animals than any other state in the USA, contributing to about one-third of all species in the country. Of the approximately 5,500 plant species in California, 40 percent are endemic, and the unique ecoregions harbor world-class biodiversity hotspots (CNRA 2024). Several recent phylogenetics projects (Goldberg et al. 2011, Toffelmier et al. 2023) explore speciation, extinction, and the evolution of regional diversity in California. However, very few studies have used spatial phylogenetics approaches to quantify the statewide phylodiversity, and those that exist focus on plant communities (Thornhill et al. 2017, Kraft et al. 2010). Here, I improve on current knowledge of spatial phylogenetics in California by extending my study to include not only plants, but also birds, mammals, squamates and butterflies endemic to the region. I also analyze, at a finer ecological scale, unique ecoregions within the state, many of which have been identified as areas of high biodiversity and endemism (Ricketts et al. 2003). In this paper I explore the following questions:

1. How do patterns of community phylogenetic diversity across large taxonomic groups compare to what might be expected under random assembly across California, and under random assembly within Ecoregions?
1. What are the relationships, if any, between patterns of phylogenetic dispersion across taxonomic groups?
2. How well do California protected areas capture phylogenetic dispersion?

Based on the literature, we might expect to see more clustering in less mobile clades, and stronger clustering under the California null model than the Ecoregion null model. This follows the evidence that at smaller scales, density-dependent interactions dominate, shifting to environmental filtering at intermediary habitat scale, with a further transition to domination of biogeographical processes at even larger scales (Cavender-Bares et al. 2009). Concerning outcomes in taxonomic group comparisons, existing studies have found little correlation between plant taxa and animal taxa in patterns of phylodiversity (Zupan et al. 2014), but some evidence of correlation within these groups and between internal clades (González-Orozco et al. 2015). While the five clades in this analysis have not yet been studied in tandem, we might expect low correlation in patterns of phylodiversity significance between plant and animal groups and higher correlation within clades in these groups. Finally, in examining California's protected areas, following previous findings of low or incomplete coverage of evolutionary distinctiveness, we expect that current reserve placement is not optimized to capture evolutionary history across the state (Aguilar-Tomasini et al. 2021, Saraiva et al. 2018).

METHODS

Overview of Community Phylogenetic Analysis at Speedy Time (C-PHAST)

The novel C-PHAST pipeline builds on existing methods to provide a tool to move from a collection of observation lists (ie: species in different reserves across California) to a high-resolution spatial analysis of phylodiversity. A key goal of spatial phylogeography is to identify regions where the community evolutionary history is either exceptionally overdispersed (meaning that the sample has more evolutionary history than expected given the size of the community) or underdispersed (meaning the opposite: that the sample has captured less evolutionary history than expected given its size). C-PHAST streamlines the calculation of the 95% confidence limits of expected evolutionary history for arbitrarily sized communities. To do this the pipeline relies on three core observations:

1. In order to estimate whether the evolutionary history of a community is over- or under-dispersed ($p < 0.05$), the true phylodiversity statistic of a given community can be compared to the expected 95% confidence interval of the statistic using a non-parametric bootstrapping test (Lean et al. 2016).
2. Of all possible alpha diversity metrics, significance testing of PD, MPD and MNTD perform best in detecting ecological processes such as habitat filtering and competitive exclusion (Miller et al. 2017).
3. For PD, MPD and MNTD, the 95% CI appears to vary smoothly as a saturating function of tree size (Miller et al. 2017).

These observations are key to understanding how C-PHAST enables rapid calculations of spatial phylogenetics. Using a traditional community phylogenetic approach on a focal spatial cell would require: 1) the identification of all species in that cell, 2) extraction of the subtree containing those taxa and calculation of true diversity metrics (PD, MPD, MNTD) from that tree, and 3) a series of randomizations constructing random communities of equal size cells sampled from the regional pool to determine if the observed diversity value falls into the tails of the expected distribution. Steps 1-3 are needed for every spatial cell of the mapped region, making the calculation of diversity metrics at fine scale computationally expensive. C-PHAST greatly accelerates these calculations by replacing the time-consuming simulations in step 3 with a lookup-table of the imputed values of the 95% confidence limits for a subtree matching the size of the community at hand. This is made possible by the well-defined behavior of PD, MPD and MNTD as they vary with tree size, behavior that can be predicted with a continuous function. C-PHAST accommodates both range and observational datasets and can be used to construct novel Contours of Evolutionary History (CEHs) across clades and geospatial areas. I introduce CEHs as surfaces that consist of two equations: one predicting the upper 95% CI of a phylodiversity statistic from richness and one predicting the lower 95% CI (Figure 1). CEH's allow a user to quickly determine whether a community with richness *n* is phylogenetically distinct. In this study, I apply the C-PHAST pipeline on California taxa across five major clades in order to gain an understanding of cross-state distribution of phylogenetic dispersion and to inform the direction of future management agencies.

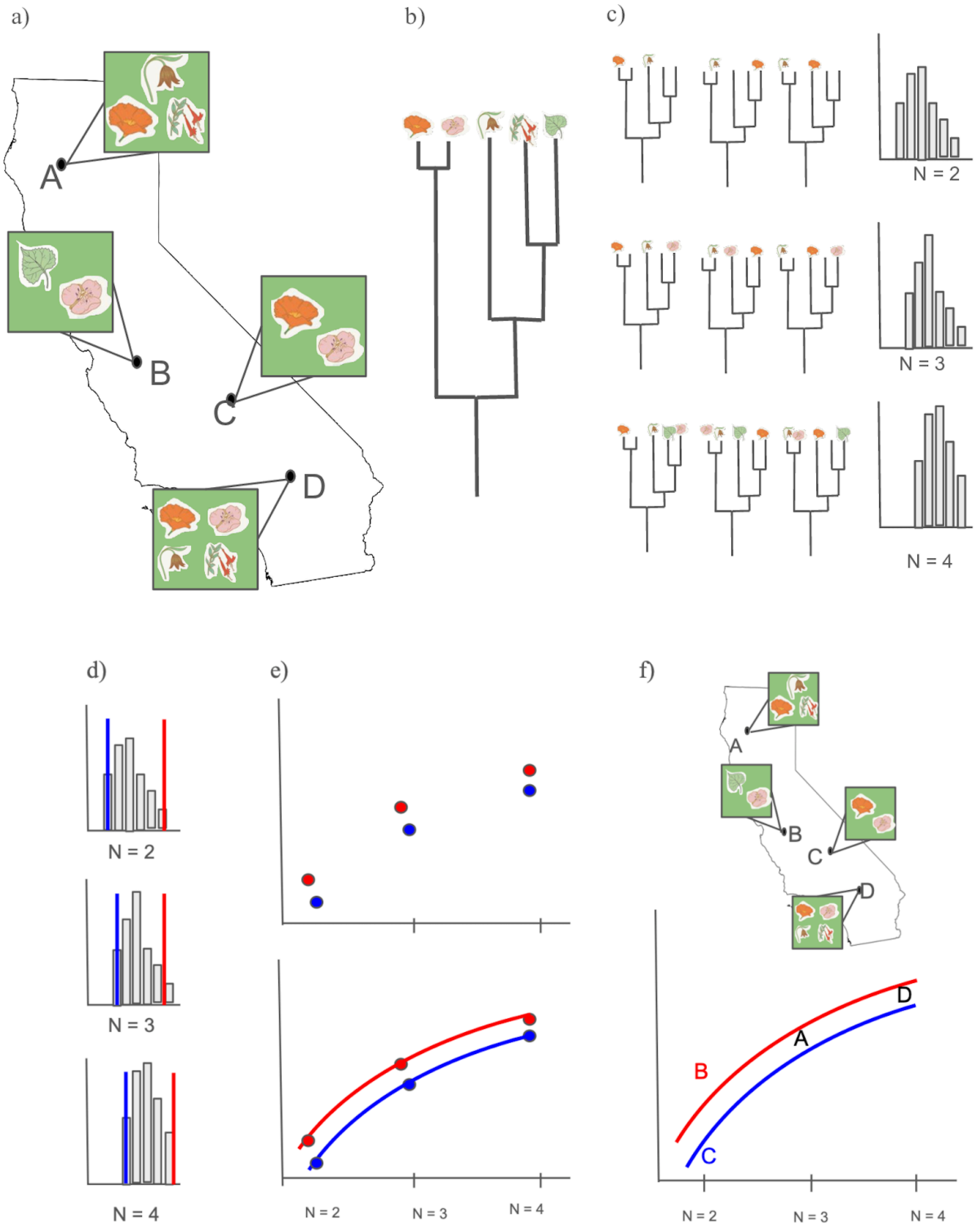

a)
A
B
C
D
b)
c)
N = 2
N = 3
N = 4
d)
N = 2
N = 3
N = 4
e)
N = 2
N = 3
N = 4
f)
A
B
C
D
D
A
B
C
N = 2
N = 3
N = 4

Figure 1: Illustration of C-PHAST pipeline. In this example, four plant communities are surveyed across California (community labels = A, B, C, D), and species composition for each community is recorded (a). A phylogeny representing all candidate species in the regional pool is constructed, and is used to calculate the true phylodiversity statistics for each community (b). These statistics can be any tree-based metrics, such as MPD, MNTD, and PD, and are not explicitly shown here. Bootstrapping randomizations to test for significance of the statistic of interest are performed for incremental richness values by reshuffling tips randomly, extracting sub-trees of size n = 2 to the maximum tip number, and calculating the given summary statistic for each sub-tree (c). In this case, three randomizations are shown, but in large trees 1000 randomizations are actually performed. The 95% confidence intervals from the bootstrapped randomizations for each incremental richness value are determined with non-parametric testing methods (d). Upper and lower CI bounds for each richness are plotted and a continuous saturating function is fit, producing two contours describing significance bounds varying with species richness, coined here as Contours of Evolutionary History (CEHs) (e). Finally, true phylodiversity statistics for each community A,B,C,D across California are compared to the imputed 95% CI bounds of the statistic as a function of species richness (f).

Partitioning California

In this analysis, I first partition California into communities in the form of geographic hexagons using the h3 package in R. Hexagons were selected at the level 6 resolution and were intersected with the boundary of California to produce a grid of 11,000, each with an area of approximately 36 km^2 (Supplementary birds, plants, mammals, squamates and butterflies 1). These hexagons represent unique geographic and ecological “communities” across California,

and can contain species from either (I.) All of California, or (II.) The ecoregion in which they are located.

The first regional species pool from which these communities can form is drawn from within the state boundary of California, determined by the CA.gov CA Geographic boundaries dataset found at https://data.ca.gov/dataset/ca-geographic-boundaries. This is appropriate for understanding how much of the total diversity in the state is captured at a local scale. The second regional species pool of comparison is based upon California's thirteen level IV ecoregions (Griffith et al. 2016). Communities compared to their respective ecoregional pools provide information about captured evolutionary history relative to that ecoregion, and might be more informative for conservation decisions at a more ecologically granular scale

Preparing California Phylogenies and Spatial Data

I used previously published phylogenies for vascular plants (Mishler et al. 2020), birds (Burleigh et al. 2014), squamates (Title et al. 2024), mammals (Upham et al. 2019), and butterflies (Kawahara et al. 2021) (Table 1). With the exception of birds, these trees were all time-calibrated in the parent source. I constructed a time-calibrated phylogenetic tree of extant bird species using the Burleigh et al. base tree of 7,000 extant taxa, and fossil calibration points from trees produced by Jarvis et al. and Oliveros et al (Burleigh et al. 2014, Oliveros et al. 2019, Jarvis et al. 2014). Details of this time calibration including the complete time-calibrated tree can be found in Supplementary Dataset I.

I collected range data from databases and previously published studies for the five focal clades. Shape data of ranges of plants, squamates and butterfly species was extracted from supplementary published data of external publications (plants: Baldwin et al. 2017, squamates:

Roll et al., 2022 butterflies: Shirey et al. 2021), while ranges of birds and mammals were collected from public databases with permission (birds: Birdlife International, mammals: IUCN). All range data was cleaned and manipulated in R using the SF package (Edzer et al. 2018). Files were first concatenated such that one large SF file was produced for each clade, in which each tuple in the dataset correlated to a species, and attributes consisted of range geometry and other species identifiers. I standardized all SF data to the WGS 84 Coordinate Reference System and all geometry was corrected using the function st_make_valid. All geometries were converted to the POLYGON or MULTIPOLYGON class, and the few remaining incompatible geometries in the form of MULTISURFACE were removed from datasets. All scripting required for data analysis can be found in the Supplementary Code.

Spatial joins between ranges and ecoregions were performed to create species-level taxa lists for each clade for all of California, and for 13 California ecoregions. Species were considered to have range overlap with an ecoregion if the range area intersecting the ecoregion covered at least 5% of the total ecoregion, or if the range area intersecting the ecoregion was greater than 20% of the total range area. This filtering excludes species from ecoregions in which they are not significantly or realistically represented, a conundrum that arose because range data was often rasterized or more coarse than ecoregion boundaries (Supplemental 2). Hexagon-level taxa lists were also created using the h3 package in R, for all hexagons covering the state of California. Hexagon-specific taxa lists for each clade were constructed through geometric intersection of clade-specific California ranger dataframes and each hexagonal area, using the sf st_intersection function. The final output of this step were regional species pools of birds, plants, mammals, squamates and butterflies for all of California, the 13 ecoregions, and the communities represented within individual hexagons.

I trimmed the global phylogenies for each clade to match the California and ecoregion species pools and created community phylogenies for each hexagon using the function drop.tips in APE (Paradis et al. 2019). The available phylogenies incompletely sampled the regional species pools based upon the shapefiles. To reduce the effects of incomplete sampling on calculation of biodiversity metrics, I collapsed phylogenies down to the level of genus for any group with more than 20 % missing species on average across hexagonal communities (Supplementary I). California-level phylogenies are available to download in Supplementary Dataset II for all five clades.

Fitting Contours of Evolutionary History

I considered a logistic family of functions as candidates to fit the contours describing the 95% CI on expected evolutionary history as a function of richness. This hypothesis space was chosen because the PD, MPD and MNTD statistics all exhibit saturating behavior within the limits of the min and max phylogeny size. I tested the following models: three, four and five parameter logistic and log logistic regressions, and the baroflex five-parameter log logistic regression (baro5). Models were fit to the bootstrapped upper and lower limits of the 95% CI's as a function of species richness across PD, MPD and MNTD for the all California regional pools for each major clade using the drc package in R (Ritz et al. 2016). I used Maximum Likelihood Estimation to determine best fit parameters for each model, and adopted an AIC-based model-selection framework across the seven selected candidate models for all clade and metric combinations. Fit was evaluated across clades and metrics using relative Akaike scores (Supplementary III). Mean Absolute Percent Error (MAPE) was also calculated (Supplementary III).

C-PHAST on California

I constructed Contours of Evolutionary History (CEHs) for birds, plants, mammals, squamates, and butterflies at both the California and Ecoregion levels. Figure 1 depicts the general pipeline for the production of CEHs, corresponding to the method detailed here. I repeated all steps detailed below for all five clades of interest.

First, I produced 11,000 community-level phylogenies—one for each hexagon of interest (Figure 1a)—and calculated phylodiversity statistics (true PD, MPD, and MNTD) for each community using the R package picante (Kembel 2010). Then, I constructed 14 unique phylogenies per clade for use in proceeding analyses—one representing all taxa possible in California, and 13 representing all taxa possible in each ecoregion (Figure 1b). I also built a cophenetic matrix for each phylogeny to optimize computation for MPD and MNTD. Next, I applied a parallelized script to each phylogeny to simulate bootstrapped randomizations across tree sizes. In each script run, I calculated 1,000 iterations of PD, MPD, and MNTD for one clade, and one particular species richness value (n). This process involved shuffling the tips on the tree 1,000 times (for PD) (Figure 1c), or the rows of the cophenetic matrix 1,000 times (for MPD and MNTD), and using the R package picante to calculate PD, MPD, and MNTD for the given richness value (n) (Kembel et al. 2010). Thus, each script produced a single bootstrapped distribution of expected PD, MPD, and MNTD under the null hypothesis that taxa in a community of richness (n) are randomly assembled from the larger regional phylogeny (either the California regional phylogeny or ecoregion regional phylogeny) (Figure 1c). For efficiency in simulation, I rarefied the computed richness values ranging from n = 0 to n = (phylogeny size - 1) in increments of 2, 5, or 10, depending on the size of the phylogeny. I incremented trees with

under 50 total taxa by 2, trees between 50 and 350 taxa by 5, and trees greater than 350 taxa by 10. I aggregated results using a Python script that cleaned the data, removed erroneous values, and filtered errors from parallelization. I then used nonparametric testing methods to calculate the bounds of the 95% confidence intervals of expected PD, MPD, and MNTD for each richness value by finding the richness values n, for which p(N<= n) = 0.025 and p(N>=n) = 0.025. The output from this script was a data file including the tree size/species richness, and the corresponding upper and lower boundary points of the 95% confidence intervals for PD, MPD, and MNTD (Figure 1d). I then exported this file to an R script and fit baro5 models for each metric (PD, MPD, and MNTD) using the R package drc (Ritz et al. 2016). I fit two models per metric: one regressing richness with lower-limit confidence values, and one regressing richness with bootstrapped upper-limit confidence values (Figure 1e), forming a Contour of Evolutionary History (CEH). The final model output consisted of the maximum likelihood best-fit model parameters for each metric/null model combination. Thus, for each combination of clade and null region of interest (e.g., California birds, Central Valley squamates), I produced a five-parameter equation that could then be used to quickly predict the upper and lower confidence limits for PD, MPD, and MNTD given a richness value.

To evaluate California phylodiversity given these null models, I compared true community phylodiversity statistics for each hexagon to expected bounds of PD, MPD, and MNTD predicted under neutral assembly across all of California, and within individual ecoregions. For each hexagon, I extracted the species richness as input to clade-specific equations to approximate the upper and lower 95% CIs expected for that richness in that clade. I then compared the true PD/MPD/MNTD values to values from interpolated 95% CI boundaries expected under the California regional model and the Ecoregion regional model. Hexagons with

true statistics above and below the confidence intervals have a value of $p < 0.05$ and are considered statistically significant (Figure 1f). In the ecoregion-level analysis, if a cell was located on an ecoregion boundary, I used a weighted average of significance compared to all overlapping ecoregions to assign a single significance value to that cell at the ecoregion level. The output from this step consisted of a dataframe containing hexagon IDs as tuples, and significance values of PD, MPD, and MNTD for birds, plants, mammals, squamates, and butterflies compared to both California and Ecoregion expected CIs.

Finally, I calculated cumulative phylodiversity significance by summing significance at each hexagon across all combinations of null model and phylodiversity statistics. For example, if hexagon ID #1 had significantly high PD in birds and mammals but normal PD in the remaining clades, the cumulative significance value for that hexagon's PD would be 2. I repeated the same for each hexagon for all three statistics. I visualized plots of cumulative California phylodiversity using the R package ggplot2 (Wickham 2016).

Comparing C-PHAST to Biodiverse

I compared the C-PHAST method to the Biodiverse package, the leading large-scale spatial phylogenetics tool. I assessed the evolutionary history of California plants using C-PHAST and compared these results to a similar analysis by Thornhill et al. 2017 using Biodiverse. Range data used in this study were identical to the range data used in Thornhill et al. While the phylogenetic treatment between the studies differs due to the OTU-based construction of the plant phylogeny in Thornhill et al, the molecular data and fossil calibrations are the same, and there is general agreement in the major features of both trees. To compare the results from these studies, I transformed Figure 2a from Thornhill et al. describing PD significance across

California to match the geography and color scale of the C-PHAST output figure, with red areas representing overdispersion, blue areas representing underdispersion and white areas representing insignificance. Color-matching was then used to identify differences and similarities between the two maps.

Cross-Clade Correlations

I analyzed processed matrices containing each hexagon across California and tuples with PD, MPD and MNTD significance across clades and null models using a Spearman's Rank Correlation Test. Pairwise p-values were calculated using the rcorr function in the Hmisc R package (Harrel et al. 2023) for each metric/regional model combination. P-values below a significance threshold of 0.05 were considered significant. I constructed six independent correlation matrices, comparing PD, MPD and MNTD significance across clades under both the California and ecoregional models.

Analysis of Protected Areas

Protected Areas in California were extracted from the CA.gov open data portal at https://data.ca.gov/harvest/california-protected-areas. I selected Protected area management agencies based on cumulative number of reserves ($n > 2$), reserve "intentions" (excluding military bases, BLM and other non-restricted land), and geographic breadth. I removed reserves managed by agencies with fewer than five statewide protected areas, and reserves with low multi-ecoregion representation. Each management system meeting these conditions was analyzed for the total protected area that covered regions identified to be either significantly overdispersed, underdispersed or not significant. I determined potential locations for future

reserves by visually identifying geographic clusters of phylogenetic overdispersion in areas that were not overlapped by at least one existing reserve.

RESULTS

California Phylogenies and Spatial Data

Five phylogenies specific to California were constructed for birds, plants, mammals, squamates and butterflies. A total of 107 squamate species were found to intersect California based on range data, of which 93 have phylogenetic information. 172 mammals were found to intersect California with a total of 145 representative species on the phylogeny. Of the 383 bird species that intersect California, 350 were represented by the larger bird phylogeny. Of the 245 butterfly species with ranges overlapping California, only 63 were represented, and of the 5220 plant species endemic to California, 2025 were represented phylogenetically. Hexagon-specific incomplete species sampling was calculated using:

$$H_{\text{loss}} = \frac{|S_{\text{tree}} \cap S_{\text{range}}|}{|S_{\text{range}}|}$$

Where S represents the set of species and |S| the cardinality of that set. This computation yielded under 20% median loss for birds, squamates, and mammals at the species level. Due to high incomplete sampling based on this analysis, butterflies and plants were analyzed at the genus level. In this reduction, 640 out of 992 plant genera are represented. Of the 97 butterfly genera that overlap with California, 86 are represented phylogenetically (Supplementary I). Reduction resulted in a median 10% hexagonal loss for butterflies and a 24% loss for plants (Table 1).

Clade	Phylogeny Source	Range Data Source	Taxonomic level used	Percent Representation
Birds	Burleigh et al. 2014	Birdlife International 2022	Species	91.3
Butterflies	Kawahara et al. 2023	Shirey et al. 2021	Genus	88.6
Mammals	Upham et al. 2019	IUCN 2023	Species	84.3
Plants	Mishler et al. 2020	Baldwin et al. 2017	Species	64.5
Squamates	Title et al. 2024	Roll et al. 2022	Species	86.9

Table 1: Data sources for parent phylogenies and range data, in addition to taxonomic level used and percent representation across clades. Percent representation is the total number of species present in a phylogeny that also intersect with the area of interest geographically, divided by the total number of species that intersect with the area.

Contours of Evolutionary History

In order to maximize likelihood and minimize overfitting, the five-parameter baroflex dose response function (baro5) was chosen as the best fit of PD, MPD and MNTD across clades (Supplementary III). The function is defined as follows:

$$\text{baro5}(r) = c + \frac{d - c}{1 + f \cdot \exp(b1 \cdot (\log(r) - \log(e))) + (1 - f) \cdot \exp(b2 \cdot (\log(r) - \log(e)))}$$

where

$$f = \frac{1}{1 + \exp\left(\frac{2 \cdot b1 \cdot b2}{|b1 + b2|} \cdot (\log(r) - \log(e))\right)}$$

Here, r is the richness, and d, c, b1/b2 and e are parameters. The outcome of this function is a weighted sum of two exponential functions, each modulated by b1 and b2, transitioning smoothly between the limits c and d. The dynamic parameter f adjusts how much influence b1and b2 have based on the value of r. As r changes, f changes, which in turn changes the contributions of the exponential terms in the denominator, ultimately affecting the overall value of the output. Average AIC for the baro5 model was significantly lower than all compared log logistic functions (Supplementary III). The baro5 function was consistently scored with >90% relative likelihood for both PD and MPD across clades. MNTD was divisive, with no clear majority, but relatively strong performance between LL.4, LL.5 and baro5. Numerical analysis yielded no difference in prediction outcomes for MNTD significance across clades using the three model candidates (LL.4, LL.5 and baro5), as all hexagons quantified as significantly low or high in a metric were exactly the same across model fits. Thus, baro5 was chosen across clades and metrics to maximize pipeline efficiency. In addition, baro5 yielded the lowest Mean Absolute Percent Error (MAPE) across models and clades, with a 1.06 % maximum deviation from predicted values and an average MAPE of 0.48% (Supplementary III).

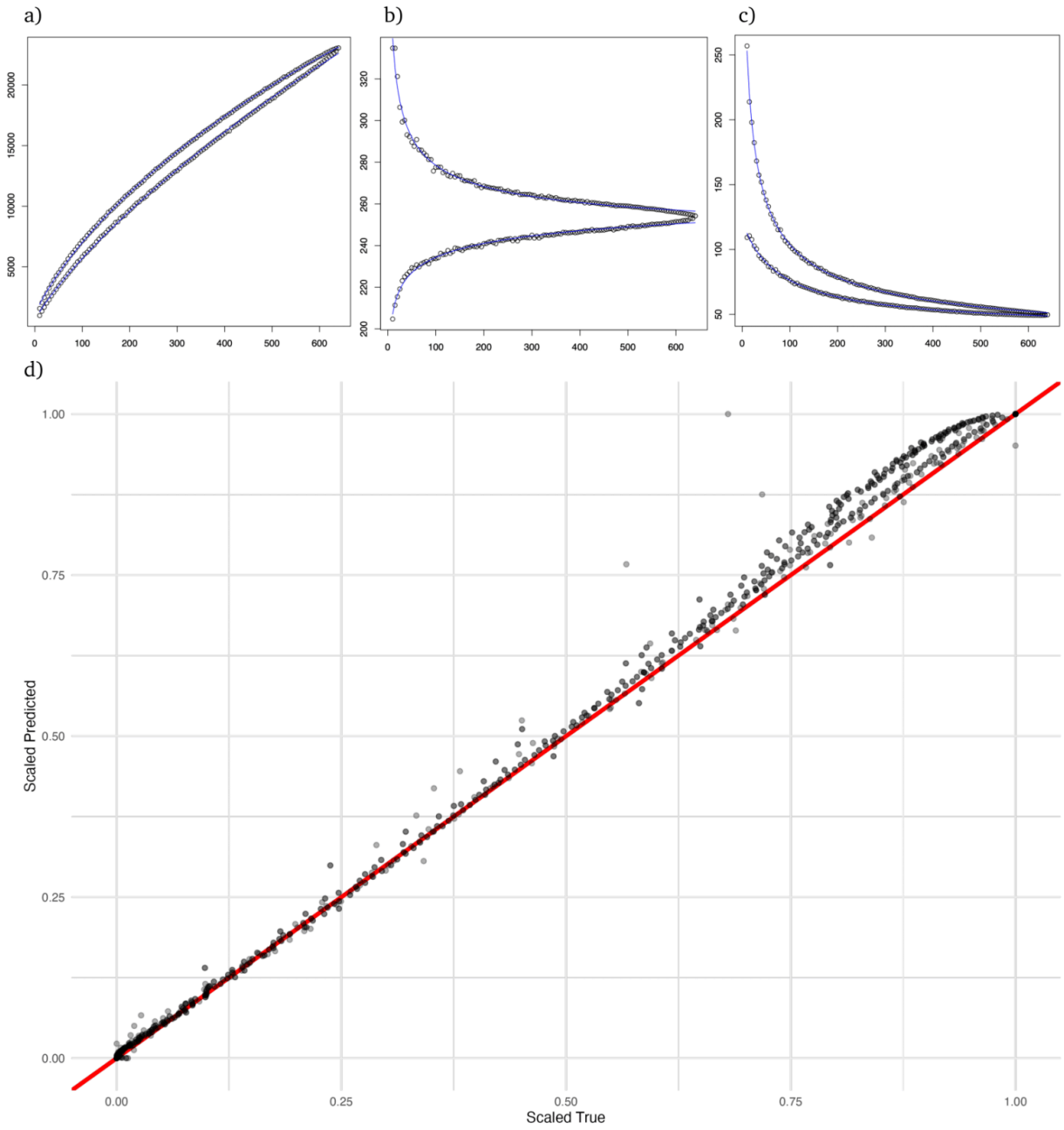


Figure 2: Baroflex five parameter log logistic model (baro5) fits for PD (a), MPD (b) and MNTD (c).These model fits correlate to the 95% CI's for plants of California, but a matrix of all model fits for birds, plants, squamates, mammals and butterflies under the California regional

model can be found in Supplementary III as can R^2 values for all model fits. Baro5 predicted value vs. true value plots of all models scaled on {0,1} and overlaid with the expected 1:1 ratio (d). The red line represents the perfect fit: scaled = predicted, and points represent data from all California models of PD, MPD and MNTD for birds, plants, mammals, squamates and butterflies. Individual true vs. predicted plots for metrics of all California models across all clades can be found in Supplementary III.

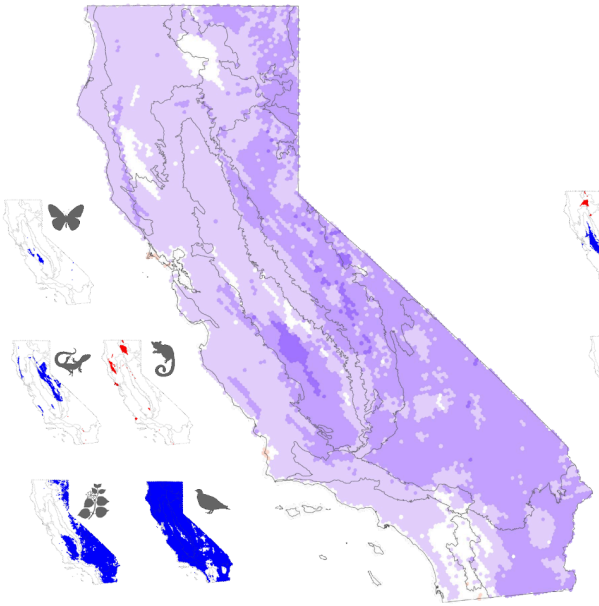
C-PHAST On California

Clade-wide analyses of birds, plants, mammals, squamates and butterflies were performed across 11,000 hexagonal communities within the state of California. When evaluated with the C-PHAST pipeline, communities that have true metrics that fall above the expected 95% CI for a statistic capture more phylodiversity than expected and are thus considered overdispersed. When the true statistic falls below the 95% CI, that community has less phylodiversity than expected and is considered underdispersed. No communities in California were significantly over- or under-dispersed in all five clades when compared to either the California or ecoregion-specific null models. However, several areas of interest are identified as having abnormally high or low cumulative phylodiversity across clades, particularly when compared to ecoregion-specific expectation (Figure 3). There appears to be high cumulative overdispersion in the ecoregion-border areas of the Southern California Mountains, the Baja California Mountains, and the Sonoran Basin and Range in the southern part of the state. Desert regions tend to be clustered in a fragmented way, with some rare patches of overdispersion in the Mojave basin and range. Northern California, at the intersection between the Cascade Ranges and the High North Coast Range, is overdispersed, as is the Western border of the Central Valley

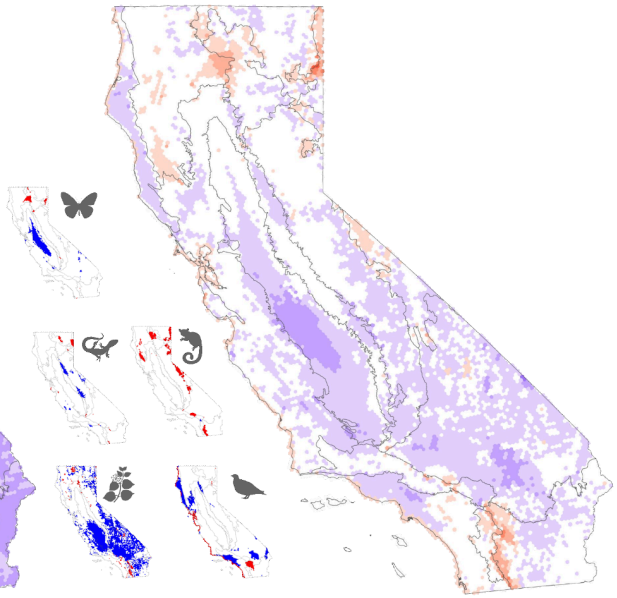
where it intersects with the Western Coast Range. Low clustering signal is present in some patchy areas within the Central Valley, and low signal present largely from overdispersion of birds is exhibited along the coastline. Finally, overdispersion is evident at the eastern edge of the Northern basin and range (Figure 3).

When compared to the null model of all of California, most communities are either underdispersed (phylogenetically clustered) or as expected under neutral assembly of all California taxa for PD and MNTD (Figure 3). Patterns of MPD were more even between patterns of cross-clade clustering and insignificance. Some small patches proved overdispersed in select clades and metrics, including the strip of the Coast range in Northern California (MNTD), and the Southern California/Baja Coast range in the South (MPD). In comparing the California regional model to the Ecoregion regional model, we see a reduction of coverage of significantly clustered hexagons across all three metrics in the latter, in congruence with an increased frequency of significantly overdispersed communities. Regional differences standout in the loss of overdispersion in the strip of the Coast range in Northern California and the Southern California/Baja Coast range in the South. In addition, the emergence of patchy overdispersion in Central Northern California in the ecoregion null model is not as strongly reflected in the California null model. Many regions like the Central Valley, which were phylogenetically underdispersed by the statewide model, fell into the expected range under the ecoregion-specific hypothesis.

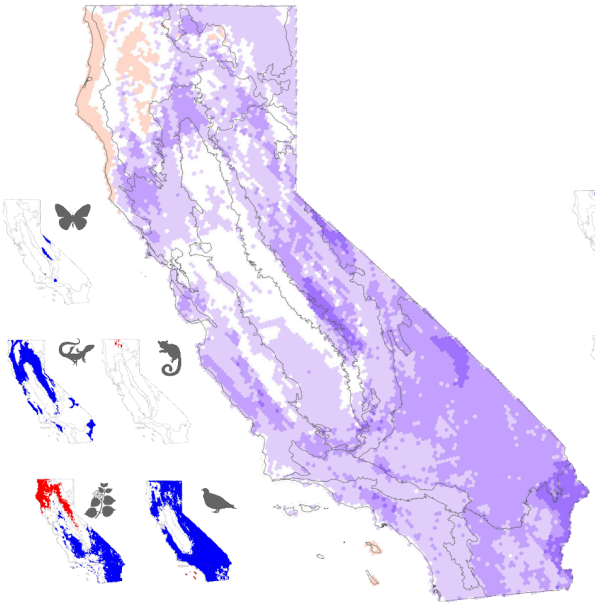
a)



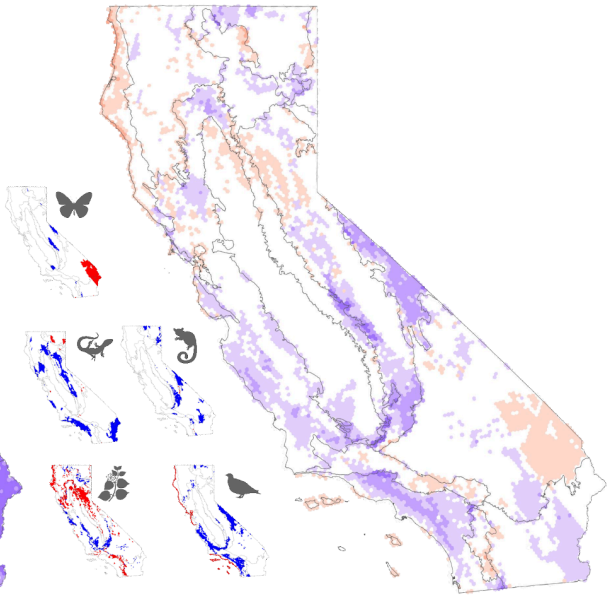
b)



c)



d)



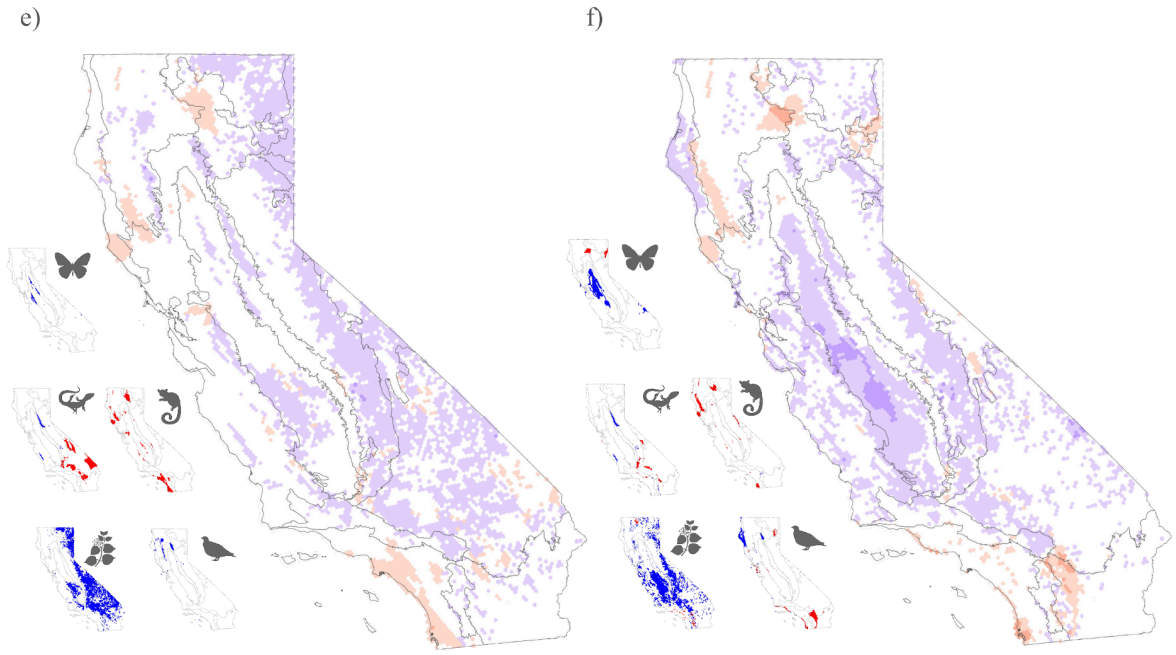


Figure 3: Cumulative Significance of PD compared to California (a), and compared to ecoregions (b), MPD compared to California (c) and compared to ecoregions (d), and MNTD compared to California (e) and compared to ecoregions (f) across all five clades of interest (plants,birds, mammals, squamates and butterflies). Significant cells relative to the California null model (left), and significant cells relative to ecoregion-specific null models (right). Cumulative phylodiversity per cell was calculated by summing significance values (1 if significantly overdispersed, -1 if significantly underdispersed and 0 if normal) for each cell. Theoretical ranges thus scale from -5 (dark blue) to 5 (dark red), although no single cell was significantly high (+5) or low (-5) across all five clades.

Comparing C-PHAST with Biodiverse

Major patterns in phylodiversity across California flora produced by C-PHAST mimic those found in a similar study by Thornhill et al. using the Biodiverse package. In order to

benchmark C-PHAST, a map of the Phylogenetic diversity of California plants was created using the new method and compared to a similar map produced by Thornhill et al. 2017 (Figure 2). While Thornhill formalized 1,000 OTU phylogeny-based taxonomic units, the C-PHAST method simplifies all taxonomic information to the genus-level, representing a total of 640 plant genera. Major regions of significance including the Mojave Basin, Central Basin and Eastern Cascades reflect highly similar patterns in phylogenetic clustering between the two models (Figure 4b). In addition, both models predict scattered overdispersion in central northern California and in some desert regions near the Southern California Mountains and Desert intersections. Key differences include Mishler's identification of the North-Central Valley along the foothills and Coastal Mountain intersection as a hotspot for overdispersion, and C-PHAST's identification of the Southern Central Valley as a significantly underdispersed region. In addition, C-PHAST analysis has much higher resolution, with about 11,000 unique hexagons across the state, so areas of high and low diversity appear less clustered.

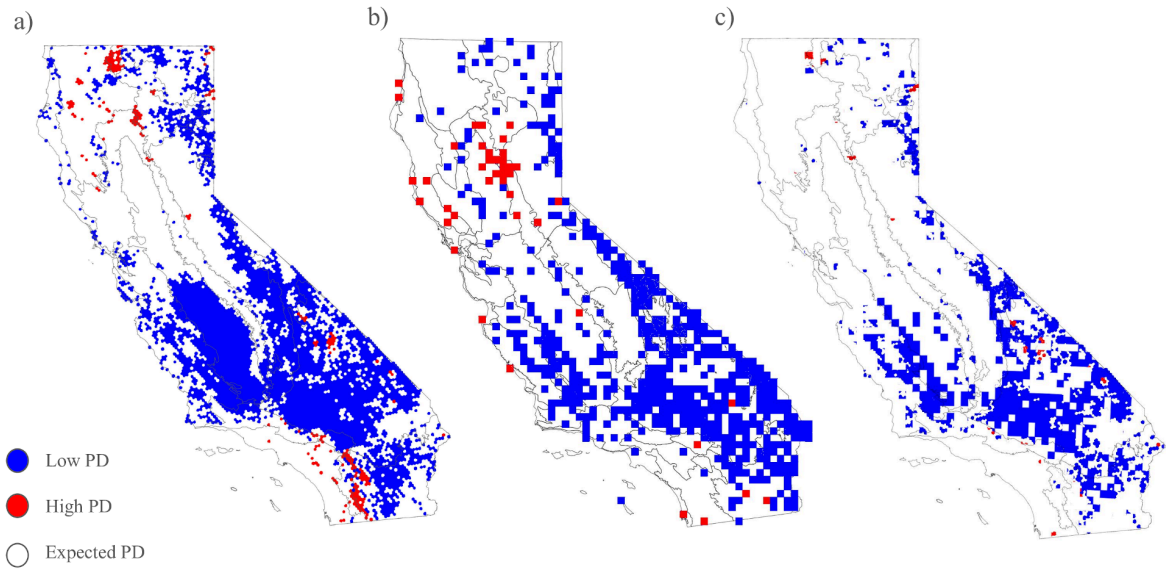


Figure 4: C-PHAST analysis of significantly phylogenetically over- (red) and under- (blue) dispersed areas across California (a) compared with Biodiverse model output from Thornhill et al. 2017 (b). Areas of agreement in both over- and under- dispersed regions between the two studies (c). Red regions represent areas significantly higher than expected in PD with $p < 0.05$, and areas in blue represent significantly lower than expected PD with $p < 0.05$. High agreement between the two figures in clustered areas is observed, with some inconsistencies in overdispersed regions.

Cross-Clade correlations

Pairwise Spearman's rank tests resulted in many significant correlations between clustering and/or overdispersion across clades (Figure 5). Of these correlations, none were found to be particularly strong, with the largest magnitude correlation -0.37 between plant MPD and squamate MPD in the California null model (Figure 5). When considering ecoregional models, there was pervasively less strength in correlation, and more correlations were found to be insignificant than when cells were ranked using the California regional model. Overall, 75% of pairwise comparisons were found to be significant in the context of both null models, though most correlations were relatively low. Most correlations maintained their sign across null models, with 25% switching or changing significance. A few discrepancies of note include a switch from weak positive correlation of MPD of squamates and butterflies in the California replicate to a weak negative correlation of MPD of squamates and butterflies in the Ecoregion replicate. MNTD showed the most variability in correlations between the California replicates and the Ecoregion replicates (Figure 5).

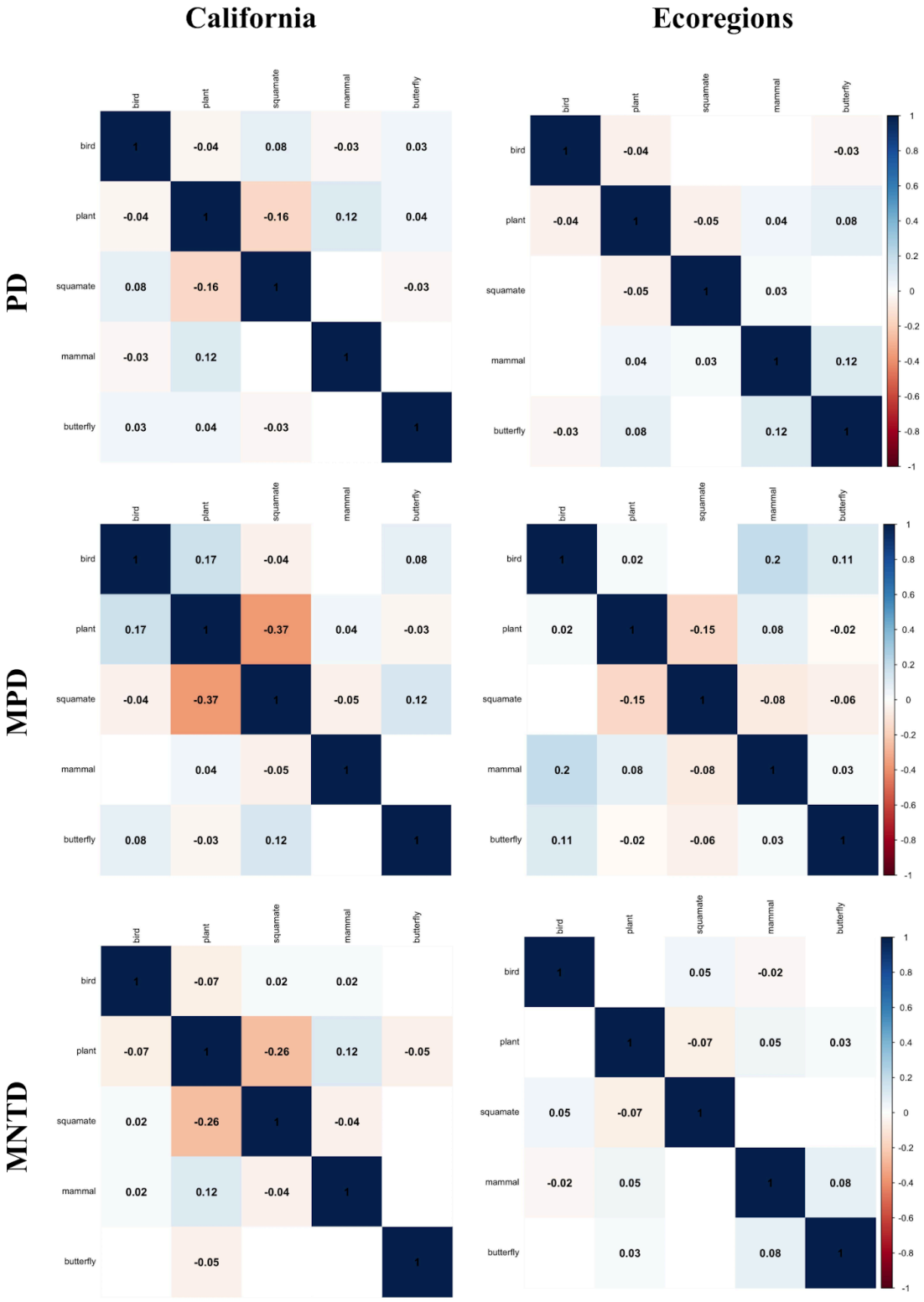

California
Ecoregions
PD
MPD
MNTD
bird
plant
squamate
mammal
butterfly
1
-0.04
0.08
-0.03
0.03
-0.04
1
-0.16
0.12
0.04
0.08
-0.16
1
-0.03
0.12
1
0.03
0.04
-0.03
1
1
-0.04
-0.03
-0.04
1
-0.05
0.04
0.08
1
-0.05
0.03
0.04
0.03
1
0.12
-0.03
0.08
0.12
1
1
0.17
-0.04
0.08
0.17
1
-0.37
0.04
-0.03
0.08
-0.37
1
-0.05
0.12
0.04
-0.05
1
0.08
-0.03
0.12
1
1
0.02
0.2
0.11
0.02
1
-0.15
0.08
-0.02
-0.15
1
-0.08
-0.06
0.2
0.08
-0.08
1
0.03
0.11
-0.02
-0.06
0.03
1
1
-0.07
0.02
0.02
-0.07
1
-0.26
0.12
-0.05
0.02
-0.26
1
-0.04
0.02
0.12
-0.04
1
-0.05
1
0.05
-0.02
0.05
1
-0.07
0.05
0.03
0.05
-0.07
1
-0.02
0.05
1
0.08
0.03
0.08
1
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Figure 5: Symmetric Spearman's rank correlation matrices for pairwise comparisons between all clades for all metric significances. Correlation tests were performed on discrete ranked data valued in the set {-1,0,1}. Rows represent metrics (top = PD, middle = MPD, bottom = MNTD) and columns represent null models evaluated (left = California, right = Ecoregions). Darker red values indicate a negative correlation, meaning as rank of one variable increases, rank of the pairwise variable decreases. Darker blue values indicate a positive correlation, meaning when the rank of one variable increases, then the rank of the pairwise variable also increases. A value of -1 means perfect negative correlation, and a value of 1 means perfect positive correlation. Correlations found to be non-significant ($p > 0.05$) are shown as blank.

Phylodiversity of California's Protected Areas

Of the 19 management agencies of protected areas in California analyzed, the majority of reserve lands cover neutral or under-dispersed regions (Supplementary V). When compared to California, reserves were overwhelmingly clustered in PD, with some representation of neutral and overdispersed regions in MPD and MNTD. When compared to ecoregions, the majority of reserve area covered neutrally dispersed hexagons, with roughly 10% overdispersed area across metrics on average. Of all reserve management systems, the Northcoast Regional Land Trust had the highest relative area covering overdispersed zones, and the Forest Service, Department of Fish and Wildlife, and National Audubon Society all had close to 25% land overdispersed in at least one metric.

University of California reserves performed relatively poorly in capturing areas of high phylogenetic diversity. Of the reserve area occupied, 11% had significantly high PD, 14% had

significantly high MPD and 1% had significantly high MNTD. Conversely, UC does an adequate job in capturing low phylodiversity and clustering, with 51% coverage for significantly low PD, 72% coverage for low MPD and 78% for low MNTD. Remaining UC reserve areas lie in overlapping regions with no significance (Figure 6).

95.5% of California communities were identified as significantly clustered in at least one metric when analyzed under the statewide null model. Conversely, 9.0% of hexagons covering California were found to be overdispersed. When compared to ecoregions, 57.6% of hexagons were clustered and 20.2% were overdispersed in at least one metric. In the context of the ecoregion comparison, 77.2% of the hexagons identified as overdispersed by a significantly high value of PD, MPD and/or MNTD overlap with a protected area or reserve. Of the clustered regions, reserves cover about 57.09% of available land.

In order to capture the maximum possible amount of evolutionary history (phylogenetic overdispersion across clades), potential areas of interest for future reserves were reduced to five main loci (Figure 7). The first area of interest is the boundary between the Central California foothills and the Northern Coast Range, which stands out in the North. Next, the High Lava Plains and Pluvial Lake Basin area in the north-east harbor high evolutionary history. The Sierra Madre Mountains at the border of the Central foothills and the Southern California Coastal Mountains are another candidate region, as are the Central Sonoran/Colorado Desert Mountains and the Eastern Mojave Arid Regions (Figure 6). These areas were chosen both because of their

net positive significance in evolutionary history and for their land availability.

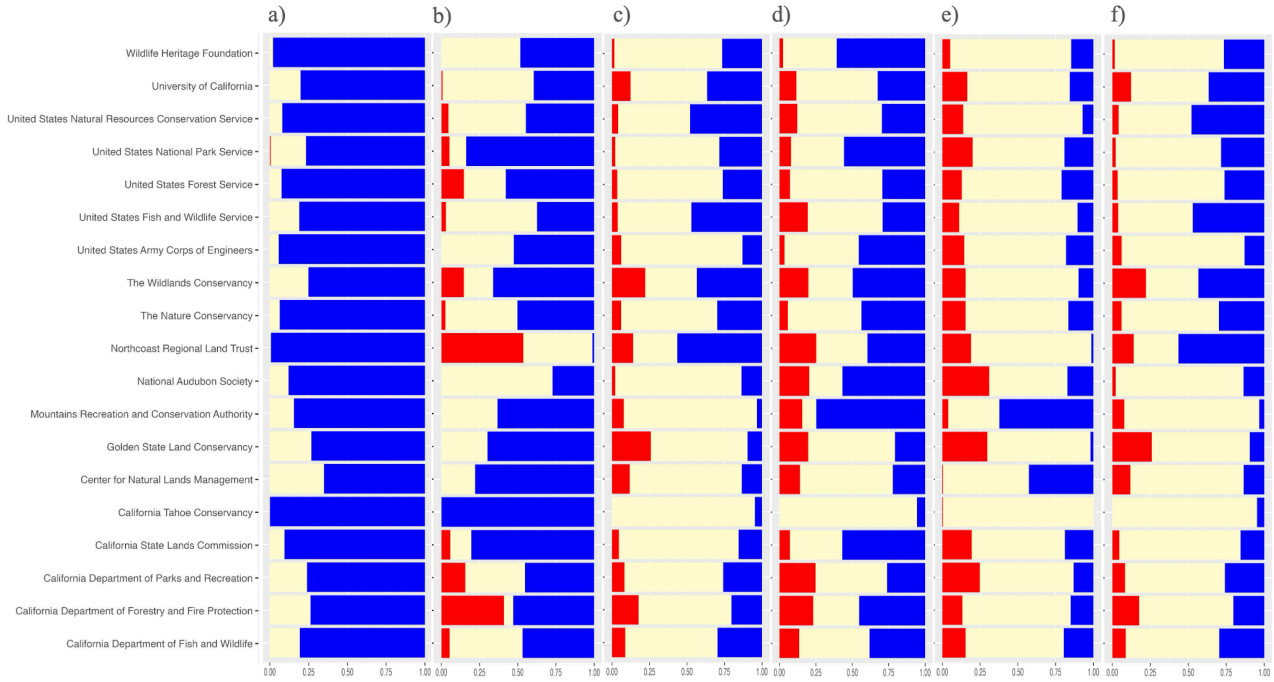


Figure 6: Proportion of reserve areas covering phylogenetically clustered (blue), neutral (cream), and overdispersed (red) areas across null models and metrics. From left to right panels represent: PD relative to California (a), MPD relative to California (b), MNTD relative to California (c), PD relative to Ecoregions (d), MPD relative to Ecoregions (e), MNTD relative to Ecoregions (f). Reserve management agencies that are represented must have at least five distinct sites across California.

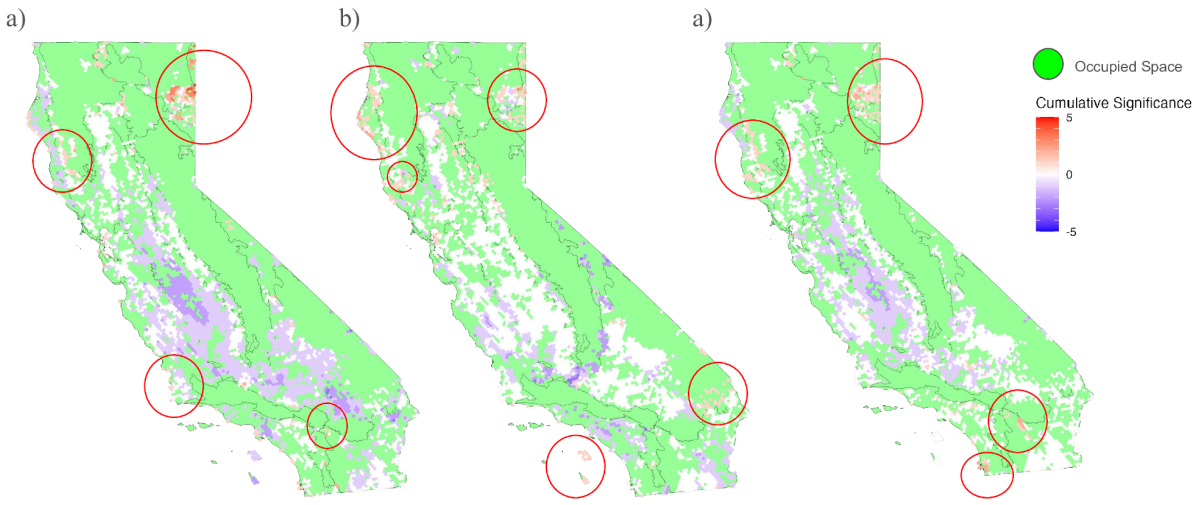


Figure 7: Areas of high, low and expected phylodiversity in PD (a), and MPD (b), and MNTD (c), as compared to the Ecoregion model, overlaid with pixels occupied by already managed areas (green). Red circles represent areas of interest and focus for the planning of future reserves, which are not already managed by one of the 19 participating management agencies and which exhibit cumulative cross-clade phylogenetic overdispersion.

DISCUSSION

Overview of the Project Goals

Quantifying wide-scale community composition is integral to understanding how humans can work to preserve existing biodiversity in the midst of a changing landscape and climate, and shifts in global ecology (Mokany et al. 2015). Two key phenomena of community ecology: clustering and overdispersion, can be computed using phylogenetic algorithms and can provide insight into evolutionary and ecological drivers of assembly.

Many studies have focused on spatial phylogenetic dispersion and community composition in a single clade (Emerson et al. 2008). This study extends this analysis framework to multiple taxonomic groups (birds, plants, mammals, squamates and butterflies), providing a cross-clade analysis of the phylodiversity present within California ecosystems. The novel C-PHAST pipeline presents an optimized tool that links large geospatial patterns of cross-clade phylogenetic diversity to a potential conservation plan.

C-PHAST takes advantage of the universal shape of expected PD, MPD and MNTD (as these metrics vary with tip number) to construct continuous functions of richness best fit by the baroreflex five parameter log-logistic growth function. I introduce a new phylogeny attribute– the Contour of Evolutionary History (CEH)– that provides a quick and accurate estimate of the bounds of a 95% confidence interval of PD, MPD and/or MNTD for any given species richness. C-PHAST compares favorably to methods relying on bootstrap randomization, and accurately identifies general regions of significant phylogenetic signal at a much more granular and geometrically accurate level than existing frameworks.

I deployed C-PHAST to analyze the cross-clade community phylogenetic structure of California. California has an unusual richness of diversity and is often the focus of conservation-driven solutions to biodiversity loss. All phylodiversity metrics studied showed high variability and unique trends among and between clades. From these results, I identified areas of significant cumulative phylodiversity –places that capture an unexpectedly high amount of evolutionary history– as target areas for ecosystem preservation.

Plants

Plants exhibited a high level of clustering, particularly in PD and MNTD when compared to both California and Ecoregion-specific null models (Figure 3). Phylogenetic clustering shows some correlation with tree size and mobility– species capable of moving are more likely to disperse from their biogeographical history and thus on a large phylogenetic scale, the low dispersal ability of plants may predict high clustering across large geospatial areas (Cavender-Bares et al. 2009). Results analyzing floral PD of California are similar to those presented in Mishler et al. 2017, highlighting the reliability of C-PHAST and how it builds on previous work with higher granularity and more dynamic flexibility in regional model choice.

Environmental factors like fire and aridity in the Mediterranean climate, have been shown to force phylogenetic clustering in resilience traits like Hardseededness (Santana et al. 2020). Aridity can also cause different patterns of clustering within plant sub-clades (Massante et al. 2021). My results bolster this understanding, with evidence of phylogenetic clustering in some arid fire-prone mountainous regions (Figure 3). In addition, California deserts have been identified as hotspots of young lineages with low neoendemic age; a finding that correlates to my observation of high floristic phylogenetic clustering in the Mojave region (Figure 3; Kraft et

al. 2010). One novel result of C-PHAST is the tendency for overdispersed areas in MPD to overlap with ecoregion boundaries (Figure 3d). As these boundaries have been shown to reasonably dictate distinct biotic assemblages, it follows that edge habitats will present unique ecological patterns (Smith 2018). Biogeographic ecotones and crossroads are hotspots of beta diversity, richness, and functional radiation, and inhabiting species often have high trait plasticity and adaptability (Spector et al. 2002). The C-PHAST finding of high overdispersion at these boundaries supports the consideration of ecotones as urgent conservation priorities for preserving California flora.

Birds

Birds displayed significantly low PD and MPD compared to neutral community assembly across the state. Avian spatial phylodiversity is informed both by geographically-linked radiation events and by biotic trends. Patterns of clustering have been observed for northern temperate latitudinal regions, while overdispersion has been seen in more tropical areas (Barnagaud et al. 2014). C-PHAST findings in the California mode follow this pattern of underdispersion across large geographic regions, particularly in PD and MPD (Figure 3a). At the more granular ecoregion level, patterns of high coastal overdispersion emerge and are likely explained by the abundance of phylogenetically distinct shorebirds along the coast of the Pacific Ocean and Salton Sea.

Loss of bird phylodiversity in the modern era is connected to the pervasive increase in monocropping farmland, and land-sparing techniques in farming and land management can considerably slow this progression (Hendershot et al. 2020, Edwards et al. 2015). Thus, in considering preservation of functional diversity in birds, it is important to focus on regions with

expected or high phylodiversity in prospective agricultural areas. Incorporation of phylogenetic history and evolutionary distinctiveness into island-style reserve systems in birds has predicted longer term persistence of larger phylogenetic tree structure, supporting the validity of phylogeny-based conservation efforts in this clade (Jetz et al. 2014).

Mammals

Phylodiversity of mammals on both large and small geographic scales follows general patterns of overdispersion, as reflected by behavior of PD and MPD at the statewide and ecoregion-specific levels. Unexpectedly high areas of evolutionary history in mammalian communities could be explained by high-trophic level competition between closely related taxa where traits are conserved, and this pattern is consistent across spatially explicit phylodiversity studies (Cooper et al. 2008, Emerson et al. 2008).

Identification of mammalian lineages that capture a disproportionate amount of evolutionary history and display ecologically distinct and unique traits is instrumental in pinpointing high-risk target species (Isaac et al. 2007). The added layer of phylogeography in this analysis allows us to correlate these high risk species with their common environment, and directs focus toward geographic action in conservation. The North Coast is a strong candidate for mammalian phylodiversity conservation, likely due to the unique old growth forest ecosystems providing habitat refuge across niches (Figure 3).

Butterflies

Butterflies followed a pattern of relatively neutral phylodiversity across both the California and Ecoregion-specific null model. However, two notable regions of interest may be

ecologically significant – the Central Valley tends to be underdispersed in butterfly diversity across metrics and models, while the south-eastern desert region of the Sonoran Basin exhibits overdispersion in some scenarios. This pattern of high desert phylodiversity is reflected in the results of Earl et al. 2021, and may be explicable by seasonal migratory patterns of butterflies from Baja California to southern desert regions. Patchy overdispersion in northern regions is also reflected in Earl et al, although C-PHAST primarily shows neutral dispersion. This might be a result of scale– Earl et al. analyzes butterflies across all of North America while this study focuses on a much more specific general study area. In addition, spatial area per cell is much smaller in this analysis, which lends to higher precision.

Despite their driving ecological importance and strong co-evolution with endemic flora, insects are generally phylogenetically understudied (Buckley 2007). A deeper understanding of insect genomics is essential to advance phylogeny-based conservation efforts in this clade (Buckley 2007). Interestingly, patterns of butterfly and plant phylodiversity are not significantly correlated according to this study, raising questions of how co-radiation and endemism might drive diversification of regional plants and pollinators.

Squamates

As ectotherms, squamates have the unusual tendency to exist in high diversity and richness in arid and desert regions, as the cost of thermoregulation is relatively inexpensive (Šmíd et al. 2021). While spatial patterns of phylodiversity are not consistent across metrics in the case of MNTD (Figure 3 e,f) we see a tendency of overdispersion in arid Mojave and high northern desert regions. In addition, while there is evidence of correlation between increased phylogenetic diversity and richness on high mountains in the Arabian Peninsula (Šmíd et al.

2021), no work has studied whether patterns of dispersion in squamates is statistically significant or whether any observed patterns are connected to geography and climate. In this study, squamates consistently exhibited significant clustering in the High Sierra mountainous region. Past phylogenetic analysis of the Squamata clade has shown deep genetic divisions identified in regions such as the Transverse Ranges, Monterey Bay, Sacramento-San Joaquin Delta, and southern Sierra Nevada, suggesting that these barriers to dispersal have influenced the phylogeographical patterns of these species (Feldman et al. 2022). Thus, the high levels of clustering found in some of these regions might be expected for this clade.

Observed patterns of competition and coexistence in the Squamata clade is dispersal-limited in lizards, while in snakes sympatric patterns of coexistence persist in island populations (Alencar et al. 2023). Squamates are evidently not monolithic, and behavior, evolution, and functionality of independent sub-clades might explain the patchiness of over- and under- dispersion observed by C-PHAST.

Patterns of Cross-Clade Correlation

C-PHAST analyses follow existing studies that find weak evidence of correlation between plant taxa and animal taxa in patterns of phylodiversity (Zupan et al. 2014, Laity et al. 2015). Of note in both models across metrics, many correlations are significant but with low relative magnitude, and inconsistencies within pairwise comparisons and between metrics mirror existing inter-taxonomic analyses. Generally, correlation strength hovered below 0.15, with many pairwise comparisons (particularly in the ecoregion-level analysis) having insignificant or extremely low magnitude correlations. In European Protected Areas, weak overall covariance between clades was recorded for PD across birds, mammals and amphibians, with little overlap

in phylogenetically significant regions across clades (Zupan et al. 2014). This suggests that observed discrepancies in phylodiversity metrics across clades and low relative correlation may be attributed to phylogeny structure and diversification history, environmental barriers to habitat, and/or species mobility (Zupan et al. 2014). C-PHAST results align with cross-clade analyses across the globe that point to high variability in patterns of spatial phylodiversity across clades (Zupan et al. 2014, González-Orozco et al. 2015, Laity et al. 2015).

One result particularly visible in the California regional model is a tendency across metrics (PD, MPD and MNTD) for squamates to covary negatively with plants. This essentially means that as plants approach values of overdispersion, squamates cluster and vice versa. Although to our knowledge no study has explicitly identified squamates and plants as negative covariates in the same spatial area, across spatial phylodiversity studies there is a trend of in clustering of plants and a weak trend of overdispersion in salamanders (reviewed in Emerson et al. 2008). Squamate richness tends to align with more arid environments (Šmíd et al. 2021) while plants are generally avoidant of these regions, a phenomenon that is reflected in the high clustering of plants in the Mojave desert region in MNTD, and the concurrent overdispersion of squamates (Figure 3 e,f).

Toward Biodiversity Conservation

In the context of large-scale environmental change, considering multi-trophic and phylogenetically diverse clades in conjunction can help shed light on how anthropogenic or climate-driven changes in one clade might affect changes in another. An example illustrated by Cavender-Bares et al. 2009 details the following scenario. Suppose there is an endemic grassland with a native herbivore. When herbivores are present, they exert pressure on plant species,

favoring those with traits that confer defense or tolerance against herbivory. This selection can lead to the dominance of plant species with these traits, resulting in phylogenetic clustering within the community. Conversely, if herbivores are excluded or their pressure is reduced–say by the introduction of a strategic competitor, the dynamics within the plant community can change. This may allow for the proliferation of plant species that were previously suppressed due to unfettered herbivory pressure. In this scenario, the dominance of certain plant species may shift, potentially leading to changes in phylogenetic structure and community composition. Critically, both the plant and herbivore communities must be examined jointly, as they exist in a dynamic equilibrium that can only be understood using cross-clade phylogenetic analyses. The strong negative correlation observed in this study between California squamate and bird phylodiversity reflects the necessity of conducting these joint analyses across disparate taxonomic groups that co-occur environmentally, and begs questions about the underlying ecological interactions that produce this pattern. This result, however, is not universal, and the phylogeographic inconsistencies in clustering and overdispersion that we observe here speak to underlying variability in evolution and adaptability (González-Orozco et al. 2015). We might see correlation between plants and squamates, but why do we see nothing between butterflies and plants, or birds and butterflies, despite observed environmental co-occurrence? These results force us to think critically about what kind of diversity we are actually conserving when we consider umbrella species or taxonomic groups as representative of a larger functional ecosystem.

In order to preserve functional diversity and evolutionary distinctiveness across the tree of life, it is critical to focus on protecting areas that capture a disproportionate amount of cross-clade evolutionary history. Highly overdispersed communities have been identified as stable climate refugia, and thus are imperative in maintaining biodiversity in a rapidly changing

global climate (Mastrogianni et al. 2019). California conservation land management agencies were found to cover a total of 77% of the California area identified as hosting particularly high phylodiversity. While this is promising, 23% of cross-clade overdispersed regions are still unprotected. This result follows expectations. Reserves generally poorly or stochastically capture evolutionary history, as reserve management structures do not regularly highlight phylodiversity in planning. Consequently, protected areas have low relative global coverage of phylogenetic distinctiveness across taxonomic groups (Aguilar-Tomasini et al. 2021, Saraiva et al. 2018, Carvalho et al. 2016). Consideration of these phylodiversity metrics is vital in strategically selecting small geographic pockets that disproportionately capture evolutionarily distinct lineages, particularly in zones of anthropogenic development (Pollock et al. 2015). My recommendation of areas with particularly high PD, MPD or MNTD across clades is corroborated by an established method that maximizes overall intra- and inter-specific phylodiversity in spatial prioritization (Carvalho et al. 2017).

Using C-PHAST, I identify currently unprotected candidate regions of exceptionally high cross-clade phylodiversity within California as the the boundary between the Central California foothills and the Northern Coast Range, the High Lava Plains and Pluvial Lake Basin area in the north-east, the Sierra Madre Mountains, the Central Sonoran/Colorado Desert Mountains and the Eastern Mojave Arid Regions, and the Southern California Coastal Mountains. While Vandergast et al. 2008 identified several regions of the Southern California mountains as interspecific evolutionary hotspots of genetic divergence, the evolutionary potential of many other regions highlighted here are less well studied. Together, these diverse ecosystems are expected to house overdispersed ecological communities and thus are vital loci to target in future management strategies and choices.

Study Limitations

Range data and phylogenetic data were both major limitations of the study. Range data is often interpolated from observational data with variable degrees of accuracy and probability, and in this study range data were combined from several different sources employing different methods. Methods for range interpolation are often criticized in their assumption, the pseudoabsence approach, that a plant, animal or other organism could exist in a place where it has not been seen (VanDerWal et al. 2009). An alternative to range data is observational data, which has risen in popularity over the past decade. Observational data was not used in this study as it is highly variable in quality across and within clades, and is difficult to standardize as many observational data hubs are operated with a citizen science framework (Kosmala et al. 2016). Thus, improving the method in which hexagons in California are assigned species lists is imperative to standardizing this method, and maintaining objectivity in results.

Phylogenies collected in this study were the most complete of each clade to date, but incomplete sampling still ranged from 20-80% at the species level. It is thus imperative to continue to expand phylogenetic datasets to enable cross-clade comparisons to fill in gaps in phylogenetic diversity, and to improve coverage of evolutionary history. Although this study covers a large scope of the tree of life, many integral clades are missing from the analysis due to an incomplete checklist and range data, or incomplete phylogenetic data. As phylogenies and range data become more complete, this model will become more and more accurate in evaluating statewide evolutionary history.

Finally, the metrics of choice may be considered limiting in truly capturing information useful to biodiversity conservation. Many other metrics of phylogenetic diversity and endemism

exist, and while PD is a fundamental statistic in determining evolutionary value of a community (Faith 1992), there is contention as to whether PD truly follows patterns of functional diversity and attrition (Mazel et al. 2018). One potential metric of interest for future analyses is Evolutionary Distinctiveness (ED), which is useful in conservation analyses (Tucker et al. 2016). Importantly, the flexible framework of C-PHAST allows for the addition of new metrics like this over time, as long as their confidence bounds can be expressed as some function of species richness. Thus, this foundational tool opens a door to future incorporation and analysis of more complex community level analyses.

Conclusion

There exists a strong body of work exploring endemism, evolutionary distinctiveness and the role of spatial phylogenetics in conservation, and these methods are slowly extending to understanding the complex biodiversity of California. This study helps to tie together the collective lattice of research scattered across taxonomic groups and metrics, facilitating the construction of a cohesive baseline understanding of California's cross-clade endemic phylodiversity. C-PHAST breaks computational barriers to present an elegant and succinct pipeline for spatial phylogeographic analysis, and has revealed novel patterns in inter-clade phylodiversity correlations, cross-clade patterns, and conservation management efficacy. Future development of this program will further accommodate new clades, geographics, and statistics, and will continue to provide insight into patterns of evolutionary history vital for the conservation of our earth's biodiversity.

APPENDIX

Supplementary I: Regional Models (California and Ecoregions)

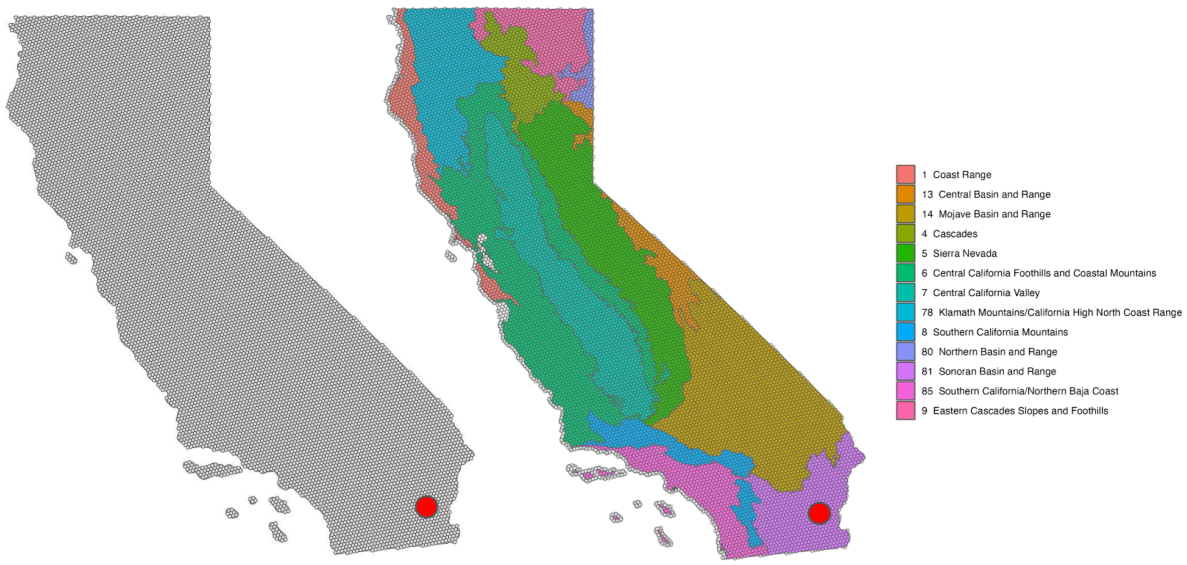


Figure 1: level 6-resolution hexagons covering California using the uber h3 package (left) and with overlaid ecoregions (right). Let's say that we are interested in understanding how the evolutionary history captured in the community represented by a red dot compares to expected evolutionary history. Null contours of PD, MPD and MNTD varying with species richness according to the figure (left) answer the question: What are patterns of phylodiversity under a statewide assumption of neutral community assembly? Contours varying with species richness according to the figure (right) answer the question: What are patterns of phylodiversity under an ecoregion-specific assumption of neutral community assembly? As seen in the figure (right) , instead of assuming random community assembly across California in the null hypothesis, we assume random community assembly within the Sonoran Basin and Range, and compare the red dot only to species available in that ecoregion.

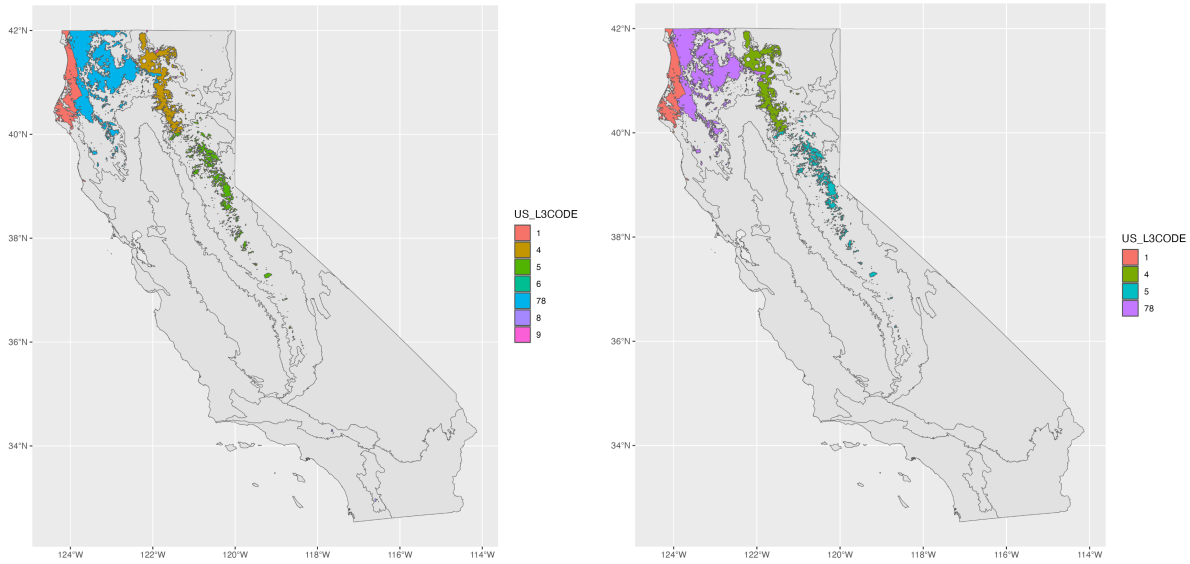


Figure 2: *Abies amabilis* range before ecoregion filtering (left) and after ecoregion filtering (right). As seen, the four obvious ecoregions (1,4,5,78) are preserved but artifact ecoregions that have extremely low representation in range data (6,8,9) are removed. This removal only informs the ecoregion regional species pool and does not affect the taxa occurring in each individual pixel.

Supplementary II: Phylogeny Selection and Trimming

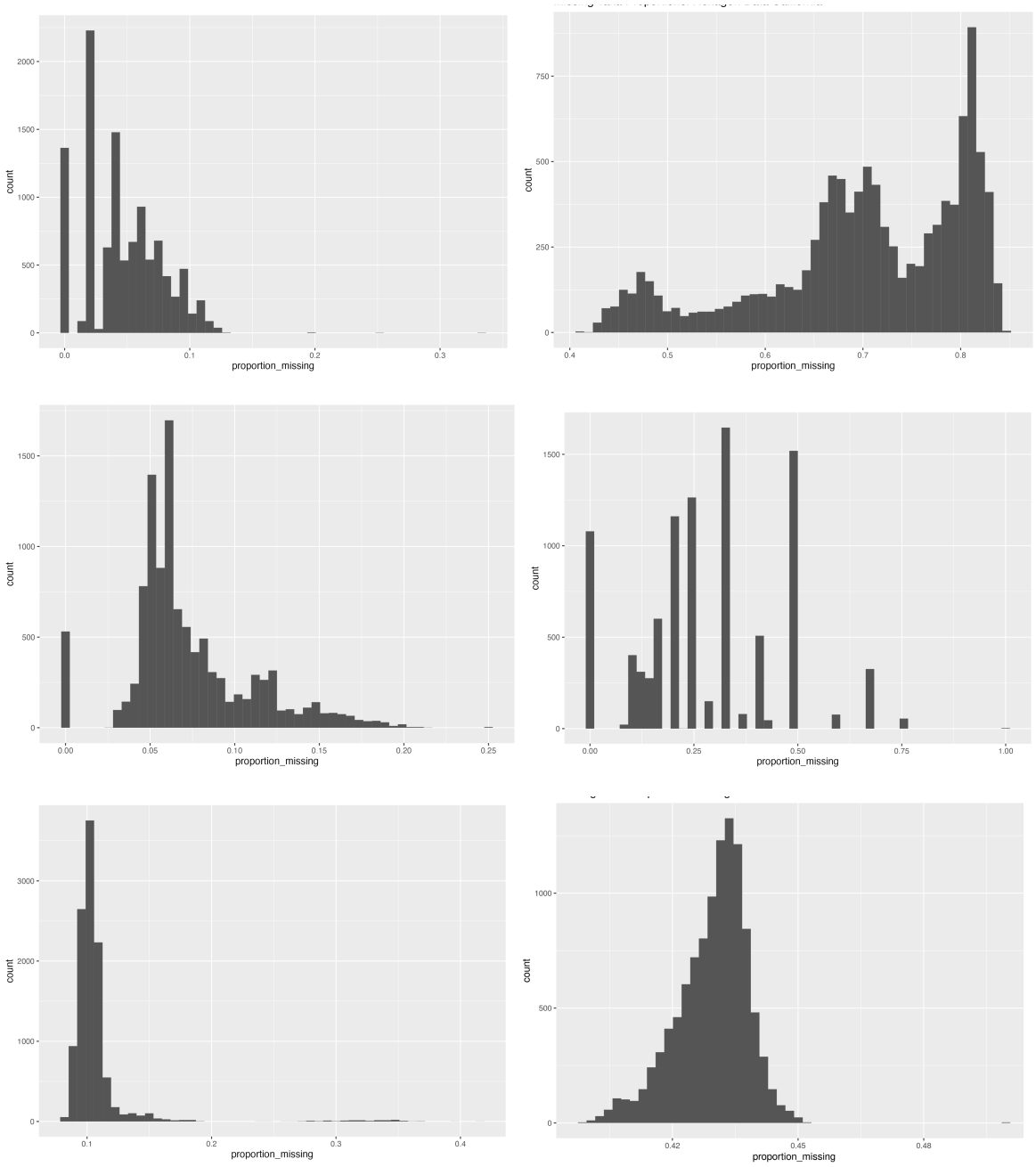


Figure 3: Histograms of incomplete sampling across hexagons for all 5 clades at the species level. From left to right and top to bottom: mammals, butterflies, squamates, amphibians, birds, plants. Amphibians were omitted from final analysis due to sparsity in range coverage.

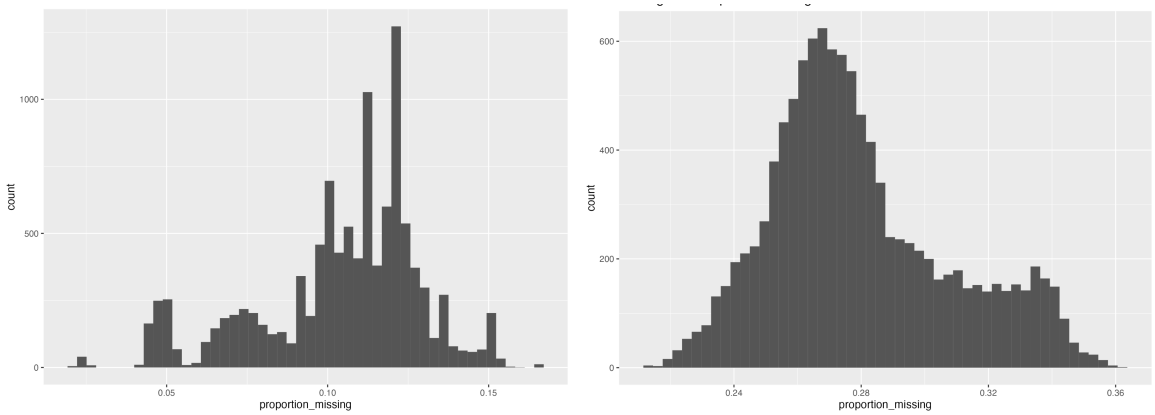


Figure 4: Genus-level histograms of incomplete sampling across hexagons for butterflies (left) and plants (right) after analysis was trimmed to the genus level.

Supplementary III: Model Fits

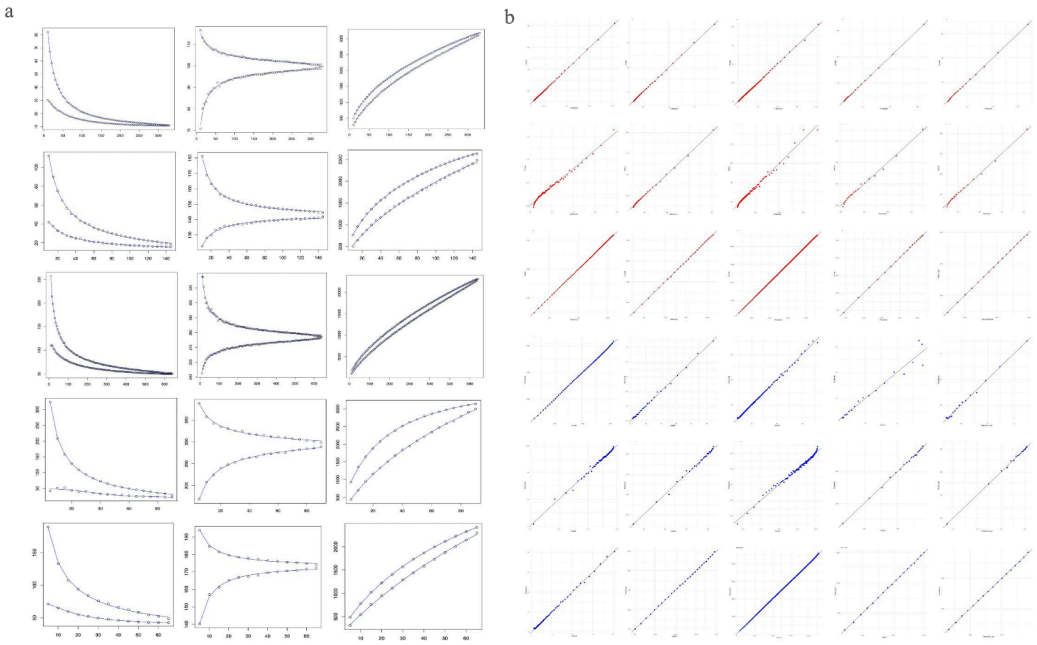
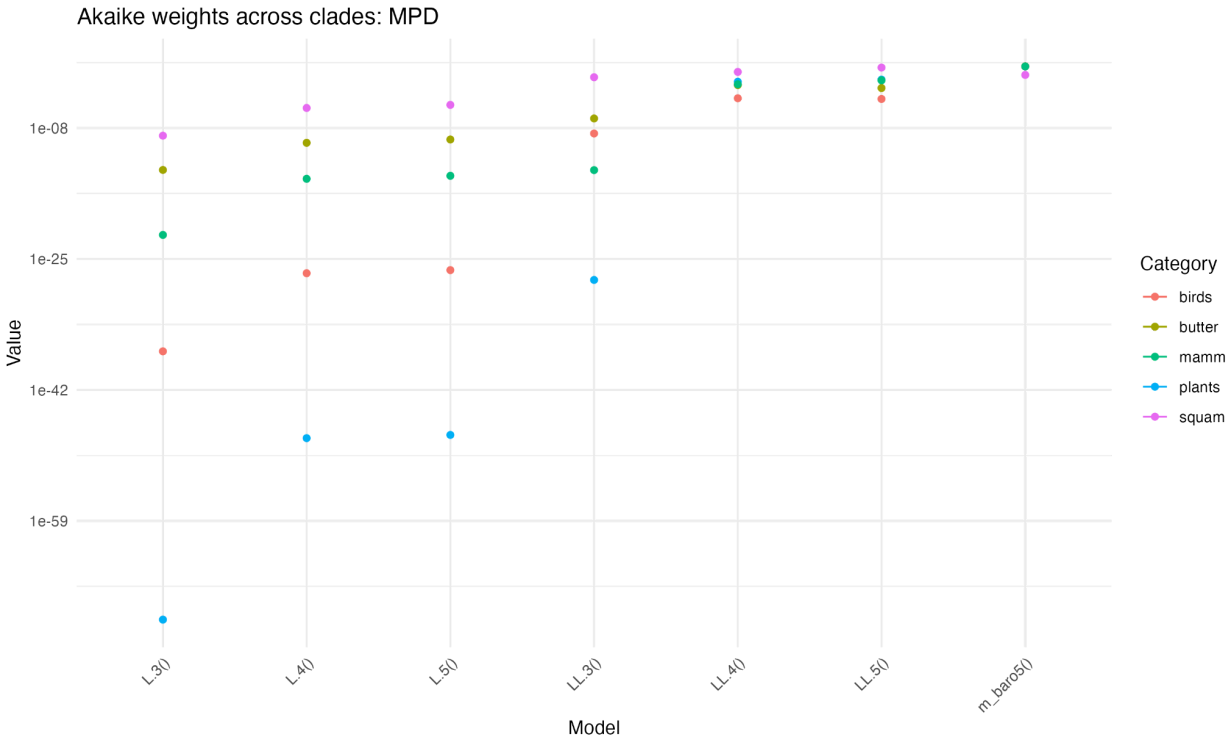
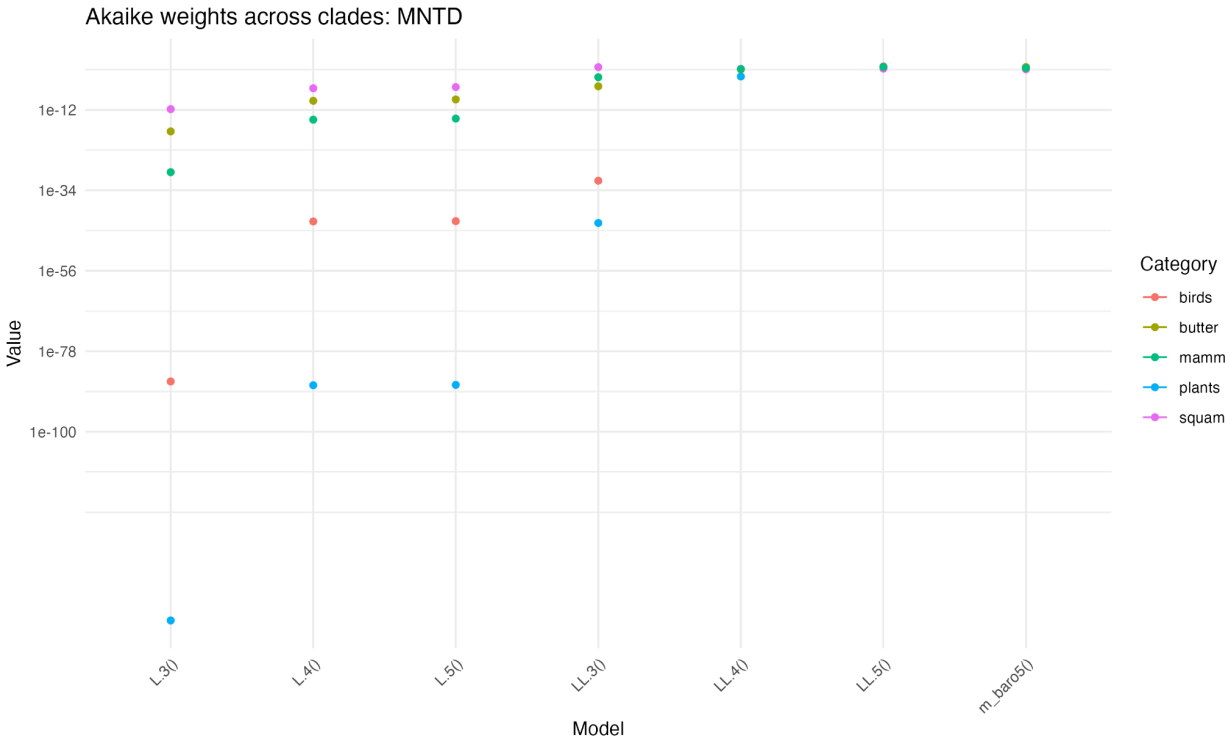


Figure 5: Contour of Evolutionary History (CEH) fits of California taxa across clades and metrics, fit with a 4-parameter log logistic growth function using the R drc package (a). Points show empirical bootstrapped values. Rows represent each clade: birds, plants, mammals, squamates and butterflies (top to bottom). Null models were fit to species pools considering all of California, but were also applied to all ecoregions. Plots of predicted vs. true values for all individual model fits (b) . Columns represent Clades in the order of birds, plants, mammals, squamates and butterflies (left to right) , and rows represent a model fit in the order of MNTD, MPD and PD 97.5 percentile fits followed by MNTD, MPD and PD 2.5 percentile fits (top to bottom).


Akaike weights across clades: MPD
Value
1e-08
1e-25
1e-42
1e-59
L.3()
L.4()
L.5()
LL.3()
LL.4()
LL.5()
m_baro5()
Model
Category
birds
butter
mamm
plants
squam


Akaike weights across clades: MNTD
Value
1e-12
1e-34
1e-56
1e-78
1e-100
L.3()
L.4()
L.5()
LL.3()
LL.4()
LL.5()
m_baro5()
Model
Category
birds
butter
mamm
plants
squam

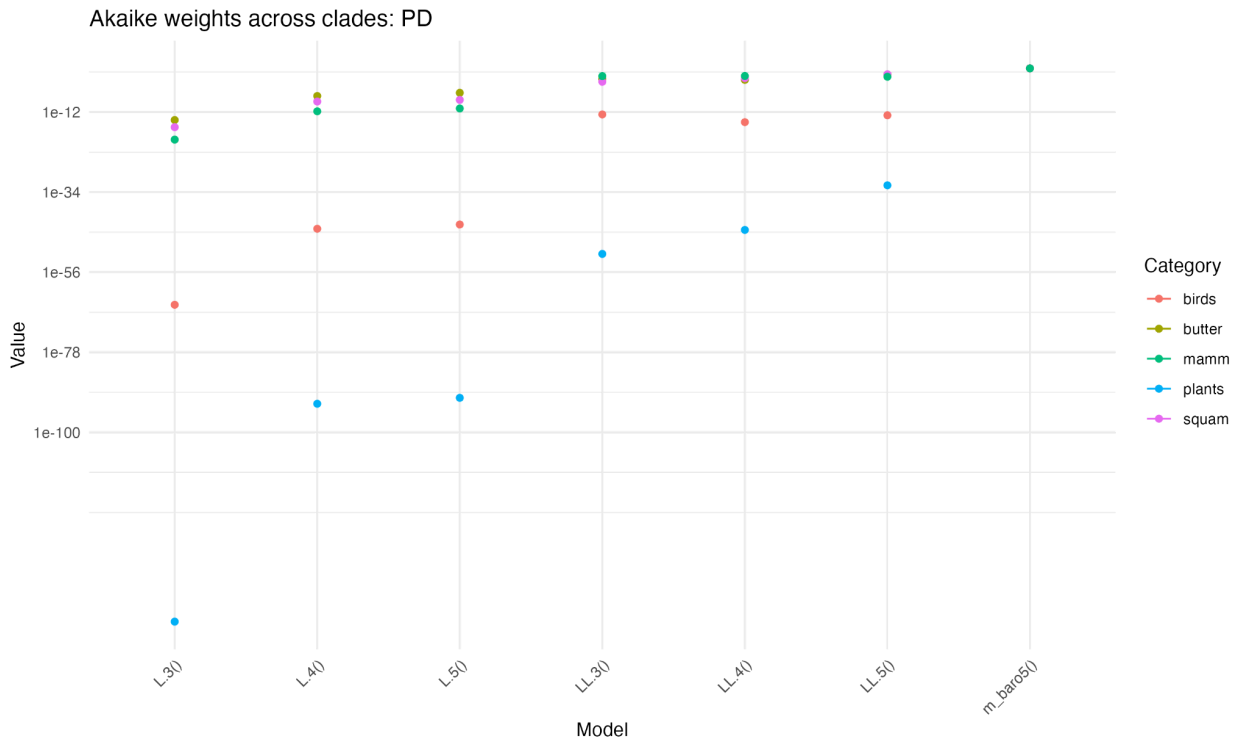


Figure 6: Akaike weights for all 7 tested models across all five clades for MPD (top), MNTD (middle) and PD (bottom). Model fits are shown for the California regional models.

	model	birds	plants	butter	squam	mamm
1	L.3()	1.097E-65	9.396E-153	6.226E-15	6.821E-17	2.644E-20
2	L.4()	8.293E-45	6.883E-93	2.566E-08	7.580E-10	1.582E-12
3	L.5()	1.241E-43	2.9121E-91	1.889E-07	2.135E-09	9.469E-12
4	LL.3()	2.189E-13	1.025E-51	0.001298	0.000210	0.007474
5	LL.4()	1.659E-15	3.989E-45	0.000666	0.002101	0.008031
6	LL.5()	1.274E-13	7.304E-33	0.020789	0.022997	0.004951
7	baro5()	0.9999999	1	0.977246	0.974691	0.979542

	model	birds	plants	butter	squam	mamm
1	L.3()	1.0314E-37	1.547E-72	3.568E-14	9.944E-10	1.324-22
2	L.4()	1.411E-27	5.594E-49	1.196E-10	3.9549E-06	2.524E-15
3	L.5()	3.585E-27	1.447E-48	3.201E-10	9.931E-06	6.199E-15
4	LL.3()	1.927E-09	1.888E-28	1.719E-07	0.0387859	3.415E-14
5	LL.4()	7.278E-05	0.0100256	0.0037538	0.1900041	0.0046378
6	LL.5()	5.819E-05	0.0188240	0.0015309	0.6941762	0.01523804
7	baro5()	0.9998690	0.97115031	0.99471503	0.0770197	0.98012415

	model	birds	plants	butter	squam	mamm
1	L.3()	5.592E-87	2.496E-152	1.291E-18	1.596E-12	9.336E-30
2	L.4()	3.115E-43	5.123E-88	3.110E-10	8.036E-07	2.095E-15
3	L.5()	3.7316E-43	6.555E-88	7.0674E-10	1.726E-06	4.1305E-15
4	LL.3()	4.245E-32	1.195E-43	2.845E-06	0.4851949	0.00085020
5	LL.4()	0.1199771	0.0013733	0.1269991	0.1644951	0.14361325
6	LL.5()	0.7548709	0.5287106	0.3591225	0.1873673	0.57476628
7	baro5()	0.1251519	0.4699159	0.5138754	0.1629400	0.28077025

Table 1: Raw Akaike weight values for all 7 tested models across all five clades for PD (top),MPD (middle) and MNTD (bottom). Model fits are shown for the California regional models.

Plants

	Model	MAPE_PD_Low	MAPE_PD_High	MAPE_MPD_Low	MAPE_MPD_High	MAPE_MNTD_Low	MAPE_MNTD_High	row_means
1	L.3()	10.531	7.516	0.634	1.934	7.3812	15.505	7.250
2	L.5()	3.088	2.470	0.617	0.988	1.810	7.020	2.666
3	L.4()	3.030	2.415	0.617	0.988	1.829	7.023	2.650
4	LL.3()	1.157	0.975	0.312	0.847	1.233	1.855	1.063
5	LL.4()	1.319	0.825	0.275	0.328	0.424	0.672	0.641
6	LL.5()	0.829	0.715	0.269	0.328	0.454	0.620	0.536
7	baro5()	0.417	0.194	0.265	0.303	0.426	0.514	0

Squamates

	Model	MAPE_PD_Low	MAPE_PD_High	MAPE_MPD_Low	MAPE_MPD_High	MAPE_MNTD_Low	MAPE_MNTD_High	row_means
1	L.3()	6.010	2.842	0.657	0.825	2.832	8.875	3.674
2	L.4()	1.900	1.088	0.619	0.301	1.193	3.342	1.407
3	L.5()	1.925	1.062	0.619	0.298	1.178	3.278	1.393
4	LL.3()	0.754	0.613	0.373	0.180	0.948	0.974	0.640
5	LL.4()	0.829	0.384	0.380	0.152	0.939	1.003	0.615
6	LL.5()	0.739	0.172	0.377	0.151	0.935	0.882	0.543
7	baro5()	0.692	0.316	0.278	0.147	0.937	0.826	0.533

Birds

	Model	MAPE_PD_Low	MAPE_PD_High	MAPE_MPD_Low	MAPE_MPD_High	MAPE_MNTD_Low	MAPE_MNTD_High	row_means
1	L.3()	7.586	4.735	1.068	0.925	11.108	25.850	8.546
2	L.5()	3.011	2.259	1.013	0.515	1.427	8.322	2.758
3	L.4()	2.924	2.239	1.013	0.515	1.463	8.355	2.752
4	LL.3()	1.060	0.564	0.423	0.280	2.701	2.794	1.304
5	LL.4()	0.999	0.644	0.411	0.248	0.711	0.605	0.603
6	LL.5()	0.950	0.558	0.405	0.244	0.596	0.601	0.559
7	baro5()	0.582	0.304	0.275	0.238	0.662	0.604	0.444

Butterflies

	Model	MAPE_PD_Low	MAPE_PD_High	MAPE_MPD_Low	MAPE_MPD_High	MAPE_MNTD_Low	MAPE_MNTD_High	row_means
1	L.3()	7.292	5.223	0.827	0.990	5.439	13.481	5.543
2	L.5()	2.531	1.866	0.767	0.381	1.632	5.150	2.05
3	L.4()	2.252	1.801	0.767	0.382	1.654	5.189	2.008
4	LL.3()	2.060	0.516	0.296	0.445	1.576	1.647	1.090
5	LL.4()	1.897	0.709	0.289	0.138	0.763	1.122	0.819
6	LL.5()	1.105	0.636	0.283	0.139	0.760	0.864	0.631
7	baro5()	0.875	0.382	0.136	0.119	0.759	0.831	0.517

Mammals

	Model	MAPE_PD_Low	MAPE_PD_High	MAPE_MPD_Low	MAPE_MPD_High	MAPE_MNTD_Low	MAPE_MNTD_High	row_means
1	L.3()	4.074	2.585	0.480	2.117	9.273	17.177	5.951
2	L.5()	2.012	1.174	0.459	0.702	2.183	6.589	2.187
3	L.4()	1.957	1.135	0.460	0.707	2.221	6.634	2.186
4	LL.3()	1.022	0.347	0.218	1.175	1.730	1.140	0.939
5	LL.4()	1.072	0.363	0.225	0.194	0.914	1.248	0.669
6	LL.5()	1.027	0.360	0.221	0.169	0.928	1.107	0.635
7	baro5()	0.671	0.341	0.157	0.152	0.926	1.051	0.550

Table 2: Mean Absolute Percent Error across all models for birds, plants, mammals, squamates and butterflies (top to bottom)

Supplementary IV: Correlation Tests

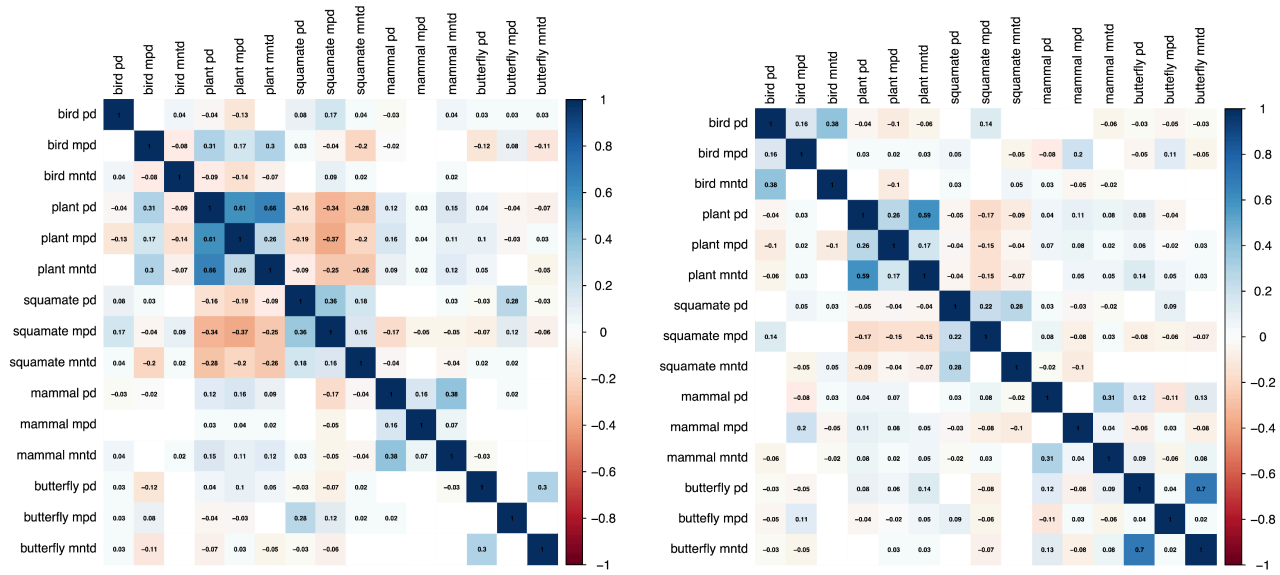


Figure 7: Symmetric Spearman's rank correlation matrices for metric/clade combinations ranked in comparison to all of California (left), and ecoregions (right). Darker red values indicate a negative correlation and darker blue values indicate a positive correlation. Ranked values exist in the set [-1, 0, 1], so 1 means that overdispersion or underdispersion always co-occur together, and values of -1 mean that overdispersion of one clade entirely co-occurs with underdispersion of the compared clade and vice versa.

Supplementary V: Reserve Analysis tables

Metric	Region Of Comparison	Clustered	Neutral	Overdispersed
PD	California	0.850	0.149	0.0002
MPD	California	0.550	0.361	0.088
MNTD	California	0.158	0.732	0.108
PD	Ecoregion	0.401	0.464	0.133
MPD	Ecoregion	0.171	0.679	0.148
MNTD	Ecoregion	0.259	0.654	0.086

Table 3: Average percent cover of clustered, neutral and overdispersed areas across reserve management systems in California

Supplementary Datasets

Supplementary Datasets I: Bird TreePL Time Calibration parameterization and output

https://ucla.box.com/s/q9941iilk1x41ojzri8nkn1f8zxl6g6j

Supplementary Datasets II: California-Trimmed phylogenies for birds, plants, mammals, squamates and butterflies

https://ucla.box.com/s/couxzo9436d8kitk5houa1l8ms315ebd

Supplementary Code

https://github.com/MayaGigChari/bird.git

REFERENCES

Aguilar-Tomasini, Maria A., Michael D. Martin, and James DM Speed. "Assessing spatial patterns of phylogenetic diversity of Mexican mammals for biodiversity conservation." *Global Ecology and Conservation* 31 (2021): e01834.

Baldwin, B. G., et al. (2017). Master spatial file for native California vascular plants used by Baldwin et al. (2017 Amer. J. Bot.) [Dataset]. Dryad. https://doi.org/10.6078/D16K5W.

Barnagaud, Jean‐Yves, et al. "Ecological traits influence the phylogenetic structure of bird species co‐occurrences worldwide." Ecology letters 17.7 (2014): 811-820.

Birch, Colin P.D., Oom, Sander P., and Beecham, Jonathan A. Rectangular and hexagonal grids used for observation, experiment, and simulation in ecology. Ecological Modelling, Vol. 206, No. 3–4. (August 2007), pp. 347–359.

Buerki, S., et al. (2015). Incorporating evolutionary history into conservation planning in biodiversity hotspots. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 370(1662), 20140014. https://doi.org/10.1098/rstb.2014.0014.

Buckley, Thomas R. "Applications of phylogenetics to solve practical problems in insect conservation." *Current opinion in insect science* 18 (2016): 35-39.

Burleigh, J. G., Kimball, R. T., & Braun, E. L. (2015). Building the avian tree of life using a large-scale, sparse supermatrix. Molecular Phylogenetics and Evolution, 84, 53-63.

Cadotte, M. W., Carscadden, K., & Mirotchnick, N. (2011). Beyond species: functional diversity and the maintenance of ecological processes and services. Journal of Applied Ecology, 48(5), 1079-1087.

Cadotte, M. W., et al. (2010). Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance, and evolutionary history. Ecology Letters, 13(1), 96-105.

Carvalho, S., Velo-Antón, G., Tarroso, P. *et al.* Spatial conservation prioritization of biodiversity spanning the evolutionary continuum. Nat Ecol Evol 1, 0151 (2017). https://doi.org/10.1038/s41559-017-0151

Cavender-Bares, J., et al. (2009). The merging of community ecology and phylogenetic biology. Ecology Letters, 12(7), 693-715.

Ceulemans, Ruben, et al. "The effects of functional diversity on biomass production, variability, and resilience of ecosystem functions in a tritrophic system." Scientific Reports 9.1 (2019): 7541.

Cooper, Natalie, Jesús Rodríguez, and Andy Purvis. "A common tendency for phylogenetic overdispersion in mammalian assemblages." Proceedings of the Royal Society B: Biological Sciences 275.1646 (2008): 2031-2037.

California Natural Resources Agency (CNRA). "Protecting Biodiversity." California Natural Resources Agency, https://resources.ca.gov/Initiatives/Protecting-Biodiversity. Accessed 3 June 2024.

data.ca.gov. CA Geographic Boundaries. data.ca.gov, 9 Aug. 2019, https://data.ca.gov/dataset/ca-geographic-boundaries. Accessed 8 June 2024.

Earl, C., et al. (2021). Spatial phylogenetics of butterflies in relation to environmental drivers and angiosperm diversity across North America. iScience, 24(4).

Edzer Pebesma, 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10:1, 439-446.

Edwards, David P., et al. "Land-sparing agriculture best protects avian phylogenetic diversity." Current biology 25.18 (2015): 2384-2391.

Emerson, Brent C., and Rosemary G. Gillespie. "Phylogenetic analysis of community assembly and structure over space and time." Trends in Ecology & Evolution, 23(11), 619-630.

Faith, Daniel P. "Conservation evaluation and phylogenetic diversity." Biological conservation 61.1 (1992): 1-10.

Ferrier, Simon, et al. "Mapping more of terrestrial biodiversity for global conservation assessment." BioScience 54.12 (2004): 1101-1109.

Goldberg, Emma E., Lesley T. Lancaster, and Richard H. Ree. "Phylogenetic inference of reciprocal effects between geographic range evolution and diversification." *Systematic biology* 60.4 (2011): 451-465.

González-Orozco, C.E., Mishler, B.D., Miller, J.T., Laffan, S.W., Knerr, N., Unmack, P., Georges, A., Thornhill, A.H., Rosauer, D.F. and Gruber, B. (2015), Assessing biodiversity and endemism using phylogenetic methods across multiple taxonomic groups. Ecol Evol, 5: 5177-5192. https://doi.org/10.1002/ece3.1747

Gotelli, Nicholas J., and Robert K. Colwell. "Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness." Ecology letter*s* 4.4 (2001): 379-391.

Hartmann, K., & André, J. (2013). Should evolutionary history guide conservation? Biodiversity and Conservation, 22, 449-458.

Hendershot, J. Nicholas, et al. "Intensive farming drives long-term shifts in avian community composition." *Nature* 579.7799 (2020): 393-396.

Howard, Christine, et al. "Improving species distribution models: the value of data on abundance." *Methods in Ecology and Evolution* 5.6 (2014): 506-513.

Isaac, Nick JB, et al. "Mammals on the EDGE: conservation priorities based on threat and phylogeny." PloS one 2.3 (2007): e296.

IUCN. (2023). The IUCN Red List of Threatened Species. Version 2023-1. https://www.iucnredlist.org. Accessed on [05/06/2024]

Jarvis, E. D., et al. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. Science, 346(6215), 1320-1331. https://doi.org/10.1126/science.1253451.

Jetz, W., & Pyron, R. A. (2018). The interplay of past diversification and evolutionary isolation with present imperilment across the amphibian tree of life. Nature Ecology & Evolution, 2.

Jetz, Walter, et al. "Global distribution and conservation of evolutionary distinctness in birds." *Current biology* 24.9 (2014): 919-930.

Kawahara, A. Y., et al. (2023). A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts, and biogeographic origins. Nature Ecology & Evolution, 7(6), 903-913.

Kembel, S., Cowan, P., Helmus, M., Cornwell, W., Morlon, H., Ackerly, D., Blomberg, S., & Webb, C. (2010). Picante: R tools for integrating phylogenies and ecology. Bioinformatics, 26, 1463-1464.

Kosmala, Margaret, et al. "Assessing data quality in citizen science." Frontiers in Ecology and the Environment 14.10 (2016): 551-560.

Kraft, Nathan JB, et al. "Trait evolution, community assembly, and the phylogenetic structure of ecological communities." The American Naturalist 170.2 (2007): 271-283.

Kraft, Nathan JB, Bruce G. Baldwin, and David D. Ackerly. "Range size, taxon age and hotspots of neoendemism in the California flora." Diversity and Distributions 16.3 (2010): 403-413.

Laffan, S. W., Lubarsky, E., & Rosauer, D. F. (2010). Biodiverse, a tool for the spatial analysis of biological and related diversity. Ecography, 33, 643-647. https://doi.org/10.1111/j.1600-0587.2010.06303.x.

Laity, Tania, et al. "Phylodiversity to inform conservation policy: An Australian example." Science of the Total Environment 534 (2015): 131-143.

Lean, Christopher, and James Maclaurin. "The value of phylogenetic diversity." Biodiversity Conservation and Phylogenetic Systematics: Preserving our evolutionary heritage in an extinction crisis (2016): 19-37.

Massante, J. C., et al. (2021). Phylogenetic structure of understorey annual and perennial plant species reveals opposing responses to aridity in a Mediterranean biodiversity hotspot. Science of The Total Environment, 761, 144018.

Mazel, F., et al. (2018). Prioritizing phylogenetic diversity captures functional diversity unreliably. Nature Communications, 9(1), 2888.

Miller, Eliot T., Damien R. Farine, and Christopher H. Trisos. "Phylogenetic community structure metrics and null models: a review with new methods and software." *Ecography* 40.4 (2017): 461-477.

Miller, J. T., et al. (2018). Phylogenetic Diversity Is a Better Measure of Biodiversity than Taxon Counting. Journal of Systematics and Evolution, 56(6), 663-667. https://doi.org/10.1111/jse.12436.

Mishler, B. D., et al. (2014). Phylogenetic measures of biodiversity and neo-and paleo-endemism in Australian Acacia. Nature Communications, 5(1), 4473. https://doi.org/10.1038/ncomms5473.

Mishler, Brent D. "Spatial phylogenetics." Journal of Biogeography 50.8 (2023): 1454-1463.

Mishler, Brent et al. (2020). Data from: Spatial phylogenetics of the North American flora [Dataset]. Dryad. https://doi.org/10.6078/D1709V

Mokany, Karel, et al. "Linking changes in community composition and function under climate change." Ecological applications 25.8 (2015): 2132-2141.

Montaño‐Centellas, Flavia A., Bette A. Loiselle, and Morgan W. Tingley. "Ecological drivers of avian community assembly along a tropical elevation gradient." Ecography 44.4 (2021): 574-588.

Moreno, Claudia E., et al. "Measuring biodiversity in the Anthropocene: a simple guide to helpful methods." Biodiversity and Conservation 26 (2017): 2993-2998.

Nitta, J. H., et al. (2022). Spatial phylogenetics of Japanese ferns: Patterns, processes, and implications for conservation. American Journal of Botany, 109(5), 727-745.

Oliveros, C. H., et al. (2019). Earth history and the passerine superradiation. Proceedings of the National Academy of Sciences, 116(16), 7916-7925.

Pollock, Laura J., et al. "Phylogenetic diversity meets conservation policy: small areas are key to preserving eucalypt lineages." Philosophical Transactions of the Royal Society B: Biological Sciences 370.1662 (2015): 20140007.

Province, Intermountain Semi-Desert. "Bioregions and the California Landscape." Fire in California's Ecosystems (2006): 2.

Revell, L. (2012). phytools: An R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution, 3, 217-223. https://doi.org/10.1111/j.2041-210X.2011.00169.x.

Ritz, C., Baty, F., Streibig, J. C., & Gerhard, D. (2015). Dose-response analysis using R. PLoS One, 10(12), e0146021.

Rosauer, Dan F., et al. "Phylogenetically informed spatial planning is required to conserve the mammalian tree of life." Proceedings of the Royal Society B: Biological Sciences 284.1865 (2017): 20170627.

Ricketts, Taylor, and Marc Imhoff. "Biodiversity, urban areas, and agriculture: locating priority ecoregions for conservation." *Conservation ecology* 8.2 (2003).

Santana, Victor M., et al. "Climate, and not fire, drives the phylogenetic clustering of species with hard-coated seeds in Mediterranean Basin communities." Perspectives in Plant Ecology, Evolution and Systematics 45 (2020): 125545.

Saraiva, Daniel Dutra, et al. "How effective are protected areas in conserving tree taxonomic and phylogenetic diversity in subtropical Brazilian Atlantic Forests?." *Journal for Nature Conservation* 42 (2018): 28-35.

Scherson, R. A., et al. (2017). Spatial phylogenetics of the vascular flora of Chile. Molecular Phylogenetics and Evolution, 112, 88-95.

Shirey, V., Belitz, M.W., Barve, V. and Guralnick, R. (2021), A complete inventory of North American butterfly occurrence data: narrowing data gaps, but increasing bias. Ecography, 44: 537-547. https://doi.org/10.1111/ecog.05396

Šmíd, Jiří, et al. "Diversity patterns and evolutionary history of Arabian squamates." Journal of Biogeography 48.5 (2021): 1183-1199.

Smith, Jeffrey R., et al. "A global test of ecoregions." Nature Ecology & Evolution 2.12 (2018): 1889-1896.

Spector, Sacha. "Biogeographic crossroads as priority areas for biodiversity conservation." *Conservation Biology* 16.6 (2002): 1480-1487.

Srivastava, Diane S., et al. "Phylogenetic diversity and the functioning of ecosystems." Ecology *letters* 15.7 (2012): 637-648.

Swenson, Nathan G. "The assembly of tropical tree communities–the advances and shortcomings of phylogenetic and functional trait analyses." *Ecography* 36.3 (2013): 264-276.

Thornhill, Andrew H., et al. "Continental‐scale spatial phylogenetics of Australian angiosperms provides insights into ecology, evolution and conservation." Journal of Biogeography43.11 (2016): 2085-2098.

Thornhill, Andrew H., et al. "Spatial phylogenetics of the native California flora." BMC biology 15 (2017): 1-18.

Title, P., et al. (2024). Data from: The macroevolutionary singularity of snakes [Dataset]. Dryad. https://doi.org/10.5061/dryad.p5hqbzkvb.

Toffelmier, E., Beninde, J., & Shaffer, H. B. (2022). The phylogeny of California, and how it informs setting multispecies conservation priorities. Journal of Heredity, 113(6), 597-603. https://doi.org/10.1093/jhered/esac045.

Tucker, C. M., et al. (2017). A guide to phylogenetic metrics for conservation, community ecology, and macroecology. Biological Reviews, 92(2), 698-715.

Upham, N. S., Esselstyn, J. A., & Jetz, W. (2019). Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation [Dataset]. Dryad. https://doi.org/10.5061/dryad.tb03d03.

Vamosi, Steven M., et al. "Emerging patterns in the comparative analysis of phylogenetic community structure." *Molecular ecology* 18.4 (2009): 572-592.

Vandergast, Amy G., et al. "Are hotspots of evolutionary potential adequately protected in southern California?." *Biological conservation* 141.6 (2008): 1648-1664.

VanDerWal, Jeremy, et al. "Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know?." Ecological modelling 220.4 (2009): 589-594.

Veron, Simon, et al. "The use of phylogenetic diversity in conservation biology and community ecology: A common base but different approaches." The Quarterly Review of Biology94.2 (2019): 123-148.

Webb, Campbell O., et al. "Phylogenies and community ecology." *Annual review of ecology and systematics* 33.1 (2002): 475-505.

Wilson, E. O. (1988). Biodiversity.

Winter, M., Devictor, V., & Schweiger, O. (2013). Phylogenetic diversity and nature conservation: where are we? Trends in Ecology & Evolution, 28(4), 199-204.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. Y. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution, 8, 28-36. https://doi.org/10.1111/2041-210X.12628.

Zhang, X. X., et al. (2022). Spatial phylogenetics of the Chinese angiosperm flora provides insights into endemism and conservation. Journal of Integrative Plant Biology, 64(1), 105-117.

Zupan, Laure, et al. "Spatial mismatch of phylogenetic diversity across three vertebrate groups and protected areas in Europe." Diversity and Distributions 20.6 (2014): 674-685.