Title

Landscape analyses using eDNA metabarcoding and Earth observation predict community biodiversity in California

Journal

Ecological Applications, 31(6)

ISSN

1051-0761

Authors

Lin, Meixi
Simons, Ariel Levi
Harrigan, Ryan J
et al.

Publication Date

2021-09-01

DOI

10.1002/eap.2379

Peer reviewed

HHS Public Access

Author manuscript

Ecol Appl. Author manuscript; available in PMC 2022 July 20.

Published in final edited form as:

Ecol Appl. 2021 September ; 31(6): e02379. doi:10.1002/eap.2379.

Landscape analyses using eDNA metabarcoding and Earth observation predict community biodiversity in California

Meixi Lin[1], **Ariel Levi Simons**[2,3], **Ryan J. Harrigan**[4], **Emily E. Curd**[1], **Fabian D. Schneider**[5], **Dannise V. Ruiz-Ramos**[6,7], **Zack Gold**[1], **Melisa G. Osborne**[8], **Sabrina Shirazi**[9], **Teia M. Schweizer**[1,10], **Tiara N. Moore**[1,11], **Emma A. Fox**[1], **Rachel Turba**[1], **Ana E. Garcia-Vedrenne**[1], **Sarah K. Helman**[1], **Kelsi Rutledge**[1], **Maura Palacios Mejia**[1], **Onny Marwayana**[1,12], **Miroslava N. Munguia Ramos**[1], **Regina Wetzer**[13,14], **N. Dean Pentcheff**[13], **Emily Jane McTavish**[7], **Michael N. Dawson**[7], **Beth Shapiro**[9,15], **Robert K. Wayne**[1], **Rachel S. Meyer**[1,9,16]

[1]Department of Ecology and Evolutionary Biology, University of California-Los Angeles, Los Angeles, California 90095 USA

[2]Department of Marine and Environmental Biology, University of Southern California, Los Angeles, California 90089 USA

[3]Institute of the Environment and Sustainability, University of California-Los Angeles, Los Angeles, California 90095 USA

[4]Center for Tropical Research, Institute of the Environment and Sustainability, University of California-Los Angeles, Los Angeles, California 90095 USA

[5]Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, California 91009 USA

[6]Columbia Environmental Research Center, U.S. Geological Survey, Columbia, Missouri 65201 USA

[7]Department of Life & Environmental Sciences, University of California-Merced, Merced, California 95343 USA

[8]Department of Molecular and Computational Biology, University of Southern California, Los Angeles, California 90089 USA

[9]Department of Ecology and Evolutionary Biology, University of California-Santa Cruz, Santa Cruz, California 95064 USA

[10]Department of Biology, Colorado State University, Fort Collins, Colorado 80523 USA

[11]School of Environmental and Forestry Sciences, University of Washington, Seattle, Washington 98195 USA

[16] rsmeyer@ucla.edu .

Supporting Information

Additional supporting information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/eap.2379/full

Open Research

Scripts and data (Lin 2021) associated with the analyses are archived in Zenodo: https://doi.org/10.5281/zenodo.4516670. The raw sequencing data is deposited in the NCBI Sequence Reads Archive under Bioproject PRJNA702201.

[12]Museum Zoologicum Bogoriense, Research Center for Biology, Indonesian Institute of Sciences (LIPI), Cibinong, Bogor 16911 Indonesia

[13]Research and Collections, Natural History Museum of Los Angeles County, Los Angeles, California 90007 USA

[14]Biological Sciences, University of Southern California, Los Angeles, California 90089 USA

[15]Howard Hughes Medical Institute, University of California-Santa Cruz, Santa Cruz, California 95064 USA

Abstract

Ecosystems globally are under threat from ongoing anthropogenic environmental change. Effective conservation management requires more thorough biodiversity surveys that can reveal system-level patterns and that can be applied rapidly across space and time. Using modern ecological models and community science, we integrate environmental DNA and Earth observations to produce a time snapshot of regional biodiversity patterns and provide multi-scalar community-level characterization. We collected 278 samples in spring 2017 from coastal, shrub, and lowland forest sites in California, a complex ecosystem and biodiversity hotspot. We recovered 16,118 taxonomic entries from eDNA analyses and compiled associated traditional observations and environmental data to assess how well they predicted alpha, beta, and zeta diversity. We found that local habitat classification was diagnostic of community composition and distinct communities and organisms in different kingdoms are predicted by different environmental variables. Nonetheless, gradient forest models of 915 families recovered by eDNA analysis and using BIOCLIM variables, Sentinel-2 satellite data, human impact, and topographical features as predictors, explained 35% of the variance in community turnover. Elevation, sand percentage, and photosynthetic activities (NDVI32) were the top predictors. In addition to this signal of environmental filtering, we found a positive relationship between environmentally predicted families and their numbers of biotic interactions, suggesting environmental change could have a disproportionate effect on community networks. Together, these analyses show that coupling eDNA with environmental predictors including remote sensing data has capacity to test proposed Essential Biodiversity Variables and create new landscape biodiversity baselines that span the tree of life.

Keywords

beta diversity; biomonitoring; citizen science; community ecology; ecological modeling; environmental DNA; gradient forest; remote sensing; zeta diversity

Introduction

Species are being rapidly lost worldwide (Pimm et al. 2014, Ceballos et al. 2015, Díaz et al. 2019) with many key habitats that harbor high biodiversity (Myers et al. 2000) threatened by climate change and environmental degradation. The scientific community needs rapid bioinventory tools to provide critical baseline biodiversity data with minimal cost and effort that can be applied globally (Bush et al. 2017). Essential Biodiversity

Variables (EBVs; Pereira et al. 2013) are a minimal set of measurements needed to support multi-purpose, long-term planning at various scales. Example EBVs include community composition, genetic composition, and ecosystem structure, which can be extrapolated from in situ and remote sensing observations. Scaling up from in situ biological measures to enable system-wide projections remains challenging (Pereira et al. 2013). Bioinventories remain often taxonomically or spatiotemporally restricted because technical feasibility limits large scale monitoring (Cristescu 2014), and thus, very few studies attempt to assess the complex composition of the total biotic environment (Karimi et al. 2018, George et al. 2019) that could provide unbiased EBVs needed to aid systems-level biodiversity conservation.

Technology-assisted citizen and community science (CCS) is a growing means to obtain in situ biodiversity observations to complement those made by taxonomic experts, and CCS observations from photographs and sounds have already eclipsed other biomonitoring data records such as physical collections (Theobald et al. 2015, Kobori et al. 2016). However, most CCS observations favor diurnal macroscopic species and often omit cryptic and microbial taxa (Theobald et al. 2015). In response, our program, CALeDNA (by the University of California Conservation Genomics Consortium; CALeDNA 2021), and several other fledging programs, have focused on giving community scientists the capacity to sample environmental DNA (eDNA) from their surroundings (Biggs et al. 2015, Miralles et al. 2016, Meyer et al. 2021), which can be probed for nearly any taxonomic group using multi-locus metabarcoding methods (Bohmann et al. 2014, Deiner et al. 2016, Thompson et al. 2017, Franklin et al. 2019).

Multi-locus metabarcoding of eDNA from surface soil and sediment retains a record of taxa recently present in the local area, including bacteria and archaea, often-overlooked meiofauna, protozoans, non-vascular plants, algae, and fungi in addition to the vertebrate and vascular plant communities that are easier to observe directly. These methods are increasing in accuracy as reference DNA sequence databases grow and informatic tools improve, and are decreasing in cost as library preparation and sequencing technology become less expensive. Community-powered eDNA surveys can be coupled with remote sensing measures of ecosystem properties to model community composition, generate EBVs and advance ecological theories about how community diversity is regulated by biotic and abiotic traits (Yamasaki et al. 2017). On the ground and space-based technologies yield increasingly copious and accessible abiotic data (Pettorelli et al. 2014, Schimel et al. 2019) on land cover, topography, soil property (Hengl et al. 2017), bioclimate (Fick and Hijmans 2017), human impact (WCS and CIESIN 2005), and vegetation (e.g., Sentinel-2; European Space Agency), which can be used to model eDNA biodiversity changes across landscapes (Crowther et al. 2019, van den Hoogen et al. 2019). Biotic-abiotic interactions among soil properties (e.g., pH and nutrient availabilities), climate, plant coverage, and habitat type have been shown to affect soil alpha and beta diversity in different taxonomic groups (Fierer and Jackson 2006, Ranjard et al. 2013, George et al. 2019, White et al. 2020) from tropical mountains to temperate ecosystems (Thompson et al. 2017, Karimi et al. 2018, Montagna et al. 2018, Peters et al. 2019). However, these studies have largely focused on a single habitat, region, or phylogenetic clade with few exceptions, notably, a national-scale soil eDNA survey in England showed that animal and microbial richness responded to different

environment factors but beta-diversity trends were shared across taxonomic groups (George et al. 2019).

Our study attempts to use multi-locus metabarcoding from CCS-collected eDNA in a biodiversity-ecological response model that spans kingdoms and habitats of California. Similar to other biodiversity hotspots, we expect discontinuous environmental clines and high endemism (Myers et al. 2000, Thompson et al. 2017) to be apparent in eDNA community patterns. Our objectives are threefold. First, we identify the taxonomic occurrence patterns recovered in eDNA surveys and assess their reliability and concordance with traditional observations. Second, we assess the relationship of eDNA alpha, beta, and zeta diversity to environmental measures to determine how the environment filters species richness and community composition. Third, we apply joint-species gradient forest and ecological co-occurrence network modeling to generate a community turnover map of the entire state of California and characterize the taxonomic families that are found to be most sensitive to environmental filtering. These analyses reveal the abiotic and biotic variables that are the most predictive of community composition patterns and provide a framework for using CCS-generated eDNA with remote sensing to refine static maps of ecological delineations and provide effective EBVs.

Methods

Sampling design

Volunteers for CALeDNA sampled biodiversity from a wide variety of habitats, including coast, shrub, and lowland forest sites across the state of California using target sampling and eDNA metabarcoding. Sample location metadata were collected by a smartphone webform made in Kobo Toolbox and included a photograph (software *available online*).[17] Surface samples were collected by filling three 2-mL tubes with substrate from <2 cm depth, each 30 cm apart. Samples were frozen at −80°C immediately upon their return to CALeDNA headquarters at UC Los Angeles.

To minimize the potential effect of seasonal variations in eDNA profiles, we selected samples from March 2017 to June 2017, with two-thirds of samples collected in April. We classified the predominant biome using photographs and a variety of geolocation data. We selected 100 samples from each of three transect types, coast, shrub/scrub (abbreviated as “shrub”), and forest, that covered the broadest latitudinal range possible. Samples with ambiguous metadata were removed, resulting in a total of 278 samples (98 coast, 89 shrub, and 91 forest) used in subsequent analyses (Table 1; Data S1).

Compilation of environmental variables

We assembled environmental variables across six main categories: location, habitat, bioclimate, soil properties, topography, and vegetation (including surface reflectance properties) variables (Appendix S1: Supplemental Methods, Figs. S1, S2; Data S1). Uncertainty layers were downloaded if available as well (Appendix S1: Fig. S3). All raster

[17] kobotoolbox.org

layers were aligned and projected to a unified 100 × 100 m grid from Google Earth Engine (Coordinate Reference System for this project: ESPG 4326, WGS84). Layers were stacked and clipped to California's extent, and used for point extraction. For coastal sites outside of the raster's geographical coverage, values were extracted by the closest point available in 0.5 km radius or assigned a value of "NA" if not available. All computation and analyses were performed in R version 3.5.3 (R Core Team 2019). Raster operations were performed using R package raster (v. 2.8–19; Hijmans 2019).

Considering that many environmental variables are correlated, we evaluated the Pearson's correlation coefficient of the 56 numerical environmental variables and hierarchically clustered the variables according to the coefficients into variable groups using R functions cor, hclust, and cutree. To reduce collinearity and improve interpretability in community modeling, we created a "reduced" set of 33 numerical environmental variables that had an $R^2 < 0.8$ (Table 1; Appendix S1: Figs. S1, S2) for downstream analysis.

DNA extraction, amplification, and sequencing

DNA extraction, amplification and sequencing followed Curd et al. (2019). Briefly, three 250-mg biological replicate soil samples from each site were fully homogenized and pooled per site. DNA was extracted using the QIAGEN DNeasy PowerSoil Kit (Qiagen, Valencia, California, USA) according to the manufacturer's instructions. Negative controls were included in every batch of 12–18 extractions. DNA was amplified by polymerase chain reaction (PCR), using primers for five barcode regions: 16S (515F and 806R; Caporaso et al. 2012), 18S (Euk_1391f and EukBr; Amaral-Zettler et al. 2009), CO1 (mlCOIintF and Folddegen-rev; Yu et al. 2012, Leray et al. 2013), fungal ITS1 ("FITS"; ITS5 and 5.8S; White et al. 1990, Epp et al. 2012), and plant ITS2 ("PITS"; ITS-S2F and ITS-S3R; Gu et al. 2013). Primer sequences and thermocycling profiles can be found in Appendix S1: Tables S1, S2. All PCR amplifications were performed in triplicate and with additional PCR negative controls. Triplicate positive amplifications confirmed by gel electrophoresis, were pooled by sample and barcode to equimolar levels, indexed and sequenced on an Illumina MiSeq (Illumina, San Diego, California, USA) using kit v3 for 2×300 bp reads (QB3-Berkeley FGL), and sequenced to a target depth of 50,000 reads/sample/metabarcode (Appendix S1: Supplemental Methods). Five of the 278 sites were processed as biological replicates by different technicians to inspect taxonomic variation in independent DNA extraction and technical processing.

Bioinformatics and data processing

We used default settings in the Anacapa Toolkit (Curd et al. 2019) for multi-locus sequence data processing and taxonomy assignment. In brief, quality control of raw sequences was performed using Cutadapt (Martin 2011) and FastX-Toolkit (Gordon et al. 2010), and inference of Amplicon Sequence Variants (ASVs) was made with DADA2 (Callahan et al. 2016). Taxonomy assignment was made on each ASV using Bowtie2 (Langmead and Salzberg 2012) and the Bayesian Lowest Common Ancestor algorithm (BLCA; Gao et al. 2017) on custom metabarcode-specific reference databases that were created using *Creating Reference libraries Using eXisting tools* (CRUX; Curd et al. 2019). Taxonomy assignments with a bootstrap confidence cutoff score over 0.6 were kept for each ASV. ASVs with

the exact same inferred LCA passing confidence filter were summed into one “taxonomic entry” as the species/phylotype/MOTU equivalent in this study (Appendix S1: Supplemental Methods).

To informatically control for contamination, we further removed all singleton or doubleton taxa, and removed taxa that occurred in more than two reads in all blank samples, from subsequent analyses. To prepare data for alpha and beta diversity analyses requiring rarefaction, we performed rarefaction in 10 replicates and took the mean using the custom_rarefaction function in the R package ranacapa (v. 0.1.0; Appendix S1: Text S1, Table S3; Kandlikar et al. 2018). Reads with no assignment were not removed before rarefaction. We also estimated concordance between biological replicates (Appendix S1: Text S2).

Comparison of eDNA taxonomic output with traditional surveys

To compare the eDNA taxonomic results to traditional surveys, we compared eDNA results to the curated species inventory of the University of California Natural Reserve System (UCNRS), which records Chordata, Arthropoda, and Streptophyta. We counted how many taxon records were shared or unique to eDNA results or traditional records at classification levels of order, family, and genus combining all reserves and within each reserve.

We developed a metric of traditional observation score (TOS) in eDNA taxonomic assignment. TOS uses all observation and collection records in the Global Biodiversity Information Facility (GBIF) database from a broad region centered on California to score whether the taxon assignment of an eDNA ASV has been observed. A TOS > 0 suggests there is support for the assignment of an ASV based on its presence in the TOS region (Appendix S1: Supplemental Methods).

Community alpha, beta, and zeta diversity relationships with environmental variables

We used the rarefied data set for alpha and beta diversity analyses to control for variations in read depth. Alpha diversity was calculated using Observed and Shannon’s Diversity Index in the R package vegan (v. 2.5–2; Oksanen et al. 2018). These two measures weigh relative sequence abundance differently. Shannon’s index penalizes rare sequences compared to the Observed index (Calderón-Sanou et al. 2020). We evaluated relationships of alpha diversity measures using the Kruskal-Wallis test for categorical environmental variables, and individual linear models and partial least squares models for numerical variables (Appendix S1: Supplemental Methods, Text S3).

Beta diversity was visualized by plotting sample relative abundance of the top 10 phyla for metabarcodes 16S, 18S, and CO1, and top 10 classes for PITS and FITS. Composition profiles were analyzed using unconstrained ordination to reveal turnover across sites. We calculated the binary Jaccard dissimilarity distance to only consider presence–absence patterns given eDNA relative abundance can be influenced by stochastic processes of DNA shedding, deposition, and decay. We performed principal coordinate analysis (PCoA; function ordinate), permutational multivariate ANOVA analysis (PERMANOVA; function adonis), and tested for the assumption of homogeneity of dispersion (function beta-disp) in the R packages phyloseq (v. 1.24.2; McMurdie and Holmes 2013) and vegan. We also

partitioned the data by the four categories in the majorhab variable (aquatic, herbaceous-, shrub- and tree-dominated habitats) and performed PCoA and PERMANOVA analyses within each major habitat. Additionally, we tested for the effects on community turnover of coastal sites and spatial correlation (Appendix S1: Text S4). Post hoc explanation of the ordination axes was performed by fitting the reduced set of numerical variables (Table 1) onto the PCoA result using functions envfit and ordisurf in the R package vegan (Appendix S1: Supplemental Methods).

Zeta diversity was used to measure the fraction of unique categories of organisms held in common among nearby sets of communities, which unlike beta diversity, considers the composition of metacommunities composed of more than two sites. We set cluster size to four nearby sites, calculated and scaled zeta four diversity (ζ_4) using the R package ZETADIV (v. 1.1.1; Latombe et al. 2018). We tested the likelihood of two model forms of the relationship between zeta diversity and sample numbers (zeta decline). Based on prior analyses (Hui et al. 2014), declines that follow a power-law of the form $\zeta_N = \zeta_1 N^{-b}$, or an exponential of the form $\zeta_N = \zeta_1 e^{b(N-1)}$, were associated with a niche differentiation or stochastic process of community assembly, respectively (Appendix S1: Supplemental Methods). Scaled ζ_4 diversity values were then plotted on a map of California using the R package Leaflet (v. 2.0.2; Cheng et al. 2018). Environmental factor groups were made by binning environmental variables according to their categories (Table 1). We used generalized linear models (GLM) to determine the variation in ζ_4 diversity attributed to either geographic distance or an environmental factor group.

Gradient forest modeling and ecological network analysis to predict and interpret community turnover across California

We used the gradient forest classification model in the R package gradientForest (v. 0.1–17; Ellis et al. 2012) to test which environmental variables best explained eDNA-detected community turnover patterns across California using all 272 sites without any missing metadata collected from three transects (six out of 278 sites excluded due to missing metadata). We chose to perform predictive modeling on beta diversity because it is less affected by molecular artefacts, such as PCR errors or tag-jumps, or variations in bioinformatics pipelines, and more likely to reflect ecologically meaningful community composition patterns compared to alpha diversity, which is more sensitive to eDNA processing strategies (Calderón-Sanou et al. 2020, Shirazi et al. 2020) and does not require the clustering of sites that zeta diversity does. Due to large variation in the coastal sites, we also performed additional gradient forest analyses excluding all coastal sites using the same methods. The gradient forest model was built with the reduced set of 33 numerical environmental variables (Table 1). We fit a classification-tree-based gradient forest model using default settings to the eDNA-derived biological matrix, but increased the number of trees to 2,000 per family to increase the stability of the model (Breiman 2001). To assess model robustness, we repeated the gradient forest model 20 times. To assess model power and reliability, we randomized the predictor matrix 100 times and ran the model with the same settings (Bay et al. 2018; Appendix S1: Supplemental Methods).

To visualize the community turnover inferred from the gradient forest model over space, we used the input of all 33 environmental variables from 100 × 100 m grids in the extent of California without extrapolation (Pitcher et al. 2011). We used the top three principal components from the transformed environmental variables and visualized them by red, green, and blue (RGB) bands (Ellis et al. 2012). To differentiate model performance from the high-dimensional nature of the environmental variable matrix and to provide prediction uncertainty estimates, we scaled the environmental variables and performed the same PCA and visualization procedure without using the model ("uninformed map") and performed a mantel test and a monotonic regression between the biological matrix and either the uninformed map or gradient-forest-informed map. We also estimated which area contained more uncertainty by mapping the sites in the gradient-forest-informed map to the biological matrix using a Procrustes rotation and evaluated the residuals (Ellis et al. 2012; Appendix S1: Supplemental Methods).

To explore the biotic interactions underlying the gradient forest patterns, results for each metabarcode were summarized by family, filtered on read depth and frequency, and used in ecological co-occurrence network analysis using the R package SpiecEasi (v. 0.1.4; Kurtz et al. 2015) for cross domain analysis that incorporates all five metabarcodes into one complex network (Tipton et al. 2018). Topological parameters were determined in Cytoscape (v. 3.6.1; Shannon et al. 2003) using the NetworkAnalyzer tool. To observe the relationship between network degrees and the prediction R^2 of each family from gradient forest, an ordinary least squares (OLS) linear regression model was made using the lm function in R and interactions were visualized with the R package Interactions (v. 1.1.1; Long 2020). To evaluate the co-occurrence and gradient forest predictor patterns in a phylogenetic framework, the 915 families used in the gradient forest modeling were mapped onto the Open Tree of Life and a synthetic tree was generated using synthesis release v12.3 (*available online*).[18] Phylogeny tips were annotated with data using the Interactive Tree of Life (*available online*).[19]

Results

eDNA metabarcoding recovered taxonomic entries across 86 phyla

The 278 selected samples from coast, shrub, and forest areas across California (Fig. 1A) were sequenced with five metabarcodes. Each metabarcode recovered their target groups as expected (Fig. 1C; Appendix S1: Table S1), with 16S amplifying Bacteria and Archaea, 18S and CO1 broadly amplifying eukaryotes including Animalia, Chromista, Fungi, Protozoa, and some Plantae, ITS1 amplifying Fungi (FITS) from Ascomycota, Basidiomycota, and other phyla, and the ITS2 region amplifying plants (PITS) across both Chlorophyta and Streptophyta.

Sequencing the 278 samples, five repeated biological replicate samples, and 23 negative controls as PCR blanks or extraction blanks amounted to 75,830,796 reads for the five metabarcoding loci and averaged 54,554 reads per sample per metabarcode. After several

[18] tree.opentreeoflife.org
[19] https://itol.embl.de/

steps of quality control, taxonomic assignment, and sequence decontamination, a total of 16,157,425 reads were assigned to 16,118 unique taxonomic entries, i.e., best taxonomic hypotheses (Data S2). The median assigned read depth was 7,717 (Appendix S1: Fig. S4) and mean taxa identified was 778 per sample. Assignments spanned 86 phyla with most reads and taxonomic entries being assigned to Proteobacteria, Ascomycota, and Basidiomycota (Fig. 1B, C). Despite fairly deep sequencing, stringent sample filtration and validation on eDNA result concordance were necessary to meet quality metrics practiced by the metabarcoding community (Goldberg et al. 2016, Taberlet et al. 2018; Appendix S1: Text S2, Fig. S5; Data S3). Sequence rarefaction for diversity analyses that require even read depth across samples was able to be set near the taxon accumulation curve asymptote, suggesting we did not undersample during sequencing, although we did have to remove a small number of sample sites to meet the depth requirement (Appendix S1: Text S1, Figs. S6, S7, Table S3).

Comparison with traditional surveys: eDNA results partially overlap with traditional observations

Our first objective to assess the concordance between eDNA surveys and traditional observations initially utilized the UC Natural Reserve System curated species list of Streptophyta, Arthropoda and Chordata made by traditional surveys. Forty-four Streptophyta families were only found in eDNA, 77 were only in traditional observations, 65 were recovered from both methods. We found that 110 Arthropoda families were only recovered from eDNA, 139 were only in traditional observations, and 16 were recovered from both methods. No Chordata families were jointly recovered from both methods, since our metabarcoding markers did not specifically target Chordata. Evaluating concordance at order, family, and genus levels, we determined that family was the classification level that could be best validated by traditional observation at our UCNRS sample sites (Data S4).

To further evaluate eDNA taxa and traditional observation concordance without relying on restricted local surveys, we assigned a Traditional Observation Score (TOS) for eDNA taxon entries using the GBIF records from Western North America and the Eastern Pacific, which represent hypotheses of correct matches if eDNA entries overlap with the region specific GBIF records. Only taxonomic entries resolved to at least the level of order were assigned a TOS, hence 1,700 eDNA entries were omitted. Results showed only 5.6% of eDNA entries had an adjusted TOS of 0 (no GBIF support for assignment), and 50.0% of entries had an adjusted TOS of 1 (strong GBIF support for assignment; Data S5). Partial concordance was found in the remaining entries. No relationship was found between TOS and the frequency at which a taxon was found in eDNA samples (Pearson's $R^2 = 0.004$; $P < 1 \times 10^{-5}$), suggesting the TOS is not heavily biased toward common or ubiquitous taxa. As with the UCNRS comparison, the TOS was highest at the family level, so we selected family level classification for downstream gradient forest and network analyses.

Beta and zeta diversity are structured by minor habitat and vegetation variables

We examined relationships of alpha, beta, and zeta diversity to environmental measures as our second objective. Alpha diversity varies at the local scale and across the terrestrial-marine interface (Appendix S1: Fig. S8), with high spatial stratification among loc (reported

location names) and minorhab (minor habitat) variables for all metabarcodes besides CO1 (Appendix S1: Fig. S9). Stratification for the clust variable (neighboring cluster of sites within a radius of 0.5 km) according to the Shannon Index for 16S and FITS (Data S6), indicated bacterial and fungal alpha diversity are locally constrained in California. Post-hoc Dunn tests of categorical groups (Appendix S1: Figs. S10–S13; Data S6), as well as individual linear regressions (Data S7) and partial least squares models (Data S8) of observed richness and Shannon diversity indices with numerical environmental observations showed alpha diversity is predicted by many environmental variables and is most strongly predicted in fungi (FITS; Appendix S1: Text S3, Fig. S14; Data S6–S8).

Similarly, beta diversity patterns exhibited variations by habitat characteristics and were structured by environmental filtering. We found visually apparent differences in dominant taxa by habitat grouping (Appendix S1: Fig. S15). In community dissimilarity analyses, beta diversity was significantly different across major habitat groups despite many overlapping sites in the ordination plots (PERMANOVA; Fig. 2A, B; Appendix S1: Figs. S16–S19; Data S9). In particular, samples from aquatic environments were more dispersed in the ordination (Fig. 2A, B). Beta dispersion also showed significant heterogeneity of multivariate dispersion (variance) within groups for all metabarcode and category combinations except loc, majorhab, transect, and clust for the PITS metabarcode (Data S9).

Further investigation into beta diversity patterns revealed that minor habitat (minorhab) composition within each of the four major habitats contributed strongly to dissimilarity in all markers (PERMANOVA, adjusted $P < 0.01$; Fig. 2C; Appendix S1: Fig. S20; Data S10). Jaccard dissimilarity PCoA revealed finerscale habitat partitions for some, but not all, minor habitat categories, suggesting eDNA may be useful to evaluate minor habitat classifications as distinct management units based on community types (McKnight et al. 2007). For example, within aquatic major habitat, many of the marine nearshore categories overlapped, while marine and freshwater lacustrine and riverine sites separated (Fig. 2C; Appendix S1: Fig. S20). Patterns of environmental filtering remained after exclusion of coastal sites and spatial correlation effects (Appendix S1: Text S4, Figs. S21, S22; Data S11, S12). For numerical variables, post hoc explanation of the ordination axes showed that photosynthetic activities (NDVI32 and greenness) were most highly correlated with 16S, 18S, and FITS (Table 2; Appendix S1: Fig. S23; Data S13). Soil organic carbon content (orcdrc) was most highly correlated with CO1, and Isothermality (bio3) was most highly correlated with PITS (Table 2).

Zeta diversity describes the degree of overlap in the number of unique categories of organisms held in common between N sites or communities (ζ_N; Appendix S1: Fig. S24A), which, as N increases, captures more variation due to turnover. This framework allows for an assessment in trends in regional scale turnover of relatively common organisms, which are less biased toward the presence of rare, or spuriously detected taxa (Hui et al. 2018). Environmental factor groups explained 1–32% of the observed variation in ζ diversity (Table 3). Vegetation variables were among the top predictors for 18S, CO1, FITS, and PITS data sets, with the highest variance explained at 32% for the FITS data set. Variables related to small-scale location describe minimal variation (<1%) in ζ_4 diversity for communities (Table 3). To better understand the likeliest processes associated with the spatial assembly of

communities, two models of zeta diversity decline were tested using the power law model and the exponential model. The power law model was found to be a better fit for more than 83% communities described in all but the PITS metabarcode results, 31% of which followed the exponential model, suggesting lower spatial autocorrelation in plant and algal communities (Appendix S1: Fig. S24; Data S14).

Gradient forest models map high-resolution biodiversity turnover in California

Our third objective used gradient forest and ecological co-occurrence network modeling to map and characterize the taxonomic families that are predicted by the environment. Our gradient forest model included 272 sites × 915 eDNA-derived families as a response variable matrix and 272 sites × 33 environmental variables as a predictor matrix (Data S15). The gradient forest model explained 35% of variation in the biotic matrix, and all 915 families were able to be effectively modeled (i.e., had an $R^2 > 0$) with high stability across 20 replicated runs (Average $R^2 = 0.349 \pm 0.0004$; Average families effectively modeled = 915 ± 0; Data S16). Using a permutation approach, we confirmed the mean overall R^2 and number of families with positive R^2 for true observations were significantly higher than all the permuted runs (Appendix S1: Fig. S25). Many of the most responsive families were from marine aquatic sites, and some of these were low in observation frequency (Fig. 3B; Appendix S1: Fig. S26).

Gradient forest provides information on the rate of community turnover along environmental gradients (Ellis et al. 2012). We plotted the relative density of splits and cumulative importance for environmental variables. Within the top three environmental variables, we found nonlinear community changes. For elevation, rapid community turnover (high splits density) occurred at 0 m and above 1,000 m (Fig. 3C, D). For sand percentage, important splits were mainly distributed at 23%, 43%, and 74% sand (local maxima with the highest density; Fig. 3C, D), which have similarity to the soil texture triangle in the USDA system (Groenendyk et al. 2015). For photosynthetic activities (NDVI32), important splits were mainly distributed along −0.16, 0.05, and 0.28 (scale: −1 to 1; Fig. 3C, D).

Our map of California biodiversity resembled EPA North America Level II and California Level III Ecoregion maps (U.S. Environmental Protection Agency 2010, 2012), which were created with different input data and methods (Fig. 4C–E). For example, in the gradient forest map (Fig. 4A), the majority of central and southwestern California community type (red) corresponded to Mediterranean California (Fig. 4C, pale green, Level II 11.1.), characterized by medium photosynthetic activities (NDVI32), lower elevation (elev), higher precipitation seasonality (bio15) and higher mean temperature of wettest quarter (bio8).

We assessed the model prediction robustness and prediction uncertainties by regenerating our community turnover map of California without using any information obtained from eDNA surveys (Fig. 4B), and the resulting map neither resembled California published maps such as the EPA North America Level II Ecoregion map (U.S. Environmental Protection Agency 2010, Omernik and Griffith 2014; Fig. 4C) nor did it separate regions as sharply as the eDNA-informed map (Fig. 4A). This purely physical approach of community turnover mapping showed adding eDNA improves gradient forest informed mapping by a 1.4% reduction in stress performance statistics and a 5.6% increase in Mantel correlation

R^2 (Appendix S1: Fig. S27). We quantified the prediction uncertainties at each site by Procrustes rotation errors and found that predictions for coastal sites harbor more deviation from real eDNA communities (Dunn test, $P < 0.001$; Appendix S1: Fig. S28). We also were curious how robust our map was when coastal sites were removed, since several of the most predicted families were marine, and found that we could still explain 30% of the variation in the biotic matrix (Appendix S1: Text S5, Fig. S29).

Biotic co-occurrence has a weak positive relationship with gradient forest predictability

To characterize the biotic relationships of families across the spectrum of their predictability in the gradient forest models, which indicates environmental filtering (Horner-Devine et al. 2007), we modeled the relationship between each family's ecological co-occurrence network degrees and their predictor R^2 using an OLS linear model. Co-occurrence patterns reflect biotic niche processes that maintain biodiversity patterns that theoretically hold no expected relationship with abiotic environmental filtering. A family-level co-occurrence network produced 916 edges connecting 290 nodes (families) out of the total 304 families that met minimum frequency thresholds for analysis (Fig. 5A; Data S17). In the OLS linear model, interaction effects of site frequency were also considered. Model results showed a modest positive relationship (adjusted $R^2 = 0.22$) between the number of edges and gradient forest R^2 for families, indicating the families determined by gradient forest to be under the most environmental filtering were also the families most integrated in ecological networks based on their numbers of degrees. However, the interaction between frequency in sites and network degrees was also significant ($P < 0.02$; Fig. 5B). In a phylogenetic analysis of these patterns, we observed that families with high network degrees and high gradient forest predictor values were widely distributed across clades and kingdoms, but most frequent in the clades containing the class Flavobacteriia and the SAR supergroup (Stramenopiles, Alveolates, and Rhizaria; Fig. 5C), suggesting ecological networks containing these families might have the lowest resilience under abiotic change.

Discussion

Species observations by the public will continue to outpace both field collections and on-the-ground observations made by scientists (Theobald et al. 2015). With eDNA as a CCS tool (Biggs et al. 2015, Miralles et al. 2016, Larson et al. 2020), broader taxonomic inventories and assessments from minimally invasive environmental collections can be accomplished. Soils and sediments used in this study, collected by CCS volunteers, had an average of 778 taxonomic lineages identified in each DNA sample, and were easily obtained from a broad area within a seasonal snapshot. Co-analysis of eDNA from these collections and readily available environmental data provides predictor values for hundreds of families that evade traditional observations.

Our first objective concerning the concordance between eDNA results and traditional observations revealed relatively low overlap with UCNRS surveys, despite high support by GBIF traditional observation score, which suggests eDNA CCS surveys complement but do not replace traditional surveys. Ongoing efforts to sequence species and build a global taxonomic biodiversity reference database in the next decade (e.g., the Earth BioGenome

Project [Lewin et al. 2018], the Centre for Biodiversity Genomics [Hobern 2021]) are positioned to ameliorate shortcomings of current DNA reference sequences. Emerging alternatives to metabarcoding may additionally help mitigate detection bias currently in favor of small body size in eDNA studies (Fig. 1; Data S4, S5). For example, DNA capture approaches to target larger organisms (Seeber et al. 2019) may improve detection of large-bodied species, but these are not yet as cost-effective for CCS as multi-locus metabarcoding is. Another challenge is that different DNA extractions from the same soil or sediment sample exhibit heterogeneity (Appendix S1: Text S2; Data S3). We are examining stability and stochasticity of taxonomic profiles under varied sample processing (Castro et al. 2021) and DNA library preparation steps (Shirazi et al. 2020) in response to calls for research about these potential biases (Prosser 2010, Goldberg et al. 2016). In this study we used several standard approaches for reducing these biases.

Our second aim to test predictors of alpha, beta, and zeta diversity revealed that most environmental categories can significantly partition samples according to taxonomic composition (Fig. 2; Appendix S1: Figs. S15–S20; Data S6–S13), suggesting that surface communities are largely filtered by ecological rather than neutral processes (Bahram et al. 2018). These patterns remained significant after exclusion of coastal sites and location effects (Appendix S1: Figs. S21, S22; Data S11, S12). However, we found substantial overlap in community composition ordinations, as has been shown in the global Earth Microbiome Project (Thompson et al. 2017) and regional soil biodiversity ordination plots (George et al. 2019; Fig. 2A, B; Data S9). In our ordinations, groups separated from each other when fine-scale categories are used, such as minor habitat within partitioned major habitat, suggesting a large amount of community partitioning is harbored within major habitats categories (Appendix S1: Fig. S20). We found prokaryotic diversity was particularly diagnostic of minor habitats in ordinations (Fig. 2C; Appendix S1: Fig. S20). We propose eDNA-based composition could be EBVs for planning management units such as minor habitat delineations and for detecting ecotones (Jetz et al. 2019).

Environmental variables (Tables 2, 3) can have power to predict general biotic patterns and can illuminate possible drivers of community turnover (Appendix S1: Fig. S23) because they can readily be compared across studies (Omernik and Griffith 2014). For example, photosynthetic activities (NDVI32/greenness) had the highest correlation with the observed fungal alpha diversity pattern and beta diversity structure in bacteria (16S), eukaryotes (18S) and fungi (FITS) in the envfit analyses (Table 2; Appendix S1: Fig. S23). We note indices of photosynthetic activity have not been included as part of most microbiome studies (Bahram et al. 2018, Karimi et al. 2018, George et al. 2019) so their importance is still being discovered. For the subset of studies we found that had included NDVI as a predictor, it was observed to be important in modulating soil fungal and herbivore nematodes communities (Timling et al. 2014, Delgado-Baquerizo et al. 2016, Yang et al. 2017, van den Hoogen et al. 2019). Isothermality (bio3) has strong positive associations with PITS beta diversity turnover, suggesting inland arid California regions with low isothermality display nestedness in the biodiversity encompassed by these markers, as has been shown with plants in Australia (Gibson et al. 2012) and in South American seasonally dry forests (Silva and Souza 2018). Organic carbon (orcdrc) was strongly associated with CO1 community turnover, which mirrors associations reported in soil meiofaunal communities, particularly

nematodes (Jackson et al. 2019). Overall, zeta diversity largely supports the envfit results, although zeta diversity had poorer explanatory power for 16S patterns, which can be attributed to its greater sensitivity to common groups (Table 3; Simons et al. 2019) such as the nearly ubiquitous taxa in Proteobacteria.

Previous efforts have successfully integrated abiotic environmental data and models with traditional observational records such as herbarium specimens (Baldwin et al. 2017) to produce maps used to conserve threatened species (Jenkins et al. 2015), assess deforestation (Zarnetske et al. 2019) and evaluate species richness and endemism (Baldwin et al. 2017). However, remotely sensed variables such as from the Sentinel-2 instrument and local-scale eDNA observations of taxonomy biodiversity enable community mapping at a grid size finer than 5 km (Jenkins et al. 2013, 2015, Pimm et al. 2014, Baldwin et al. 2017, Zarnetske et al. 2019), which aligns better with in situ biodiversity (Wang et al. 2018). Our objective to project community composition across California's landscape achieved a higher resolution than currently available statewide maps (Fig. 4). Elevation (elev), sand percentage (sndppt), photosynthetic activities (NDVI32) and the mean temperature in the wettest quarter (bio8) were the among the most important predictors (Fig. 3A) and all of these variables had been proposed to be prominent drivers in community structures worldwide. For example, sand percentage, an inverse of clay percentage, is known to explain differences in plant community guilds (Cornelius et al. 1991), correlates with presence of halophytes (Lee et al. 2016, Moreno et al. 2018) and influences microbial community structures (Sessitsch et al. 2001, Ehrlich et al. 2015).

Space, flight, tower, and drone-based remote-sensing information are becoming increasingly available and accessible (Pettorelli et al. 2014). By providing more direct, spatially continuous measures of plant functional diversity and ecosystem functioning at regional (Schneider et al. 2017, Durán et al. 2019, Sousa et al. 2021) to global scales (Schimel et al. 2019, Schneider et al. 2020), we expect that future analyses will uncover new rules (Rocchini et al. 2021) and important environmental predictors, and will develop prediction maps on species richness (alpha diversity) or community turnover at higher dimensions (zeta diversity), expanding on the beta diversity map presented here. The eDNA composition could potentially be better predicted with more remote sensing and in situ bioinventory data from different spatial and temporal scales with improved gradient forest R^2 from what we achieved at $R^2 = 0.35$ and decreased prediction uncertainties. Bayesian hierarchical modeling and artificial neural networks are also receiving increasing attention for community modeling with more application potentials for improved spatial-temporal biodiversity predictions with associated uncertainty estimates (Hefley and Hooten 2016, Nieto-Lugilde et al. 2018, Pollock et al. 2020). We are looking forward to applying Bayesian hierarchical models in future CALeDNA meta-analyses.

Finally, we suggest eDNA ecological network analyses should be leveraged so that the biotic interaction dependence can be contrasted with dependence or sensitivity to the abiotic environment. Our work shows a weak but positive relationship between the number of degrees a family has and its propensity for environmental filtering based on gradient forest predictability. This positive relationship persists across phylogenetic groups (Fig. 5). Other

studies focused on a single kingdom have obtained similar conclusions, such as in microbial variation in an altitudinal gradient in the Atacama Desert, Chile (Mandakovic et al. 2018).

Conclusion

In conclusion, we demonstrate the emerging potential of coupling CCS observations and eDNA data from samples that CCS volunteers collect in combination with remote sensing and ecological modeling to assess community–environment interactions and ultimately map community turnover. We provide one of the most comprehensive surveys of terrestrial biodiversity across three domains of life over a large, environmentally diverse state. We show the predictive and explanatory power of environmental variables on alpha, beta, and zeta diversity across highly diverse regions and at local geographic scales. The beta diversity map for California, as a continuous surface of community turnover, shares many similar boundaries to the standard U.S. Ecoregion maps, but with nuanced detail. Computationally intensive and artificial intelligence driven models are producing maps for mitigating the challenges of global change (Harfouche et al. 2019, Pollock et al. 2020). Our approach contributes to the development of strategies to model living systems which could be directly used as Essential Biodiversity Variables for tracking biodiversity change, advancing ecological understanding, and managing ecosystems.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding for the CALeDNA sample processing, infrastructure, and personnel was provided by the University of California Research Initiatives (UCRI) Catalyst grant CA-16-376437 and Howard Hughes Medical Institute (HHMI) Professors Grant GT10483. Additional funding for personnel and computational infrastructure was provided by the National Science Foundation (NSF) 1759756. The research carried out at the Jet Propulsion Laboratory, California Institute of Technology, was under a contract with the National Aeronautics and Space Administration (80NM0018D0004). Government sponsorship is acknowledged. A. E. Garcia-Vedrenne is a postdoctoral fellow supported by UPLIFT: UCLA Postdocs' Longitudinal Investment in Faculty (Award # K12 GM106996). This is contribution number 7 of the Natural History Museum of Los Angeles County's Diversity Initiative for the Southern California Ocean (DISCO). Graduate student support was additionally provided by the National Council for Scientific and Technological Development of Brazil [Grant No. 209261/2014-5] and the University of California, Los Angeles Department of Ecology and Evolutionary Biology (EEB) Summer Research Fellowship. We thank the UC Natural Reserves System managers, other natural areas managers, and the hundreds of volunteers for collections. We thank A. Mahinan, N. Stavros, and W.-Y Kwan for assisting with spatial environmental data and A. DeVries and L. Bulbenko for preparing extractions. We thank T. Gillespie, C. M. Mueller, and Z. Kurtz for help optimizing analyses. R. S. Meyer, R. K. Wayne, B. A. Shapiro, and E. E. Curd designed the study. M. Lin, R. S. Meyer, R. J. Harrigan, A. L. Simons, M. Osborne, E. E. Curd, and Z. Gold selected analyses. R. S. Meyer and E. E. Curd coordinated public sampling. M. Lin, E. Fox, T. M. Schweizer, and R. S. Meyer made DNA libraries. M. Lin, M. P. Mejia, F. D. Schneider, A. E. Garcia-Vedrenne, D. Ruiz-Ramos, R. S. Meyer, and E. E. Curd curated environmental metadata. M. Lin, F. D. Schneider, and R. J. Harrigan generated statewide data layers. M. Lin led biodiversity and gradient forest analyses and M. Lin, A. L. Simons, and R. S. Meyer generated plots. E. J. McTavish generated the synthetic phylogeny. All authors performed analyses and interpretation. M. Lin, R. S. Meyer, and R. K. Wayne wrote the manuscript with input from all authors.

Literature Cited

Amaral-Zettler LA, McCliment EA, Ducklow HW, and Huse SM. 2009. A method for studying Protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. PLoS ONE 4:e6372. [PubMed: 19633714]

Bahram M, et al. 2018. Structure and function of the global topsoil microbiome. Nature 560:233–237. [PubMed: 30069051]

Baldwin BG, Thornhill AH, Freyman WA, Ackerly DD, Kling MM, Morueta-Holme N, and Mishler BD. 2017. Species richness and endemism in the native flora of California. American Journal of Botany 104:487–501. [PubMed: 28341628]

Bay RA, Harrigan RJ, Underwood VL, Gibbs HL, Smith TB, and Ruegg K. 2018. Genomic signals of selection predict climate-driven population declines in a migratory bird. Science 359:83–86. [PubMed: 29302012]

Biggs J, et al. 2015. Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (*Triturus cristatus*). Biological Conservation 183:19–28.

Bohmann K, Evans A, Gilbert MTP, Carvalho GR, Creer S, Knapp M, Yu DW, and de Bruyn M. 2014. Environmental DNA for wildlife biology and biodiversity monitoring. Trends in Ecology & Evolution 29:358–367. [PubMed: 24821515]

Breiman L 2001. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Statistical Science 16:199–231.

Bush A, et al. 2017. Connecting Earth observation to high-throughput biodiversity data. Nature Ecology & Evolution 1:0176.

Calderón-Sanou I, Münkemüller T, Boyer F, Zinger L, and Thuiller W. 2020. From environmental DNA sequences to ecological conclusions: How strong is the influence of methodological choices? Journal of Biogeography 47:193–206.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, and Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nature Methods 13:581–583. [PubMed: 27214047]

Caporaso JG, et al. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME Journal 6:1621–1624. [PubMed: 22402401]

Castro LR, Meyer RS, Shapiro B, Shirazi S, Cutler S, Lagos AM, and Quiroga SY. 2021. Metabarcoding meiofauna biodiversity assessment in four beaches of Northern Colombia: effects of sampling protocols and primer choice. Hydrobiologia 848. 10.1007/s10750-021-04576-z

Ceballos G, Ehrlich PR, Barnosky AD, García A, Pringle RM, and Palmer TM. 2015. Accelerated modern human–induced species losses: entering the sixth mass extinction. Science Advances 1:e1400253. [PubMed: 26601195]

Cheng J, Karambelkar B, and Xie Y. 2018. leaflet: create interactive web maps with the JavaScript "Leaflet" library. https://CRAN.R-project.org/package=leaflet

Cornelius JM, Kemp PR, Ludwig JA, and Cunningham GL. 1991. The distribution of vascular plant species and guilds in space and time along a desert gradient. Journal of Vegetation Science 2:59–72.

Cristescu ME 2014. From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. Trends in Ecology & Evolution 29:566–571. [PubMed: 25175416]

Crowther TW, van den Hoogen J, Wan J, Mayes MA, Keiser AD, Mo L, Averill C, and Maynard DS. 2019. The global soil community and its influence on biogeochemistry. Science 365:eaav0550. [PubMed: 31439761]

Curd EE, et al. 2019. Anacapa toolkit: an environmental DNA toolkit for processing multilocus metabarcode datasets. Methods in Ecology and Evolution 10:1469–1475.

Deiner K, Fronhofer EA, Müachler E, Walser J-C, and Altermatt F. 2016. Environmental DNA reveals that rivers are conveyer belts of biodiversity information. Nature Communications 7:12544.

Delgado-Baquerizo M, Maestre FT, Reich PB, Jeffries TC, Gaitan JJ, Encinar D, Berdugo M, Campbell CD, and Singh BK. 2016. Microbial diversity drives multifunctionality in terrestrial ecosystems. Nature Communications 7:10541.

Díaz S, et al. 2019. Pervasive human-driven decline of life on Earth points to the need for transformative change. Science 366:eaax3100. [PubMed: 31831642]

Durán SM, et al. 2019. Informing trait-based ecology by assessing remotely sensed functional diversity across a broad tropical temperature gradient. Science Advances 5:eaaw8114. [PubMed: 31840057]

Ehrlich R, Schulz S, Schloter M, and Steinberger Y. 2015. Effect of slope orientation on microbial community composition in different particle size fractions from soils obtained from desert ecosystems. Biology and Fertility of Soils 51:507–510.

Ellis N, Smith SJ, and Pitcher CR. 2012. Gradient forests: calculating importance gradients on physical predictors. Ecology 93:156–168. [PubMed: 22486096]

Epp LS, et al. 2012. New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. Molecular Ecology 21:1821–1833. [PubMed: 22486821]

Fick SE, and Hijmans RJ. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37:4302–4315.

Fierer N, and Jackson RB. 2006. The diversity and biogeography of soil bacterial communities. Proceedings of the National Academy of Sciences USA 103:626–631.

Franklin TW, et al. 2019. Using environmental DNA methods to improve winter surveys for rare carnivores: DNA from snow and improved noninvasive techniques. Biological Conservation 229:50–58.

Gao X, Lin H, Revanna K, and Dong Q. 2017. A Bayesian taxonomic classification method for 16S rRNA gene sequences with improved species-level accuracy. BMC Bioinformatics 18:247. [PubMed: 28486927]

George PBL, et al. 2019. Divergent national-scale trends of microbial and animal biodiversity revealed across diverse temperate soil ecosystems. Nature Communications 10:1107.

Gibson N, Meissner R, Markey AS, and Thompson WA. 2012. Patterns of plant diversity in ironstone ranges in arid south western Australia. Journal of Arid Environments 77:25–31.

Goldberg CS, et al. 2016. Critical considerations for the application of environmental DNA methods to detect aquatic species. Methods in Ecology and Evolution 7:1299–1307.

Gordon A, et al. 2010. Fastx-toolkit. FASTQ/A short-reads preprocessing tools. http://hannonlab.cshl.edu/fastx_toolkit.

Groenendyk DG, Ferré TPA, Thorp KR, and Rice AK. 2015. Hydrologic-process-based soil texture classifications for improved visualization of landscape function. PLoS ONE 10: e0131299. [PubMed: 26121466]

Gu W, Song J, Cao Y, Sun Q, Yao H, Wu Q, Chao J, Zhou J, Xue W, and Duan J. 2013. Application of the ITS2 region for barcoding medicinal plants of Selaginellaceae in Pteridophyta. PLoS ONE 8:e67818. [PubMed: 23826345]

Harfouche AL, Jacobson DA, Kainer D, Romero JC, Harfouche AH, Scarascia Mugnozza G, Moshelion M, Tuskan GA, Keurentjes JJB, and Altman A. 2019. Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. Trends in Biotechnology 37:1217–1235. [PubMed: 31235329]

Hefley TJ, and Hooten MB. 2016. Hierarchical species distribution models. Current Landscape Ecology Reports 1:87–97.

Hengl T, et al. 2017. SoilGrids250m: global gridded soil information based on machine learning. PLoS ONE 12:e0169748. [PubMed: 28207752]

Hijmans RJ et al. 2019. raster: geographic data analysis and modeling. https://CRAN.R-project.org/package=raster

Hobern DG 2021. BIOSCAN: DNA barcoding to accelerate taxonomy and biogeography for conservation and sustainability. Genome 64:161–164. [PubMed: 32268069]

Horner-Devine MC, et al. 2007. A comparison of taxon co-occurrence patterns for macro- and microorganisms. Ecology 88:1345–1353. [PubMed: 17601127]

Hui C, McGeoch MA, Harrison AES, and Bronstein EJL. 2014. Zeta diversity as a concept and metric that unifies incidence-based biodiversity patterns. American Naturalist 184:684–694.

Hui C, Vermeulen W, and Durrheim G. 2018. Quantifying multiple-site compositional turnover in an Afrotemperate forest, using zeta diversity. Forest Ecosystems 5:15.

Jackson LE, Bowles TM, Ferris H, Margenot AJ, Hollander A, Garcia-Palacios P, Daufresne T, and Sánchez-Moreno S. 2019. Plant and soil microfaunal biodiversity across the borders between arable and forest ecosystems in a Mediterranean landscape. Applied Soil Ecology 136:122–138.

Jenkins CN, Pimm SL, and Joppa LN. 2013. Global patterns of terrestrial vertebrate diversity and conservation. Proceedings of the National Academy of Sciences USA 110: E2602–E2610.

Jenkins CN, Van Houtan KS, Pimm SL, and Sexton JO. 2015. US protected lands mismatch biodiversity priorities. Proceedings of the National Academy of Sciences USA 112:5081–5086.

Jetz W, et al. 2019. Essential biodiversity variables for mapping and monitoring species populations. Nature Ecology & Evolution 3:539–551. [PubMed: 30858594]

Kandlikar GS, Gold ZJ, Cowen MC, Meyer RS, Freise AC, Kraft NJB, Moberg-Parker J, Sprague J, Kushner DJ, and Curd EE. 2018. ranacapa: an R package and Shiny web app to explore environmental DNA data with exploratory statistics and interactive visualizations. F1000Research 7:1734. [PubMed: 30613396]

Karimi B, et al. 2018. Biogeography of soil bacteria and archaea across France. Science Advances 4:eaat1808. [PubMed: 29978046]

Kobori H, et al. 2016. Citizen science: a new approach to advance ecology, education, and conservation. Ecological Research 31:1–19.

Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, and Bonneau RA. 2015. Sparse and compositionally robust inference of microbial ecological networks. PLoS Computational Biology 11:e1004226. [PubMed: 25950956]

Langmead B, and Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357–359. [PubMed: 22388286]

Larson ER, et al. 2020. From eDNA to citizen science: emerging tools for the early detection of invasive species. Frontiers in Ecology and the Environment 18:194–202.

Latombe G, McGeoch MA, Nipperess DA, and Hui C. 2018. zetadiv: functions to compute compositional turnover using ζ diversity. https://cran.r-project.org/package=zetadiv

Lee J-S, Kim J-W, Lee SH, Myeong H-H, Lee J-Y, and Cho JS. 2016. Zonation and soil factors of salt marsh halophyte communities. Journal of Ecology and Environment 40:4.

Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, and Machida RJ. 2013. A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Frontiers in Zoology 10:34. [PubMed: 23767809]

Lewin HA, et al. 2018. Earth BioGenome Project: sequencing life for the future of life. Proceedings of the National Academy of Sciences USA 115:4325–4333.

Lin M 2021. Data from: Landscape analyses using eDNA metabarcoding and earth observation predict community biodiversity in California (Version 1.0). Ecological Applications. Zenodo, data set. 10.5281/zenodo.4516670

Long JA 2020. interactions: Comprehensive, user-friendly toolkit for probing interactions. https://cran.r-project.org/package=interactions

Mandakovic D, et al. 2018. Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. Scientific Reports 8:5875. [PubMed: 29651160]

Martin M 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10–12.

McKnight MW, White PS, McDonald RI, Lamoreux JF, Sechrest W, Ridgely RS, and Stuart SN. 2007. Putting beta-diversity on the map: broad-scale congruence and coincidence in the extremes. PLoS Biology 5:e272. [PubMed: 17927449]

McMurdie PJ, and Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE 8:e61217. [PubMed: 23630581]

Meyer RS, et al. 2021. The CALeDNA program: citizen scientists and researchers inventory California's biodiversity. California Agriculture 75:20–32.

Miralles L, Dopico E, Devlo-Delva F, and Garcia-Vazquez E. 2016. Controlling populations of invasive pygmy mussel (*Xenostrobus securis*) through citizen science and environmental DNA. Marine Pollution Bulletin 110:127–132. [PubMed: 27381987]

Montagna M, et al. 2018. Differential biodiversity responses between kingdoms (plants, fungi, bacteria and metazoa) along an Alpine succession gradient. Molecular Ecology 27:3671–3685. [PubMed: 30146795]

Moreno J, Terrones A, Juan A, and Alonso MÁ. 2018. Halophytic plant community patterns in Mediterranean saltmarshes: shedding light on the connection between abiotic factors and the distribution of halophytes. Plant and Soil 430:185–204.

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, and Kent J. 2000. Biodiversity hotspots for conservation priorities. Nature 403:853–858. [PubMed: 10706275]

Nieto-Lugilde D, Maguire KC, Blois JL, Williams JW, and Fitzpatrick MC. 2018. Multiresponse algorithms for community-level modelling: review of theory, applications, and comparison to species distribution models. Methods in Ecology and Evolution 9:834–848.

Oksanen J, et al. 2018. vegan: community ecology package. https://cran.r-project.org/web/packages/vegan

Omernik JM, and Griffith GE. 2014. Ecoregions of the conterminous United States: evolution of a hierarchical spatial framework. Environmental Management 54:1249–1266. [PubMed: 25223620]

Pereira HM, et al. 2013. Essential biodiversity variables. Science 339:277–278. [PubMed: 23329036]

Peters MK, et al. 2019. Climate–land-use interactions shape tropical mountain biodiversity and ecosystem functions. Nature 568:88–92. [PubMed: 30918402]

Pettorelli N, Safi K, and Turner W. 2014. Satellite remote sensing, biodiversity research and conservation of the future. Philosophical Transactions of the Royal Society B: Biological Sciences 369:20130190.

Pimm SL, Jenkins CN, Abell R, Brooks TM, Gittleman JL, Joppa LN, Raven PH, Roberts CM, and Sexton JO. 2014. The biodiversity of species and their rates of extinction, distribution, and protection. Science 344:1246752. [PubMed: 24876501]

Pitcher CR, Ellis N, and Smith SJ. 2011. Example analysis of biodiversity survey data with R package gradientForest. 16. http://gradientforest.r-forge.r-project.org/biodiversity-survey.pdf

Pollock LJ, O'Connor LMJ, Mokany K, Rosauer DF, Talluto MV, and Thuiller W. 2020. Protecting biodiversity (in all its complexity): new models and methods. Trends in Ecology & Evolution 35:1119–1128. [PubMed: 32977981]

Prosser JI 2010. Replicate or lie. Environmental Microbiology 12:1806–1810. [PubMed: 20438583]

R Core Team. 2019. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org

Ranjard L, et al. 2013. Turnover of soil bacterial diversity driven by wide-scale environmental heterogeneity. Nature Communications 4:1434.

Rocchini D, et al. 2021. From zero to infinity: minimum to maximum diversity of the planet by spatio-parametric Rao's quadratic entropy. Global Ecology and Biogeography 30:1153–1162.

Schimel D, Schneider FD, Carbon JPL, and Participants E. 2019. Flux towers in the sky: global ecology from space. New Phytologist 224:570–584. [PubMed: 31112309]

Schneider FD, Ferraz A, Hancock S, Duncanson LI, Dubayah RO, Pavlick RP, and Schimel DS. 2020. Towards mapping the diversity of canopy structure from space with GEDI. Environmental Research Letters, 15:115006.

Schneider FD, Morsdorf F, Schmid B, Petchey OL, Hueni A, Schimel DS, and Schaepman ME. 2017. Mapping functional diversity from remotely sensed morphological and physiological forest traits. Nature Communications 8:1441.

Seeber PA, McEwen GK, Löber U, Förster DW, East ML, Melzheimer J, and Greenwood AD. 2019. Terrestrial mammal surveillance using hybridization capture of environmental DNA from African waterholes. Molecular Ecology Resources 19:1486–1496. [PubMed: 31349392]

Sessitsch A, Weilharter A, Gerzabek MH, Kirchmann H, and Kandeler E. 2001. Microbial population structures in soil particle size fractions of a long-term fertilizer field experiment. Applied and Environmental Microbiology 67:4215–4224. [PubMed: 11526026]

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, and Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research 13:2498–2504. [PubMed: 14597658]

Shirazi S, Meyer R, and Shapiro B. 2020. Revisiting the effect of PCR replication and sequencing depth on biodiversity metrics in environmental DNA metabarcoding. Authorea, preprint. https://www.authorea.com/users/336873/articles/462535-pcr-replication-in-environmental-dna-metabarcoding

Silva AC, and Souza AF. 2018. Aridity drives plant biogeographical sub regions in the Caatinga, the largest tropical dry forest and woodland block in South America. PLoS ONE 13: e0196130. [PubMed: 29702668]

Simons AL, Mazor R, Stein ED, and Nuzhdin S. 2019. Using alpha, beta, and zeta diversity in describing the health of stream-based benthic macroinvertebrate communities. Ecological Applications 29:e01896. [PubMed: 31051052]

Sousa D, et al. 2021. Tree canopies reflect mycorrhizal composition. Geophysical Research Letters 48:e2021GL092764.

Taberlet P, Bonin A, Zinger L, and Coissac E. 2018. Environmental DNA: for biodiversity research and monitoring. Oxford University Press, Oxford, UK.

Theobald EJ, et al. 2015. Global change and local solutions: tapping the unrealized potential of citizen science for biodiversity research. Biological Conservation 181:236–244.

Thompson LR, et al. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463. [PubMed: 29088705]

Timling I, Walker DA, Nusbaum C, Lennon NJ, and Taylor DL. 2014. Rich and cold: diversity, distribution and drivers of fungal communities in patterned-ground ecosystems of the North American Arctic. Molecular Ecology 23:3258–3272. [PubMed: 24689939]

Tipton L, Müller CL, Kurtz ZD, Huang L, Kleerup E, Morris A, Bonneau R, and Ghedin E. 2018. Fungi stabilize connectivity in the lung and skin microbial ecosystems. Microbiome 6:12. [PubMed: 29335027]

University of California Conservation Genomics Consortium CALeDNA 2021. www.ucedna.com

U.S. Environmental Protection Agency. 2010. NA_CEC_Eco_Level2. U.S. EPA Office of Research and Development (ORD) – National Health and Environmental Effects Research Laboratory (NHEERL), Corvallis, Oregon, USA ftp://ftp.epa.gov/wed/ecoregions/cec_na/NA_CEC_Eco_Level2.zip.

U.S. Environmental Protection Agency. 2012. Level III ecoregions of California. U.S. EPA Office of Research and Development (ORD) – National Health and Environmental Effects Research Laboratory (NHEERL), Corvallis, Oregon, USA ftp://newftp.epa.gov/EPADataCommons/ORD/Ecoregions/ca/ca_eco_l3.zip.

USDA Forest Service. 2007. USDA Forest Service. 2007. USDA Ecoregion Sections, California. USDA Forest Service, Pacific Southwest Region, Remote Sensing Lab. https://databasin.org/datasets/81a3a809a2ae4c099f2e495c0b2ecc91

van den Hoogen J, et al. 2019. Soil nematode abundance and functional group composition at a global scale. Nature 572:194–198. [PubMed: 31341281]

Wang R, Gamon JA, Cavender-Bares J, Townsend PA, and Zygielbaum AI. 2018. The spatial sensitivity of the spectral diversity–biodiversity relationship: an experimental test in a prairie grassland. Ecological Applications 28:541–556. [PubMed: 29266500]

White TJ, Bruns T, Lee SJWT, and Taylor J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. PCR Protocols: A Guide to Methods and Applications 18:315–322.

White HJ, et al. 2020. Methods and approaches to advance soil macroecology. Global Ecology and Biogeography 29:1674–1690.

Wildlife Conservation Society and Center for International Earth Science Information Network, Columbia University. 2005. Last of the Wild Project, Version 2, 2005 (LWP-2): Global Human Footprint Dataset (Geographic). NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, New York, USA. 10.7927/H4M61H5F

Yamasaki E, et al. 2017. Genomics meets remote sensing in global change studies: monitoring and predicting phenology, evolution and biodiversity. Current Opinion in Environmental Sustainability 29:177–186.

Yang T, Adams JM, Shi Y. u., He J-S, Jing X, Chen L, Tedersoo L, and Chu H. 2017. Soil fungal diversity in natural grasslands of the Tibetan Plateau: associations with plant diversity and productivity. New Phytologist 215:756–765. [PubMed: 28542845]

Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, and Ding Z. 2012. Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring: biodiversity soup. Methods in Ecology and Evolution 3:613–623.

Zarnetske PL, et al. 2019. Towards connecting biodiversity and geodiversity across scales with satellite remote sensing. Global Ecology and Biogeography 28:548–556. [PubMed: 31217748]

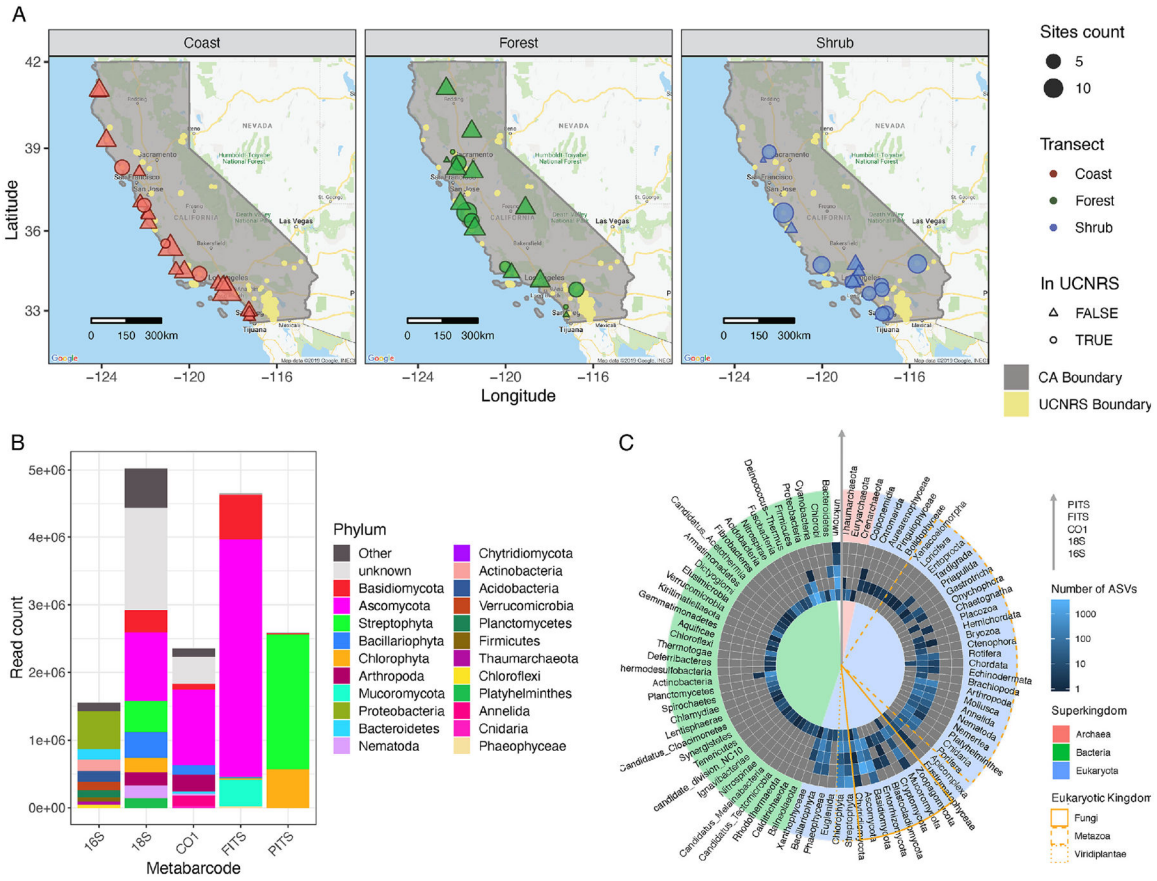

Fig. 1.

Map of 278 sites included in this study and illustration of taxonomic entries recovered with five metabarcodes. (A) Study area (gray shade) is defined within the State of California, United States. Sample sites are colored by three transect designations: coast (red), forest (green), and shrub (blue). Size of the points corresponds to the number of samples taken in the same area. The shape of the points represents areas within (circles) and outside (triangles) of the University of California's Natural Reserve System (UCNRS, yellow shade, area size not to scale for visibility). (B) Read abundance is grouped by the phylum they belong to after taxonomy assignment and decontamination for five metabarcodes targeting Bacteria and Archaea (16S), Eukaryota (18S), Metazoa (CO1), Fungi (FITS), and Viridiplantae (PITS). Only the most abundant 10 phyla are plotted for each metabarcode. All other phyla are summarized in the "Other" category. (C) Heatmap shows each metabarcode's taxonomic specificity. The results from each metabarcode (16S, 18S, CO1, FITS, PITS) are represented from inner to outer rings (gray arrow). Lighter blue in one cell represents more taxonomic entries were recovered by that metabarcode for that phylum, gray color represents no entries. Phyla are indicated on the periphery. Background color of each pie wedge denotes the superkingdom (red, Archaea; blue, Eukaryota; green, Bacteria; no background, unknown) to which the phyla belonged at the time of taxonomy assignment (taxonomy file downloaded from NCBI on 19 January 2018). For eukaryotic phyla, kingdoms are marked by different line types in an orange outline: Fungi (solid), Metazoa (dashed), and Viridiplantae (dotted).

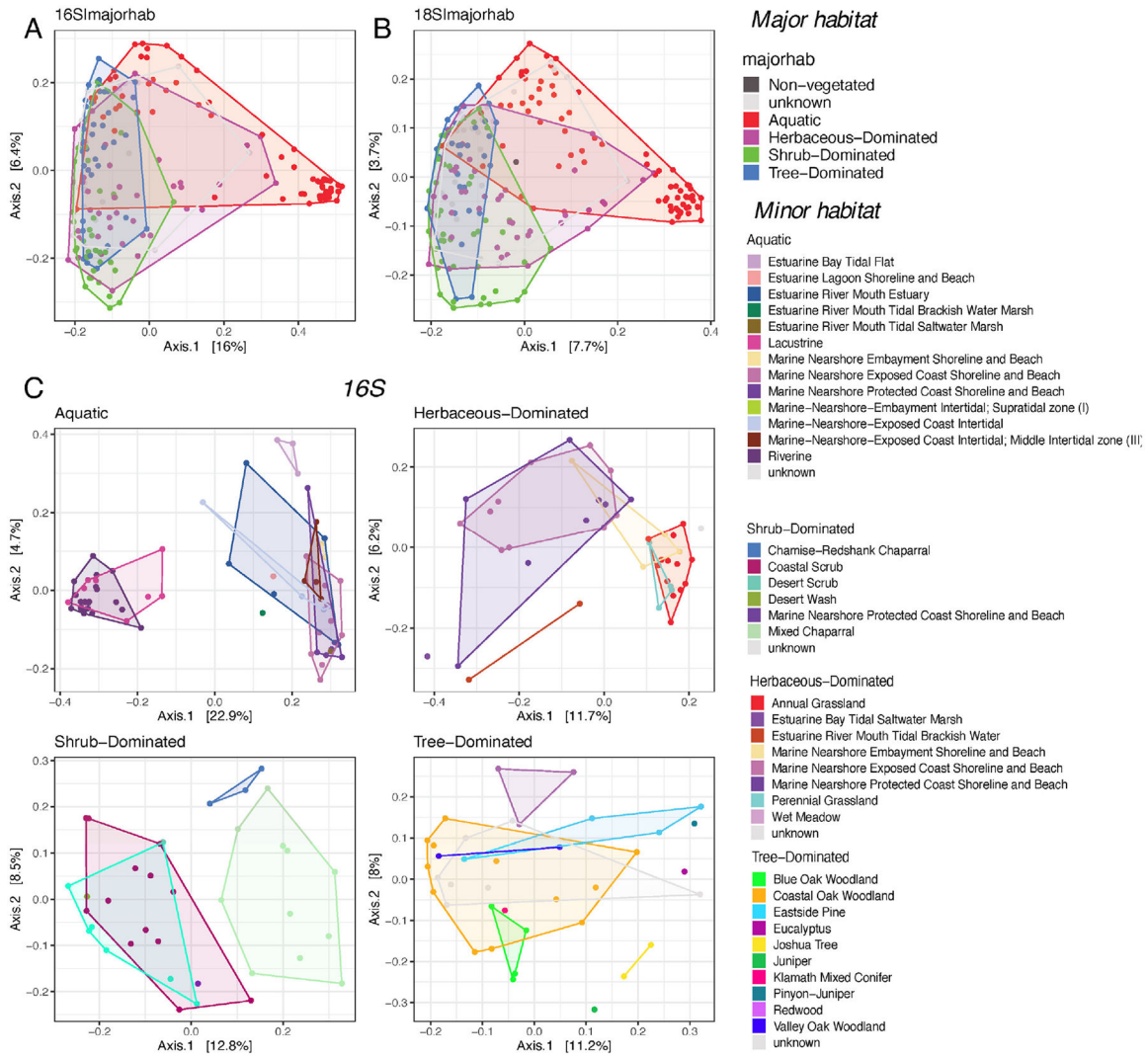

Fig. 2.

Beta diversity plots based on Jaccard dissimilarity. The first two principal coordinates are plotted with percentage of variance explained included in the axis label. We show selected Principal Coordinate Analysis (PCoA) plots from (A) 16S and (B) 18S for major habitat. Each point stands for a sample site. (C) Example PCoA plots based on Jaccard dissimilarity with samples grouped by minor habitat and plotted within aquatic major habitat for 16S metabarcode. Some minor habitat groups separate while others overlap, and patterns of compositional similarity (overlap) are different for different metabarcodes (Appendix S1: Fig. S20).

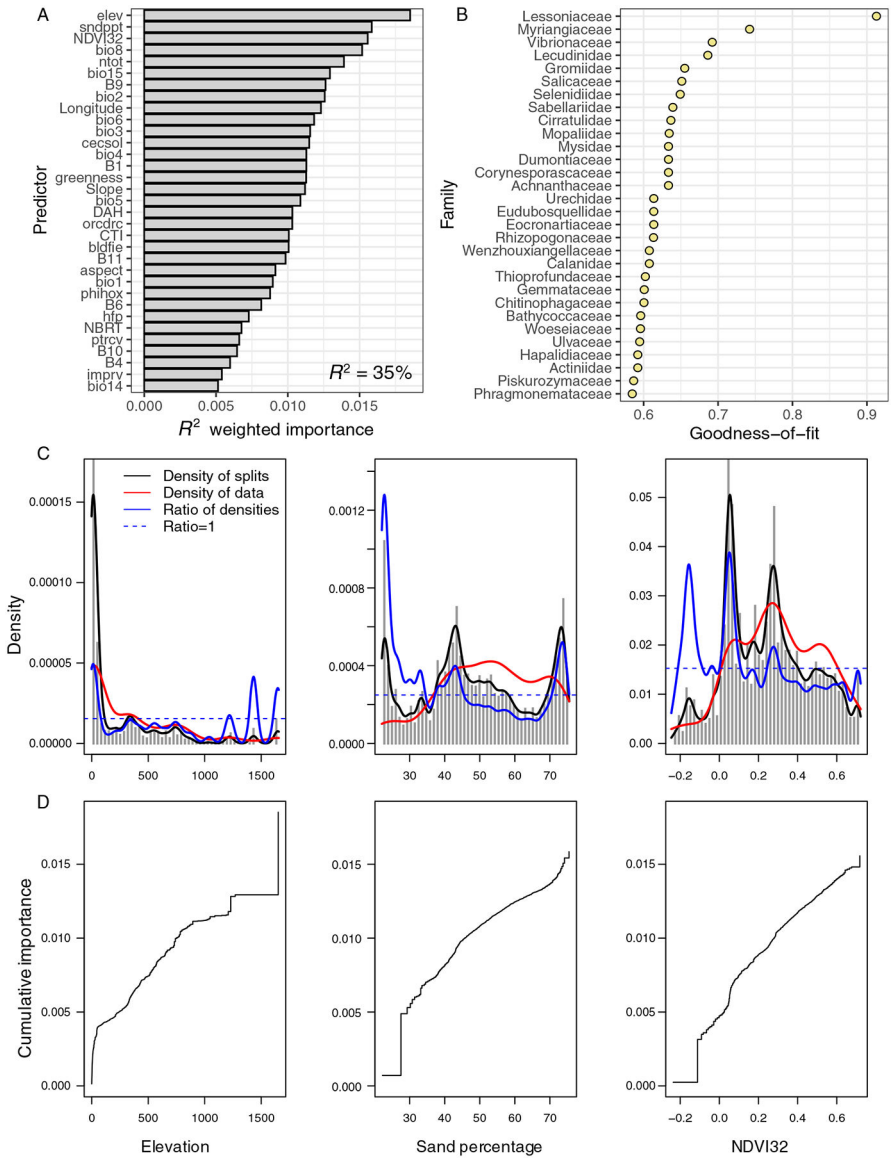

Fig. 3.

Gradient forest result for filtered CALeDNA data set. (A) Ranked overall importance for 33 environmental predictors. (B) Ranked goodness-of-fit (1 − relative error rates) for the top 30 families (response variables). (C and D) Community turnover along the three most important environmental gradients: elevation, sand percentage, and photosynthetic activity proxy (NDVI32). (C) The gray histogram shows binned split importance at each gradient. Kernel density of splits (black lines), of observed predictor values (red lines) and of splits standardized by observation density (blue lines) are overlaid. The horizontal dashed line indicates where the ratio is 1. Each curve integrates to the importance of the predictor. (D) The line shows cumulative importance distributions of splits improvement scaled by R^2 weighted importance and standardized by density of observations, averaged over all families.

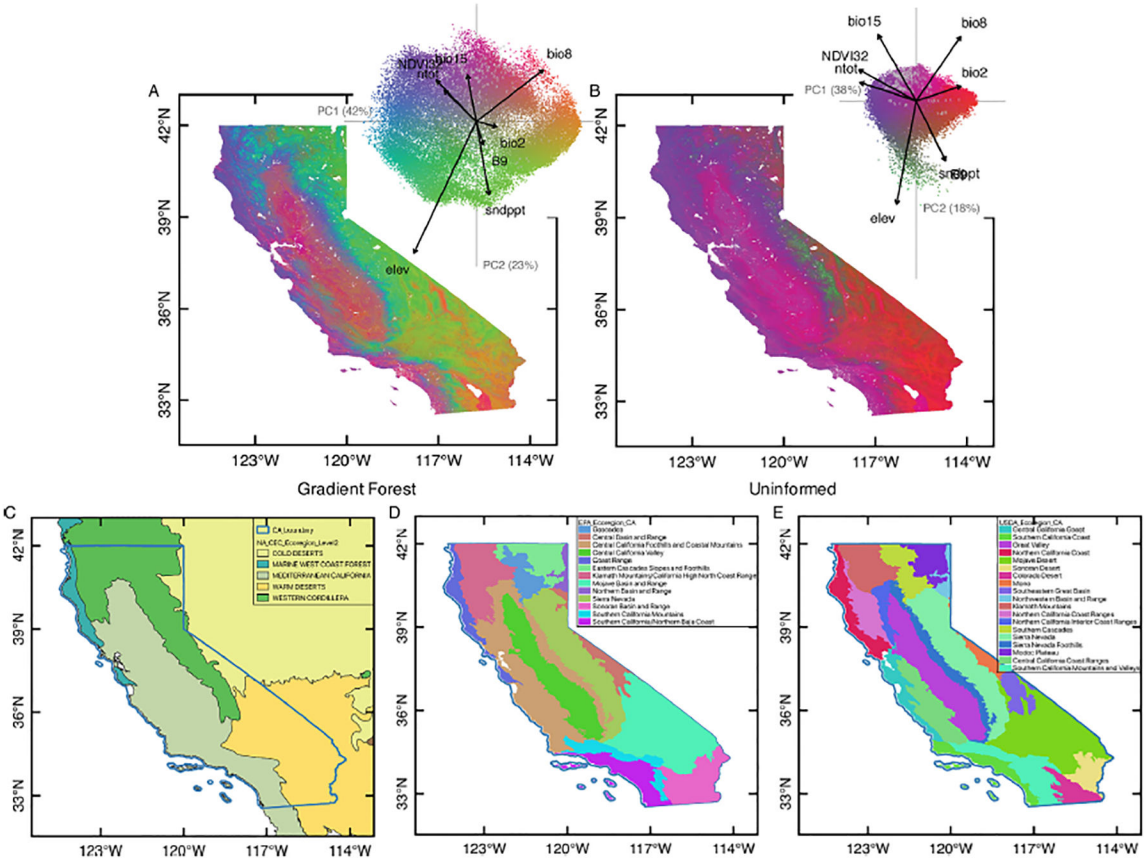

Fig. 4.

Gradient forest predicted community turnover map in California. (A) Map of transformed environmental variables following gradient forest predictions of biodiversity turnover from eDNA results compared with (B) uninformed, standardized environmental variables and (C–E) current major ecoregion maps in California. The map shows the first three principal dimensions of (A) biologically predicted or (B) uninformed community compositions with an RGB color palette with 100-m resolution. The biplot of the first two PCs of the transformed environment space with (inset A) or without (inset B) biological information provides a color key for the compositional variation (n = 50,000). Similar colors approximate similar community in the transformed environmental space. The gray crosses denote the input eDNA sites (n = 272). Vectors denote the direction and magnitude of the eight most important environmental correlates. (C–E) Selected major ecoregions maps are provided for comparisons with (A) the gradient forest map. (C) EPA Level II Ecoregions of North America (U.S. Environmental Protection Agency 2010). (D) EPA Level III Ecoregions of California (U.S. Environmental Protection Agency 2012). (E) USDA Ecoregion Sections in California (USDA Forest Service 2007).

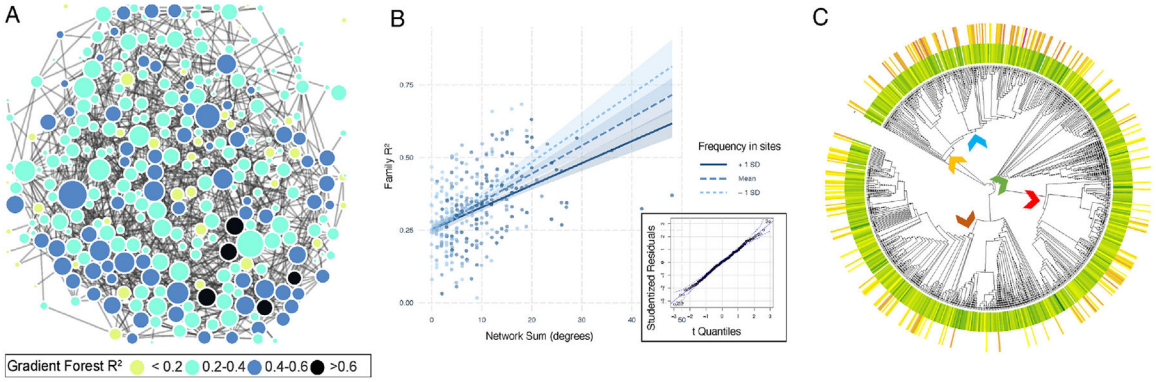

Fig. 5.

eDNA-based ecological co-occurrence network and relationship with gradient forest model goodness-of-fit R^2. (A) A total of 369 families (as nodes) are included in the network and 290 of those have at least one edge connecting them to another node. Dark blue and black nodes represent families with R^2 predictor values >0.4. The size of the node is scaled to the number of network degrees. (B) OLS linear regression and quantile-quantile plot showing the interaction between network sum of degrees and frequency of taxa in sample sites with the dependent variable of gradient forest family goodness-of-fit R^2. There were 304 families included as joint observations in gradient forest and network results. The adjusted $R^2 = 0.22$, network sum estimate = 0.01 ($t = 5.44$; $P = 0.00$), frequency in sites estimate = 0.00 ($t = 0.18$; $P = 0.86$), and interaction between network sum and frequency in sites = 0.00 ($t = -2.38$; $P = 0.02$). (C) Phylogenetic tree made with the Open Tree of Life targeting input families as tips. Heat map labels correspond to the range of gradient forest R^2 (0.078–0.913) from yellow to dark green (inner circle), and to the range of network degrees (0–48) from yellow to purple (outer circle). Families too rare to be included in the network analysis (in fewer than 28 sites) are not colored in heat maps. Arrows indicate the following clades: brown, fungi; mustard, Enterobacteriaceae; blue, Flavobacteriia; green, Streptophyta; red, SAR supergroup.

Table 1.

List of the categorical and a reduced set of numerical variables used in the diversity analysis and gradient forest modeling.

Variable	Category	Description and definition
Categorical variables		
loc	location	name of places visited reported by volunteers
clust	location	neighboring cluster of sites within a radius of 0.5 km derived from GPS record
ecoregion	habitat	EPA Level III Ecoregions of California (Conterminous United States)
majorhab	habitat	major habitat type classified according to California Wildlife Habitat Relationships System
minorhab	habitat	minor habitat type classified according to California Wildlife Habitat Relationships System
transect	habitat	original classification of the predominant biome type (coast/coastal, shrub/shrub-scrub, and forest)
NLCD	habitat	USGS national land cover classification 2011
SoS	soil properties	volunteers' classification of substrate type (sediment, soil, sand)
taxousda	soil properties	predicted most probable class in USDA soil taxonomy
Reduced set of numerical variables		
Longitude	location	longitude of sample sites
hfp	habitat	global human footprint index
bio1	BIOCLIM	annual mean temperature
bio2	BIOCLIM	mean diurnal range (mean of monthly (maximum temperature − minimum temperature))
bio3	BIOCLIM	isothermality (BIO2/BIO7) (× 100)
bio4	BIOCLIM	temperature seasonality (standard deviation × 100)
bio5	BIOCLIM	maximum temperature of warmest month
bio6	BIOCLIM	minimum temperature of coldest month
bio8	BIOCLIM	mean temperature of wettest quarter
bio14	BIOCLIM	precipitation of driest month
bio15	BIOCLIM	precipitation seasonality (coefficient of variation)
phihox	soil properties	soil pH × 10 in H_2O at depth 0.00 m
orcdrc	soil properties	Soil organic carbon content (fine earth fraction) in g/kg at depth 0.00 m
cecsol	soil properties	cation exchange capacity of soil in cmolc/kg at depth 0.00 m
sndppt	soil properties	sand content (50–2,000 μm) mass fraction in percent at depth 0.00 m
bldfie	soil properties	bulk density (fine earth) in kg/m^3 at depth 0.00 m
ntot	soil properties	mass percentage of total nitrogen at depth 0.00 m

Variable	Category	Description and definition
elev	topography	elevation of sample sites
Slope	topography	the rate of change of elevation for each digital elevation model (DEM) cell
aspect	topography	the direction of the maximum rate of change in the z value from each cell in a raster surface
CTI	topography	compound topographic index
DAH	topography	diurnal anisotropic heating
B1	vegetation	Sentinel-2 spectral band 1 (wavelength: 443.9 nm (S2A)/442.3 nm (S2B); description: aerosols)
B4	vegetation	Sentinel-2 spectral band 4 (wavelength: 664.5 nm (S2A)/665 nm (S2B); description: red)
B6	vegetation	Sentinel-2 spectral band 6 (wavelength: 740.2 nm (S2A)/739.1 nm (S2B); description: red edge 2)
B9	vegetation	Sentinel-2 spectral band 9 (wavelength: 945 nm (S2A)/943.2 nm (S2B); description: water vapor)
B10	vegetation	Sentinel-2 spectral band 10 (wavelength: 1,373.5 nm (S2A)/1,376.9 nm (S2B); description: cirrus)
B11	vegetation	Sentinel-2 spectral band 11 (wavelength: 1,613.7 nm (S2A)/1,610.4 nm (S2B); description: SWIR 1)
NDVI32	vegetation	Normalized Difference Vegetation Index in 32-d period
NBRT	vegetation	Normalized Burn Ratio Thermal index in 32-d period
greenness	vegetation	annual greenest pixel in the year of 2017
imprv	habitat	percentage of the pixel covered by developed impervious surface
ptrcv	habitat	percentage of the pixel covered by tree canopy

Note: For a complete list of variables, detailed description, and data accession information, refer to Data S1.

Table 2.

Post hoc fitting of environmental variables on PCoA ordination (Envfit) for each metabarcode.

Metabarcode	First variable	R^2	Second variable	R^2	Third variable	R^2
16S	NDVI32	0.49	greenness	0.47	B1	0.42
18S	NDVI32	0.51	greenness	0.49	B1	0.43
CO1	orcdrc	0.41	ptrcv	0.36	NBRT	0.33
FITS	greenness	0.52	B1	0.5	orcdrc	0.46
PITS	bio3	0.21	sndppt	0.2	B11	0.13

Notes: Here, we present the three significant ($P < 0.001$) environmental variables with the highest correlation coefficient. The significance of the correlation was tested by 1999 permutations. For a complete result of all variables, please refer to Data S13. The direction of changes is included in Appendix S1: Fig. S23.

Table 3.

Variation in ζ_4 (zeta) diversity attributed to geographic separation distance between site clusters (VarDistance) vs. variation in an environmental factor group between the same site clusters (VarFactor).

Metabarcode	Factor group	No. samples	VarFactor (%)	VarDistance (%)	VarUnknown (%)
16S	location	184	0.00	0.29	99.70
16S	topography	184	0.94	0.17	98.90
16S	habitat	156	1.33	0.00	98.70
16S	vegetation	169	5.92	0.00	94.10
16S	Bioclim	184	7.17	0.00	92.20
16S	soil properties	180	9.21	0.00	90.70
18S	location	184	0.14	0.00	99.90
18S	habitat	156	5.49	0.00	94.50
18S	topography	184	7.15	0.00	92.80
18S	Bioclim	184	7.30	0.00	92.70
18S	soil properties	180	15.30	0.00	84.70
18S	vegetation	169	18.50	0.00	81.50
CO1	location	184	0.11	0.22	99.60
CO1	habitat	156	1.86	0.00	98.10
CO1	topography	184	3.30	0.46	96.20
CO1	Bioclim	184	12.00	0.00	88.00
CO1	vegetation	169	18.20	0.31	81.10
CO1	soil properties	180	18.60	0.00	81.30
FITS	topography	184	0.69	0.55	98.70
FITS	location	184	0.93	0.38	98.20
FITS	habitat	156	2.24	0.37	97.10
FITS	Bioclim	184	18.50	0.00	80.40
FITS	soil properties	180	22.40	0.00	77.50
FITS	vegetation	169	32.40	1.05	66.40
PITS	location	184	0.03	0.00	100.00
PITS	Bioclim	184	1.30	0.00	98.70
PITS	habitat	156	2.16	0.00	97.80

Metabarcode	Factor group	No. samples	VarFactor (%)	VarDistance (%)	VarUnknown (%)
PITS	topography	184	2.98	0.03	96.90
PITS	soil properties	180	4.23	0.00	95.70
PITS	vegetation	169	9.00	0.00	91.00

Notes: Within each metabarcode, factor groups were ordered from lowest to highest contributions to variations in zeta diversity. Communities were defined at family levels.