## 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

**(These are the people who will show up as authors in the dataset citation.** These are the individuals who have provided intellectual or other significant contributions to the creation of this dataset, much like the authors of a research paper.)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| First Name | Middle Initial | Last Name | Organization | e-mail address | ORCID ID (optional) |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

## Other personnel names and roles

(Who should a data user contact with questions about these data? You **must** enter a person or organization name to serve as the **contact** for this dataset. You may also list other personnel who participated in the project (such as field crew, lab tech, data entry etc.) in this table with optional fields e-mail addresses, organization and ORCID ID.)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| First Name | Middle Initial | Last Name | Organization | e-mail address | ORCID ID (optional) | Role in project |
|  |  |  |  |  |  | Contact |
|  |  |  |  |  |  |  |

## Keywords

(**List keywords below and separate with commas.** Using keywords from a controlled vocabulary (CV) will improve the future discovery and reuse of your data. The LTER CV is a good source for keywords. [**Access the LTER CV here**](http://vocab.lternet.edu/vocab/vocab/index.php). Also, please determine one or two keywords that best describe your lab, station, and/or project (e.g., Trout Lake Station, NTL LTER).)

## Funding of this work:

List only the **main PI of a grant** that supported this project, starting with the main grant first. Add rows to the table if several grants were involved.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| PI First Name | PI Middle Initial | PI Last Name | PI ORCID ID (optional) | Title of Grant | Funding Agency | Funding Identification Number |
|  |  |  |  |  |  |  |

## Timeframe

* Begin date: July 2015
* End date: May 2021
* Data collection ongoing/completed:

## Geographic location

(Use **decimal degrees** to define a point or a bounding box. Use a negative symbol (-) to indicate a west longitude. Copy this block to add multiple points or areas.)

* 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;

*Opopaea sp*. (Family: Oonopidae)

*Neoscona theisi* (Family: Aranaiedae),

*Heteropoda venatoria* (Family: Sparassidae)

*Smeringopus pallidus* (Family: Pholcidae)

*Scytodes longipes* (Family: Scytodidae)

Phylum: Arthropoda, Class: Insecta;

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

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

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

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

*Mecistocephalus sp*.

## Methods

(Be specific about the study design and field and lab methods for collecting and processing the data. Include instrument descriptions and protocol citations.)

## Data Provenance

(Were these data derived from other data? If so, you will want to document this information so users know where these data came from. Please specify the source datasets used in the below **provenance table**, preferably with their DOI or URL. An example of a dataset derived from several others is [here](https://portal.edirepository.org/nis/mapbrowse?packageid=edi.101.3).)

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset title | Dataset DOI or URL | Creator (name & email) | Contact (name & email) |
|  |  |  |  |

## Data Table

(Provide a Table Name and Table Description. Each row in the below table describes one column in your data table. Complete each row as follows:

* **Column name**: This name must be exactly as it appears in the dataset. Please avoid special characters (like & or \), dashes and spaces. Underscores are permissible. Do not begin a column name with a number.
* **Description**: Please give a specific definition of the column name. This can be lengthy.
* **Unit:** Identify units for all numeric variables. Please avoid special characters and describe units in this pattern: e.g. microSiemenPerCentimeter, microgramsPerLiter, absorptionPerMolePerCentimeter
* **Code explanation**: If you use codes in your column, please explain in this way: e.g., LR=Little Rock Lake, A=Sample suspect, J=Nonstandard routine followed
* **Date format**: Please tell us exactly how the date and time is formatted: e.g. mm/dd/yyyy hh:mm:ss plus the time zone and whether or not daylight savings was observed. ISO date format of YYYY-MM-DD or YYYY-MM-DD hh:mm:ss is preferred.
* **Missing value code**: If a code for ‘no data’ is used, please specify: e.g., -99999

**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 |

## Spatial data objects

(List any geospatial data objects you would like to archive. Organize spatial data into .zip directories

and describe each.)

**Directory name:** (A short name for the data)

**Directory description:** (A brief description of the data)

|  |  |
| --- | --- |
| Attribute | Value |
| Horizontal Coordinate System Name (e.g. WGS\_1984\_UTM\_Zone\_12N) |  |
| Horizontal Accuracy Report |  |
| Vertical Accuracy Report |  |
| Cell Size X Direction |  |
| Cell Size Y Direction |  |
| Raster Origin (e.g. Upper Left) |  |
| Number of Rows |  |
| Number of Columns |  |
| Number of Verticals |  |
| Cell Geometry (e.g. pixel) |  |

## Scripts/code (software)

(List any software scripts/code you would like to archive along with your data. These may include processing scripts you wrote to create, clean, or analyze the data.)

|  |  |  |
| --- | --- | --- |
| File name | Description | Scripting language |
|  |  |  |

## Articles

(List articles citing this dataset)

|  |  |  |
| --- | --- | --- |
| Article DOI or URL (DOI is preferred) | Article title | Journal title |
|  |  |  |

## Notes and Comments