## Dataset Title

Palmyra Atoll Terrestrial Predator DNA Diet: 2015-2017

## Abstract

These are data and code associated with terrestrial predator diet DNA collected from Palmyra Atoll (2015-2017). These datasets include the raw sequencing data, all downstream datasets, body size data used to analyze predator-prey interactions, and taxonomic assignments collected from database searches on BOLD and GenBank (accessed in 2019-2020). The code includes code to reproduce all bioinformatics (merge, filter, match to taxonomies, rarefy, sort) as well as all body size determinations and the statistics and figures generated from analyses. Raw data are from DNA extractions of predator gut regions (abdomens and opisthosomas) and amplification of the CO1 gene using PCR. Predator species include common spiders, insects, and a centipede all collected via various collection methods (hand, insecticide fogging) and collected individually with sterile implements. Data were collected at the individual predator level and were collected to examine patterns in predator-prey interactions and food-webs in terrestrial invertebrate communities.

## Creators

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| First Name | Middle Initial | Last Name | Organization | e-mail address | ORCID ID (optional) |
| Ana | T | Miller-ter Kuile | UC Santa Barbara | Ana.miller.ter.kuile@gmail.com |  |
| Austen |  | Apigo | UC Santa Barbara |  |  |
| An |  | Bui | UC Santa Barbara |  |  |
| Bartholomew |  | DiFiore | UC Santa Barbara |  |  |
| Elizabeth | S | Forbes | UC Santa Barbara |  |  |
| Michelle |  | Lee | UC Santa Barbara |  |  |
| Devyn |  | Orr | UC Santa Barbara |  |  |
| Daniel | L | Preston | Colorado State University |  |  |
| Rachel |  | Behm | UC Santa Barbara |  |  |
| Taylor |  | Bogar | University of Hong Kong |  |  |
| Jasmine |  | Childress | UC Santa Barbara |  |  |
| Rodolfo |  | Dirzo | Stanford University |  |  |
| Maggie |  | Klope | UC Santa Barbara |  |  |
| Kevin | D | Lafferty | USGS, Western Ecological Research Center |  |  |
| John |  | McLaughlin | UC Santa Barbara |  |  |
| Marisa |  | Morse | UC Santa Barbara |  |  |
| Carina |  | Motta | UC Santa Barbara |  |  |
| Kevin |  | Park | UC Santa Barbara |  |  |
| Katherine |  | Plummer | Stanford University |  |  |
| David |  | Weber | Stanford University |  |  |
| Ronny |  | Young | UC Santa Barbara |  |  |
| Hillary | S | Young | UC Santa Barbara |  |  |

## Other personnel names and roles

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| First Name | Middle Initial | Last Name | Organization | e-mail address | ORCID ID (optional) | Role in project |
| Ana | T | Miller-ter Kuile | UC Santa Barbara | Ana.miller.ter.kuile@gmail.com |  | Contact |
| Magalay |  | Espinoza | UC Santa Barbara |  |  | Field crew |
| Cora |  | Johnston | UC Santa Barbara |  |  | Field Crew |
| Chelsea |  | Steel | UC Santa Barbara |  |  | Lab tech |
| Emily |  | Lutz | UC Santa Barbara |  |  | Lab tech |
| Tessa |  | Chou | UC Santa Barbara |  |  | Lab tech |

## Keywords

Allometry, arthropod, centipede, DNA metabarcoding, hunting strategy, insect, spider

## Funding of this work:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| PI First Name | PI Middle Initial | PI Last Name | PI ORCID ID (optional) | Title of Grant | Funding Agency | Funding Identification Number |
| Hillary | S | Young |  |  | National Science Foundation | DEB-1457371 |
| Rodolfo |  | Dirzo |  |  | National Geographic Society |  |
| Hillary | S | Young |  |  | UC Santa Barbara Faculty Senate Grant |  |

## Timeframe

* Begin date: July 2015
* End date: May 2021

## Geographic location

* Verbal description: Palmyra Atoll National Wildlife Refuge, Northern Line Islands
* North bounding coordinate: 5.883333
* South bounding coordinate: -162.08333

## Taxonomic species or groups

Phylum: Arthropoda, Class: Arachnida, Order: Araneae;

1. *Opopaea sp*. (Family: Oonopidae)

2. *Neoscona theisi* (Family: Aranaiedae),

3. *Heteropoda venatoria* (Family: Sparassidae)

4. *Smeringopus pallidus* (Family: Pholcidae)

5. *Scytodes longipes* (Family: Scytodidae)

Phylum: Arthropoda, Class: Insecta;

6. *Pantala flavescens* (Order: Odonata, Family: Libellulidae)

7. *Phisis holdhausi* (Order: Orthoptera, Family: Tettigoniidae)

8. *Euborellia annulipes* (Order: Dermaptera, Family: Anisolabidae)

Phylum: Arthropoda, Class: Chilopoda, Order: Geophilomorapha, Family: Mecistocephalidae

9. *Mecistocephalus sp*.

## Methods

*Field site and collections*

We conducted this work on Palmyra Atoll National Wildlife Refuge, Northern Line Islands (5º53’ N, 162º05’W). Palmyra Atoll has a well-characterized species list, and like many atolls, is relatively species poor, allowing for detailed characterization of potential diet items (Handler et al. 2007). Predator individuals were collected across habitat types, including different forest types and microhabitats (e.g., understory vegetation, canopy vegetation, and soil types). For each of these habitat types, we used a combination of methods, including individual collection during visual surveys for understory, and soil collections and canopy fogging with insecticide onto collection sheets for canopy individuals. All individuals were collected individually with sterilized implements (ethanol-burned forceps) in sterilized collection containers containing 95% EtOH to avoid contamination (Greenstone et al. 2011). All individuals were stored in 95% EtOH at -20ºC before DNA extraction.

We identified all predators to morphospecies using a species list for Palmyra Atoll (Handler et al. 2007) and later validated unique species by DNA metabarcoding sequence data. The predators sampled represent the most common predator species found in each habitat location and span a body size range of 0.2 – 998 mg (wet mass). These predators included five arachnid species (*Opopaea sp*., *Neoscona theisi*, *Heteropoda venatoria*, *Smeringopus pallidus*, and *Scytodes longipes*), one dragonfly (*Pantala flavescens*), one predatory katydid (*Phisis holdhausi*), one earwig (*Euborellia annulipes*), and one soil-dwelling centipede species (*Mecistocephalus sp*.). These predators use various hunting tools, including webs and venom and employ several different hunting strategies, including active hunting and non-active hunting (e.g., sit-and-wait or ambush).

*DNA extraction, PCR amplification, library preparation, sequencing, and denoising*

We individually measured the length of each predator (mm) and separated the thorax, opisthosoma, or trunk (depending on predator species, (Krehenwinkel et al. 2017, Macías-Hernández et al. 2018)) for DNA extraction following a modified CTAB extraction protocol (Fulton et al. 1995). While most individuals were run in separate samples (70%, n = 121/173), some individuals were too small to extract ample DNA from only one individual (mean size of 4.04 ± 0.12 mm in total length), and so we combined these individuals with other individuals from the same species, size range (within ± 0.5 mm in length), and sampling period. For these combined samples, we aimed for a minimum total sample weight of 5mg, and ideal sample weights of 10-20mg, a range we had previously determined to be sufficient for downstream DNA extraction and cleaning protocols. This resulted in a maxiumum of 12 individuals in one sample (SI Figure 6). Following methods in (Krehenwinkel et al. 2017), we standardized concentrations of 40uL of each sample to 20ng/ul and used Ampure XP (Agencourt, Beverly, MA, USA) beads to remove higher molecular weight predator DNA prior to PCR steps. We then amplified the CO1 gene, which is well-represented in online databases (Porter and Hajibabaei 2018) with general metazoan primers (mlCOIintF/Fol-degen-rev; (Yu et al. 2012, Leray et al. 2013, Krehenwinkel et al. 2017)). We ran total reaction volumes per sample of 25μL, with 9μL nuclease free water, 12.5μL GoTaq Green Master Mix (Promega Corp., Madison, WI, USA), 1.25μL of each primer (at 10mM), and 1μL of DNA template (at 10ng/μL) and ran a duplicate for each sample. We followed a PCR protocol as follows: 3 minutes at 95ºC, 35 cycles of: 95ºC for 30 seconds, 46ºC for 30 seconds, 72ºC for one minute; ending with 72ºC for five minutes. We removed reaction dimer with Ampure XP beads at 0.8x bead-to-DNA ratio. We then attached Illumina index primers (Nextera XT Index Kit v2) with 5μL of PCR product per reaction and the recommended PCR protocol for these primers (Illumina 2009). We combined and cleaned successfully amplified duplicate samples using Ampure XP beads (0.7x beads-to-DNA) and diluted each sample to 5nM in 10mM TRIS, using 1uL of each sample for sequencing.

Because of the sample size and the need for a large number of sequences per predator in order to detect rarer prey DNA ((Krehenwinkel et al. 2017), SI Figure 4), we ran samples for this study across four separate sequencing runs. All individuals within a predator species were sequenced on the same run and each run contained one to five predator species. We ran 19 samples of one predator species (*H. venatoria*) across all runs to quantify run-to-run variation in sequencing. For each run, we multiplexed all samples along with one negative control and two PCR4-TOPO TA vectors (Invitrogen, Carlsbad, CA, USA) containing the internal transcribed spacer 1 region from two fungal species as positive controls (GenBank accession numbers: MG840195 and MG840196; (Toju et al. 2012, Clark et al. 2016, Apigo and Oono 2018)). We submitted multiplexed samples for sequencing at the University of California, Santa Barbara Biological Nanostructures Laboratory Genetics Core. Samples were run on an Illumina MiSeq platform (v2 chemistry, 500 cycles, paired-end reads) with a 15% spike-in of PhiX. Following sequencing, samples were demultiplexed using Illumina’s bcl2fastq conversion software (v2.20) at the Core facility.

We merged, filtered (max ee = 1.0), and denoised (clustered) our sequences around amplicon sequence variants (ASVs) using the DADA2 algorithm in R (dada2 package version 1.1.14.0; Callahan et al., 2016). Prior to denoising with DADA2, we used cutadapt (version 1.18, (Martin 2011)) to remove primers from each sequence. We compared results to a similar protocol using the UNOISE3 algorithm (unoise3 function in unoise (Edgar 2016), but found that DADA2 gave more high-read abundance ASVs. We ran DADA2 on sequences from all sequencing runs combined but verified that this was appropriate by first ensuring that error rates per run were similar, following recommendations from the algorithm developers. We removed samples from analysis that had not been sequenced to sufficient depth using iNEXT (Hsieh et al. 2016) and a lower quantile cutoff. We rarefied remaining samples (McKnight et al. 2019) based on the sample with the lowest sequencing depth which had been sequenced with 95%+ sampling completeness based on iNEXT (version 2.0.20) interpolation and extrapolation methods (Hsieh and Chao 2017). We rarefied using the rrarefy() function in the vegan (version 2.5.6) package in R to 15,954 reads per sample.

*ASV taxonomic assignment with BLAST and BOLD*

From the output of the DADA2 algorithm, we created a list of unique ASVs which we matched to taxonomies both in the GenBank and BOLD databases. For GenBank, we used BLAST (version 2.7.1) with the blastn command for taxonomic assignment of each ASV using the computing cluster at UC Santa Barbara, comparing against the GenBank nucleotide database with an evalue of 0.01 (downloaded on November 20, 2019). We visualized and exported taxonomic alignment using MEGAN Community Edition (version 6.18.0, (Huson et al. 2016)), using default settings and selecting the subtree with all possible diet items for this species (Kingdom: Animalia, Clade: Bilateria). For BOLD taxonomic assignment, we used the BOLD IDEngine of the CO1 gene with Species Level Barcode Records (accessed May 21, 2020; 4,070,029 Sequences, 225,114 Species, and 104,607 Interim Species in database) to match each ASV list to taxonomies. We combined taxonomic assignments from both programs and discarded taxonomic assignments that were mismatched at the family level or higher (Elbrecht et al. 2017). We chose to combine prey taxonomies at the family level, similar to diet resolution in both metabarcoding and histological methods in this field (e.g. (Kartzinel et al. 2015, Brose et al. 2019, Eitzinger et al. 2019)) by summing the cumulative read abundances across the ASVs that corresponded to each diet family in each sample. Family-level data provides information comparable to previous studies. Additionally, on Palmyra, each invertebrate family corresponds to an average of 1.9 (± 0.13 SE) species, so for this system a family-level taxonomic assignment may closely mirror species-level assignments. We corrected for potential sequence jumping (‘cross-talk’) across samples by removing reads across samples that emerged in negative controls (Oono et al. 2020) and all DNA matching any predator family present on an individual sequencing run was removed as a conservative method to account for potential sequence jumping (‘cross-talk’) (van der Valk et al. 2020). We verified ASV specificity based on positive control samples.

Prior to data analyses, we verified that samples that consisted of multiple individuals (n = 53) did not represent a disproportionate number of interaction counts by comparing the number of predator-prey interactions observed for samples based on the number of individuals comprising each sample.

*Predator length-mass model*

Because we wanted to compare predator to prey mass, we had to convert the lengths taken on predators to predicted masses. We used mass data collected from predator individuals from Palmrya Atoll and from the literature (Yaninek and Gnanvossou 1993, Sohlström et al. 2018, Su et al. 2020), Miller-ter Kuile *unpublished data*, McLaughlin et al. *unpublished data*). We fit a linear mixed effects model on log10-log10 transformed mass and length data for these predator individuals. These models included predator length as a predictor of predator mass with a random intercept and slope taking into account by-species variation in the slope and intercept of this relationship (length|species in the random effects model). We assessed model fit for this model and then predicted the values for our predator individuals based on these results. We fit models with the glmmTMB package (version 1.0.2.1) in R (version 4.0.2), assessed model fit with the MuMIn (version 1.43.12) and DHARMa (version 0.3.3.0) packages and used the predict function to predict predator masses from the model results.

*Predator and prey size determination*

We measured the length of each predator individual from the front of the head to the end of the abdomen prior to DNA extraction. We converted predator lengths to wet mass using mass-length scaling relationships for each predator species from existing datasets ((Yaninek and Gnanvossou 1993, Sohlström et al. 2018, Su et al. 2020). Prey species masses were taken as the average mass for individuals across species within each family. Averaging prey size by family and using average prey masses in predator-prey mass scaling studies is a common method in the field, and though not being able to assign prey mass is a limitation of diet DNA metabarcoding data, compiling data in this way allows for comparisons with recent synthetic studies (Brose et al. 2019). In other words, here, we do not report the size of prey individuals that were eaten; rather, for the prey families that were eaten, we report their average body sizes observed in the field.

*Data analyses*

To determine whether individual predator size, species, or both predicted prey size, we fit a linear mixed effects model with the response variable of log10 prey mass (in mg) and predictor variables of log10 predator mass (in mg), species identity, and their interaction, with random intercepts by predator individual to account for dependence among multiple prey species observations within each individual predator. Then, to explore whether predator hunting traits or predator phylogenetic relatedness influences predator-prey size ratios, we divided predator-prey interactions based on whether or not the predator species uses webs to capture prey or uses venom to subdue prey. We determined the ratio of predator to prey size for each of these interactions (raw predator mass/prey mass) and then built a set of linear mixed models of this ratio (log transformed for data normality) as the response variable, with each type of predator trait as a predictor variable (one model with web-building and one with venom use). We compared these to two predator species relatedness models – choosing to compare the ratio of predator to prey size based on predator species and predator class, with the aim to determine whether, if hunting traits did not influence size selection, individuals within shared taxonomic groups had conserved size ratios. In each of these models, we used a nested random intercept term of predator individual within species. We considered the species model to be the model without any fixed effects and the random effects of the other models (i.e., including predator individual and predator species).

*Statistical model selection*

For the linear mixed effects models examining how predator size and species identity shape prey size, we performed model selection using the dredge() function in the MuMIn package in R (package version 1.43.17, (Barton 2020)) to compare nested models (n = 5 models) and chose the model with the lowest AICc value. To compare the predator trait and phylogeny models, we performed model selection by comparing AICc values for these models (along with a null model with no predictor variables [n = 5 total models]). For all models, we verified model assumptions using the DHARMa package in R (version 0.3.3.0, (Hartig 2020)). The color palette in our figures is from the calecopal package (version 0.1.0, (Bui et al. 2020)).

## Data Provenance

Some of the data in this article were generated for this article and we pulled data from other sources as well. Other data sources are listed below

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset title | Dataset DOI or URL | Creator (name & email) | Contact (name & email) |
| Data from: Applying generalized allometric regressions to predict live body mass of tropical and temperate arthropods | ﻿https://doi.org/10.5061/dryad.vk24fr1 | Sohlström, Esra H.  Lucas, Marin  Barnes, Andrew D.  Haneda, Noor F.  Scheu, Stefan  Rall, Björn C.  Brose, Ulrich  Jochum, Malte |  |
| Maximum aerodynamic force production by the wandering glider dragonfly (Pantala flavescens, Libellulidae) | <https://doi.org/10.1242/jeb.218552> | Su, Guanting  Dudley, Robert  Pan, Tianyu  Zheng, Mengzong  Peng, Liansong  Li, Qiushi |  |
| Invertebrate morphometric data collected from Palmyra Atoll National Wildlife Refuge, Northern Islands: August 2009 - November 2016 | See separate metadata file for these data | Young, Hillary  McLaughlin, John  Miller-ter Kuile, Ana  Lafferty, Kevin |  |
| Predator traits determine food-web architecture across ecosystems | ﻿http://dx.doi.org/10.1038/s41559-019-0899-x | Brose, Ulrich  Archambault, Phillippe  Barnes, Andrew D.  Bersier, Louis Felix  Boy, Thomas  Canning-Clode, João  Conti, Erminia  Dias, Marta  Digel, Christoph  Dissanayake, Awantha  Flores, Augusto A.V.  Fussmann, Katarina  Gauzens, Benoit  Gray, Clare  Häussler, Johanna  Hirt, Myriam R.  Jacob, Ute  Jochum, Malte  Kéfi, Sonia  McLaughlin, Orla  MacPherson, Muriel M.  Latz, Ellen  Layer-Dobra, Katrin  Legagneux, Pierre  Li, Yuanheng  Madeira, Carolina  Martinez, Neo D.  Mendonça, Vanessa  Mulder, Christian  Navarrete, Sergio A.  O’Gorman, Eoin J.  Ott, David  Paula, José  Perkins, Daniel  Piechnik, Denise  Pokrovsky, Ivan  Raffaelli, David  Rall, Björn C.  Rosenbaum, Benjamin  Ryser, Remo  Silva, Ana  Sohlström, Esra H.  Sokolova, Natalia  Thompson, Murray S.A.  Thompson, Ross M.  Vermandele, Fanny  Vinagre, Catarina  Wang, Shaopeng  Wefer, Jori M.  Williams, Richard J.  Wieters, Evie  Woodward, Guy  Iles, Alison C. |  |

## Data Tables

\*Note: We did not include summaries of data tables in the “outputs” folder of the data folder. These are all intermediate steps in the data preparation process and the steps to generate them are found in the scripts in the ‘2\_data\_prep’ folder and explained in more detail as they are processed in the ‘2\_data\_prep.Rmd’ file in that folder.

**Table name:** ASVs\_all.fasta

**Table description:** a list of all filtered and merged CO1 sequences from the DADA2 pipeline of all sequencing runs combined

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Description | Unit or  code explanation or date format | Missing value code |
| ASV | A CO1 sequence merged and filtered via DADA2 |  |  |

**Table name:** ASVs\_counts\_all.tsv

**Table description:**a sample by ASV matrix of all the by-sample sequence counts for each ASV from the DADA2 pipeline on all sequencing runs combined

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Description | Unit or  code explanation or date format | Missing value code |
| ASV | A CO1 sequence merged and filtered via DADA2 (matched to sequence in the ASVs\_all.fasta file) | A taxonomic unit from DADA2 |  |
| CEN10b\_S19 – SMEb\_S87 | A sample run through sequencing corresponding to one or a few predator individuals collected from the same size group, environment, and species | A raw number of DNA sequences corresponding to each ASV in each sample |  |

**Table name:** DADA2\_sum\_stats.csv

**Table description:**the filtering and merging statistics per sample for the DADA2 pipeline, in the same order as the sample names in the ASV\_counts\_all.tsv file

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Description | Unit or  code explanation or date format | Missing value code |
| Dada3\_input | The number of sequences in that sample entering the dada2 pipeline | Number of sequences (reads) |  |
| filtered | The number of DNA sequences that made it through the filtering step in each sample | Number of sequences (reads) |  |
| Dada\_f | The number of forward reads recognized as forward reads in dada2 merging step | Number of sequences (reads) |  |
| Dada\_r | The number of reverse sequence reads recognized as reverse reads in the dada2 merging step | Number of sequences (reads) |  |
| Merged | Number of merged forward-reverse sequences in each sample | Number of sequences (reads) |  |
| Nonchim | Number of non-chimeric sequences following chimera detection and removal step | Number of sequences (reads) |  |
| Final\_perc\_reads\_retained | The percent of reads that passed all steps in the dada2 merging, filtering, and cleaning protocol | Percent of sequence reads remaining |  |

**Table name:** denoised.fasta

**Table description:** a list of all filtered and merged CO1 sequences from the unoise3 pipeline of all sequencing runs combined

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Description | Unit or  code explanation or date format | Missing value code |
| ZOTU | A CO1 sequence merged and filtered via UNOISE3 |  |  |

**Table name:** zotu\_table\_a.txt – zotu\_table\_l.txt

**Table description:**a sample by ASV matrix of all the by-sample sequence counts for each ASV from the UNOISE3 pipeline on all sequencing runs combined. They are in separate tables because of the limitations of the 32-bit usearch program.

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Description | Unit or  code explanation or date format | Missing value code |
| #OTU ID | A CO1 sequence merged and filtered via UNOISE3 (matched to sequence in the denoised.fasta file) | A taxonomic unit from unoise3 |  |
| CEN01 – SME14 | A sample run through sequencing corresponding to one or a few predator individuals collected from the same size group, environment, and species | A raw number of DNA sequences corresponding to each ASV in each sample |  |

**Table name:** BOLD\_0.csv – BOLD\_17.csv & bold.csv

**Table description:** taxonomic assignments from BOLD database of all the ASVs in the DADA2 dataset (bold.csv is a combined version of the other csvs combined in R)

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Description | Unit or  code explanation or date format | Missing value code |
| Query ID | A CO1 sequence merged and filtered via DADA2 | A taxonomic unit from dada2 |  |
| Best ID | The best-matching taxonomic ID from the BOLD database | A species, genus, family, or order of the taxonomic ID of each ASV. | No match |
| Search DB | The database searched from BOLD |  |  |
| Top % | The highest match from the BOLD database | Percent of DNA sequence match |  |
| Low % | The lowest match from the BOLD database | Percent of DNA sequence match |  |

**Table name:** bold\_wID.csv

**Table description:** taxonomic assignments from BOLD database of all the ASVs in the DADA2 dataset, similar to the datasets with numbers from BOLD, but now combined (in R) and with taxonomic assignments at multiple taxonomic levels

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Description | Unit or  code explanation or date format | Missing value code |
| ASV | A CO1 sequence merged and filtered via DADA2 | A taxonomic unit from dada2 |  |
| Domain | The domain of the taxonomic assignment |  |  |
| Phylum | The phylum of the taxonomic assignment |  |  |
| Class | The class of the taxonomic assignment |  |  |
| Order | The order of the taxonomic assignment |  |  |
| Family | The family of the taxonomic assignment |  |  |
| Genus | The genus of the taxonomic assignment |  |  |
| Species | The species of the taxonomic assignment |  |  |
| ID\_bold | The original ID given by the BOLD search |  |  |
| Search DB | The database searched from BOLD |  |  |
| Top % | The highest match from the BOLD database | Percent of DNA sequence match |  |
| Low % | The lowest match from the BOLD database | Percent of DNA sequence match |  |
| X | The highest match for a second assignment from BOLD | Percent of DNA sequence match |  |
| X1 | The highest match for a second assignment from BOLD | Percent of DNA sequence match |  |
| Type | Type of DNA | Blank = predator or prey  Non-diet = definitely not a diet item (no assignment or fungi)  Unclear = assigned to different things between BOLD and GenBank search |  |

**Table name:** ncbi.csv

**Table description:** taxonomic assignments from GenBank database of all the ASVs in the DADA2 dataset. These were exported from the .rma6 file, which can be opened in MEGAN to view a taxonomic tree of these data.

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Description | Unit or  code explanation or date format | Missing value code |
| ASV | A CO1 sequence merged and filtered via DADA2 | A taxonomic unit from dada2 |  |
| Domain | The domain of the taxonomic assignment |  |  |
| Phylum | The phylum of the taxonomic assignment |  |  |
| Class | The class of the taxonomic assignment |  |  |
| Order | The order of the taxonomic assignment |  |  |
| Family | The family of the taxonomic assignment |  |  |
| Genus | The genus of the taxonomic assignment |  |  |
| Species | The species of the taxonomic assignment |  |  |

**Table name:** Pal\_UG\_mass\_length.csv

**Table description:** Mass and length data collected by Miller-ter Kuile from 2010-2015

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Description | Unit or  code explanation or date format | Missing value code |
| Island | The islet site from Palmyra Atoll where the predator was collected |  |  |
| Date | The date the sample was collected | Either a year value or a date in D/M/Y format |  |
| Tree | A tree species ID if these data were collected | PS: Phymatosorus scolopendria  PG: Pisonia grandis  TA: Tournefortia argentea  Scae: Scaevola taccada |  |
| Number | A sample ID number |  |  |
| Weight\_mg | The weight of the predator | milligrams |  |
| Length\_mm | The length of the predator | Millimeters |  |
| Order | The order of the predator sample |  |  |
| Family | The family of the predator sample |  |  |
| Genus | The genus of the predator sample |  |  |
| Species | The species of the predator sample |  |  |

**Table name:** Predator\_IDs.csv

**Table description:** taxonomic information as well as trait information for the predator species in our datasets

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Description | Unit or  code explanation or date format | Missing value code |
| Pred\_Class | The taxonomic class of the predator species |  |  |
| Pred\_Order | The taxonomic class of the predator species |  |  |
| Pred\_Family | The taxonomic class of the predator species |  |  |
| Pred\_Genus | The taxonomic class of the predator species |  |  |
| Pred\_Species | The taxonomic class of the predator species |  |  |
| Pred\_ID | The ID used colloquially to describe the predators |  |  |
| Sample\_str | The three-letter ID used to create sample names in each predator species |  |  |
| Hunting\_mode | A binary active/not active hunting mode strategy |  |  |
| venom | A yes/no of whether the predator uses venom |  |  |
| webs | A yes/no of whether the predator uses webs to catch or subdue prey |  |  |

**Table name:** Sample\_metadata.csv

**Table description:** collection information, identification, extraction, size, and isotope information for each predator sample

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Description | Unit or  code explanation or date format | Missing value code |
| Method | Method used to collect the sample (all using sterilized implements) | Hand: collected by hand  Fogging: collected via canopy insecticide fog  Net: insect net  BC: Collected off of a branch |  |
| Island | The islet site ID where the sample was collected |  |  |
| Habitat | The tree or vegetation type on or under which the sample was collected | PG: Pisonia grandis  CN: Cocos nucifera  TC: Terminalia catalpa  PF: Pandanus fischeranus  TA: Tournefortia argentea |  |
| Microhabitat | The type of habitat substrate in which the sample was collected | Soil: in the soil  Canopy: in the canopy  Understory: in understory vegetation or on the ground  Open: in the air |  |
| Year | The year the sample was collected | A year |  |
| Order | The order of the predator sample |  |  |
| ID | The species/genus ID for each predator |  |  |
| Date Collected | The date the sample was collected | D/M/YY format |  |
| Extraction ID | An ID attributed during DNA extraction, including multiple individuals from a sampling date/microhabitat sometimes |  |  |
| Extr\_ID2 | An extraction ID with a leading zero |  |  |
| No. Individuals | The number of individuals in each extraction ID sample |  |  |
| Length\_mm | The length of the predator | millimeters |  |
| Sterilized | For another study, whether the sample was surface sterilized with bleach prior to DNA extraction |  |  |
| Source | For another study, whether the sample was kept in a lab environment |  |  |
| Isotope\_ID | For another study, whether the sample was run for isotopes (Nitrogen and Carbon) |  | NA |

## Scripts/code (software)

|  |  |  |
| --- | --- | --- |
| File name | Description | Scripting language |
| Cutadapt\_primer\_trimming.txt | Code to remove primer sequences from the ends of raw sequencing data prior to bioinformatics steps (Category: Bioinformatics) | Bash/command line |
| 1b1\_dada2\_separate\_cross\_run\_error.R | This script runs the DADA2 algorithm on each sequencing run separately up to the step generating the error distributions. This was then used to compare the error structures of each run prior to pooling all samples into one dada2 pipeline combined (Category: Bioinformatics) | R |
| 1b2\_dada2\_combined\_runs.R | This script is the script used to generate the data analysed in the manuscript. This script takes all the samples run across all 4 sequencing runs and processes them through the dada2 filtering, merging, and cleaining pipeline (Category: Bioinformatics) | R |
| Unoise\_code | This code was used to merge, clean, and filter sequences through the UNOISE3 algorithm. Once we compared these data to DADA2 – we found that DADA2 generated more sequences per sample and so used DADA2-generated data instead of UNOISE3 data (Category: Bioinfomatics) | Bash/command line and the free version of usearch (32-bit) |
| Knot\_blast\_bash\_script | This was the code submitted to the super computing cluster at UCSB to compare the DADA2 ASV sequences to the NCBI nucleotide database, which was downloaded locally onto our user account on the super computer (Category: Bioinformatics) | Bash/command line |
| 2\_data\_prep.Rmd | This R markdown runs through all the steps used to clean and combine datasets prior to analyses. Each step is explained in the code, along with the generated dataset. This markdown sources nine R scripts used to generate each step in the process (Category: Data Prep) | R and R markdown |
| 2a\_remove\_negatives.R | This code examines the number of sequences attributed to negative control samples in our DNA sequencing data and removes that number of sequences across samples to correct for sequencing jumping (Category: Data Prep) | R |
| 2b\_low\_sampling\_depth\_cutoff.R | One species of predator and several samples of other species generated very low sample sequencing depths, likely due to inhibitors that were not detected during amplification steps. We generated an inflection curve to decide which samples had too-low sequencing and were removed from further analyses (Category: Data Prep) | R |
| 2c\_master\_taxonomy\_lists.R | We compared the ASVs generated via DADA2 to both NCBI GenBank nucleotide database and the BOLD database. This script combines the taxonomies from both of these sources and attributes them to either likely prey items, or things that are not prey (this included mostly fungal sequences) (Category: Data Prep) | R |
| 2d\_rarefy\_samples.R | Our samples had different sequencing depths and so we rarefied them to a consistent sampling level based on the sample that received the lowest sequencing depth (after the cutoff in step 2b above) (Category: Data Prep) | R |
| 2e\_subset\_diet\_taxonomies.R | We combined the taxonomic lists with the rarefied per-sample data in this step and then kept only those taxonomies that had been assigned to diet items for further steps (Category: Data Prep) | R |
| 2f\_palmyra\_bs\_cleaning.R | This code takes data that was shared with collaborators from the Palmyra terrestrial community and cleans up naming inconsistencies and data entry errors for both body sizes per species and the names of all the species “nodes” (Category: Data Prep) | R |
| 2g\_master\_body\_size\_list.R | This script combines all the body size data from Miller-ter Kuile et al, unpublished Palmrya community data, and published sources and combines them into one master list of body size data per species/family for the Palmyra community (Category: Data Prep) | R |
| 2h\_predator\_mass\_length\_models.R | WE only measured lengths for our predator samples and so we generated mass-length models per each species using the lists generated from the master body size list (step 2g) so that we could use mass in our analyses instead of length (Category: Data Prep) | R |
| 2i\_interactions\_w\_body\_size.R | This is the code that generates the final dataset used in statistical analyses in our study. It includes the taxonomic identity of every predator-prey pair along with the body masses of both predators and prey and the predator traits used in the trait models (Category: Data Prep) | R |
| 3a\_dada\_cross\_run\_error.R | This is the QC code to compare errors across sequencing runs for DADA2 prior to pooling them all into one DADA2 pipeline (Category: QC) | R |
| 3b\_seq\_depth\_dadavsunoise.R | This is the script comparing the per-sample raw sequencing depth per sample for both dada2 and unoise3 – showing that DADA2 generated more sequences per sample Category: QC) | R |
| 3c\_cross\_run\_sample\_comparisons.R | We sequenced 19 predator DNA samples across all 4 sequencing runs and then compared the identity and assignment of ASVs to these samples across sequencing runs to determine whether different runs may have resulted in different prey communities Category: QC) | R |
| 3d\_multiple\_individuals.R | A script comparing samples that comprise only one individual to those for which we pooled multiple individuals of the same size from the same sampling site and time Category: QC) | R |
| 3e\_control\_samples.R | A script assessing the sequencing specificity of the positive control samples and the lack of sequences in negative control samples Category: QC) | R |
| 4\_source.R | A script that loads datasets and packages needed for statistical analyses and which tidies the data for the statistical analyses (Category: Data Analyses) | R |
| 4a\_pred\_prey\_size.R | A script of the statistical analyses of predator-prey size scaling relationships (Category: Data Analyses) | R |
| 4b\_feeding\_mode\_ratios.R | A script of the statistical analyses of predator:prey size ratio based on taxonomic identity and predator hunting traits (Category: Data Analyses) | R |
| 5a\_pred\_size\_histograms.R | A script generating a histogram figure of predator sample size distributions by species (Category: Figures) | R |
| 5b\_community\_size\_dists.R | A script generating a histogram figure of community, prey community, and predator sample size distributions (Category: Figures) | R |
| 5c\_bipartite.R | A script generating a bipartite graph of predator-prey interactions (Category: Figures) | R |
| 5d\_pred-prey\_size.R | A script generating predator-prey size dotplots and boxplots (Category: Figures) | R |
| 5e\_pred\_prey\_ratios.R | A script generating predator:prey size ratio boxplots (Category: Figures) | R |
| Body\_size\_summaries.R | A script generating tables and supplementary figures of size distributions (Category: Figures) | R |
| Brose\_data\_summaries.R | A script generating summary graphs of the Brose et al. 2019 dataset from the predatory interactions from 22 invertebrate food webs (Category: Figures) | R |
| Data\_summary\_tables | A script generating data summary tables from the supplementary figures (Category: Figures) | R |

## Notes and Comments

## References

Apigo, A., and R. Oono. 2018. MG840195 and MG840196.

Barton, K. 2020. MuMIn: Multi-Model Inference.

Brose, U., P. Archambault, A. D. Barnes, L. F. Bersier, T. Boy, J. Canning-Clode, E. Conti, M. Dias, C. Digel, A. Dissanayake, A. A. V. Flores, K. Fussmann, B. Gauzens, C. Gray, J. Häussler, M. R. Hirt, U. Jacob, M. Jochum, S. Kéfi, O. McLaughlin, M. M. MacPherson, E. Latz, K. Layer-Dobra, P. Legagneux, Y. Li, C. Madeira, N. D. Martinez, V. Mendonça, C. Mulder, S. A. Navarrete, E. J. O’Gorman, D. Ott, J. Paula, D. Perkins, D. Piechnik, I. Pokrovsky, D. Raffaelli, B. C. Rall, B. Rosenbaum, R. Ryser, A. Silva, E. H. Sohlström, N. Sokolova, M. S. A. Thompson, R. M. Thompson, F. Vermandele, C. Vinagre, S. Wang, J. M. Wefer, R. J. Williams, E. Wieters, G. Woodward, and A. C. Iles. 2019. Predator traits determine food-web architecture across ecosystems. Nature Ecology and Evolution 3:919–927.

Bui, A., H. Lowman, A. S. Guerra, and A. Miller-ter Kuile. 2020. calecopal: A California-inspired Package of Color Palettes.

Callahan, B. J., P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13:581–583.

Clark, K., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. 2016. GenBank. Nucleic Acids Research 44:D67–D72.

Edgar, R. C. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv:81257.

Eitzinger, B., N. Abrego, D. Gravel, T. Huotari, E. J. Vesterinen, and T. Roslin. 2019. Assessing changes in arthropod predator–prey interactions through DNA-based gut content analysis—variable environment, stable diet. Molecular Ecology 28:266–280.

Elbrecht, V., B. Peinert, and F. Leese. 2017. Sorting things out: Assessing effects of unequal specimen biomass on DNA metabarcoding. Ecology and Evolution 7:6918–6926.

Fulton, T. M., J. Chunwongse, and S. D. Tanksley. 1995. Microprep protocol for extraction of DNA from tomato and other herbaceous plants. Plant Molecular Biology Reporter 13:207–209.

Greenstone, M. H., D. C. Weber, T. C. Coudron, and M. E. Payton. 2011. Unnecessary roughness? Testing the hypothesis that predators destined for molecular gut-content analysis must be hand-collected to avoid cross-contamination. Molecular Ecology Resources 11:286–293.

Handler, A., D. Gruner, W. Haines, M. Lange, and K. Kaneshiro. 2007. Arthropod surveys on Palmyra Atoll, Line Islands, and insights into the decline of the native tree Pisonia grandis (Nyctaginaceae). Pacific Science 61:485–502.

Hartig, F. 2020. DHARMa.

Hsieh, T. C., and A. Chao. 2017. Rarefaction and extrapolation: Making fair comparison of abundance-sensitive phylogenetic diversity among multiple assemblages. Systematic Biology 66:100–111.

Hsieh, T. C., K. H. Ma, and A. Chao. 2016. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). Methods in Ecology and Evolution 7:1451–1456.

Huson, D. H., S. Beier, I. Flade, A. Górska, M. El-Hadidi, S. Mitra, H. J. Ruscheweyh, and R. Tappu. 2016. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Computational Biology 12:1–12.

Illumina. 2009. Illumina adapter sequences. Page Illumina.

Kartzinel, T. R., P. A. Chen, T. C. Coverdale, D. L. Erickson, W. J. Kress, M. L. Kuzmina, D. I. Rubenstein, W. Wang, and R. M. Pringle. 2015. DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. Proceedings of the National Academy of Sciences 112:8019–8024.

Krehenwinkel, H., S. Kennedy, S. Pekár, and R. G. Gillespie. 2017. A cost‐efficient and simple protocol to enrich prey DNA from extractions of predatory arthropods for large‐scale gut content analysis by Illumina sequencing. Methods in Ecology and Evolution 8:126–134.

Leray, M., J. Y. Yang, C. P. Meyer, S. C. Mills, N. Agudelo, V. Ranwez, J. T. Boehm, and R. J. Machida. 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 10:1–14.

Macías-Hernández, N., K. Athey, V. Tonzo, O. S. Wangensteen, M. Arnedo, and J. Harwood. 2018. Molecular gut content analysis of different spider body parts. PLoS ONE 13:1–16.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBNet Journal 17:10–12.

McKnight, D. T., R. Huerlimann, D. S. Bower, L. Schwarzkopf, R. A. Alford, and K. R. Zenger. 2019. Methods for normalizing microbiome data: An ecological perspective. Methods in Ecology and Evolution 10:389–400.

Oono, R., D. Black, E. Slessarev, B. Sickler, A. Strom, and A. Apigo. 2020. Species diversity of fungal endophytes across a stress gradient for plants. New Phytologist 228:210–225.

Porter, T. M., and M. Hajibabaei. 2018. Over 2.5 million COI sequences in GenBank and growing. PLoS ONE 13:1–16.

Sohlström, E. H., L. Marian, A. D. Barnes, N. F. Haneda, S. Scheu, B. C. Rall, U. Brose, and M. Jochum. 2018. Applying generalized allometric regressions to predict live body mass of tropical and temperate arthropods. Ecology and Evolution 8:12737–12749.

Su, G., R. Dudley, T. Pan, M. Zheng, L. Peng, and Q. Li. 2020. Maximum aerodynamic force production by the wandering glider dragonfly (Pantala flavescens, Libellulidae). The Journal of experimental biology 223.

Toju, H., A. S. Tanabe, S. Yamamoto, and H. Sato. 2012. High-coverage ITS primers for the DNA-based identification of ascomycetes and basidiomycetes in environmental samples. PLoS ONE 7.

van der Valk, T., F. Vezzi, M. Ormestad, L. Dalén, and K. Guschanski. 2020. Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. Molecular Ecology Resources 20:1171–1181.

Yaninek, J. S., and D. Gnanvossou. 1993. Fresh and dry wei ghts of Mononychellus tanajoa (Acari: Tetranychidae): A functional description of biomass accumulation. Experimental and Applied Acarology 17:775–779.

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