Working Author list: Ana Miller-ter Kuile, Austen Apigo, Bart DiFiore, An Bui, Daniel Preston, Elizabeth Forbes, Carina Motta, Michelle Lee, John McLaughlin, Kevin Lafferty, Hillary Young

Field: Jasmine Childress, Katie Plummer, Rodolfo Dirzo, David Weber, Maggie Klope, Cora Johnston, Taylor Bogar, Magalay Espinoza

Lab: Ronny Young – others?

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 characterization of potential diet items (Handler et al., 2007). Predator individuals were collected across habitat types on the atoll, including different forest types and microhabitats, including understory vegetation, canopy vegetation, and soil habitats (supplementary summary table of habitats). For each of these habitats, we used a combination of methods, including individual collection during visual surveys for understory and soil collections and canopy fogging with insecticide (cite) onto collection sheets for canopy individuals. All individuals were collected individually with sterilized implements (ethanol-burned forceps) in sterilized collection containers containing 85% EtOH to avoid contamination (Greenstone et al., 2011). We identified all predators to morphospecies in a laboratory environment and the predator species used in this study represent the most common predator species found in each habitat type. All individuals were stored in a -20ºC freezer before DNA extraction.

*DNA extraction and removal of consumer DNA with Ampure XP beads*

We individually measured each individual predator length in a sterilized laminar flow hood with sterilized implements and sterilized surfaces, after which we removed body parts other than the abdomen and thorax (for Insecta predators), the oposthosoma (for Arachnida predators), or the trunk (for Chilopoda predators). We extracted DNA from each predator following a modified CTAB extraction protocol (Fulton et al., 1995). At least twenty-four hours following extraction, we quantified DNA using a Qubit (Invitrogen) fluorometer with the high sensitivity double-stranded DNA quantification kit using 1uL of DNA template per reaction. We used methods developed by Krehenwinkel et al., (2017) to isolate a proportion of lower molecular weight consumer or diet DNA prior to PCR steps with Ampure XP beads. Ampure XP beads preferentially bind to heavier molecules of more intact consumer DNA, leaving the smaller fragments of presumed semi-digested diet DNA in the supernatant. Thus, by keeping the supernatant, we aimed to work with a sample that had a larger proportion of lower molecular weight diet DNA after removing consumer DNA that bound to beads (Krehenwinkel et al., 2017). To do this, we diluted each DNA sample to 20ng/ul (creating a total sample volume of 40uL), mixed each sample using Ampure XP beads (0.75x bead-to-DNA ratio), and kept the supernatant from this step. With the supernatant, we repeated the CTAB protocol steps for precipitating DNA pellets with isopropanol and 5M potassium acetate and cleaned DNA pellets with ethanol washes. After at least another twenty-four hours, we quantified DNA again using a Qubit fluorometer (following the same methods as above) and diluted all samples to 10ng/uL prior to PCR steps. All DNA pellets were stored in and diluted with TE buffer.

*PCR amplification, library preparation, and sequencing*

We amplified the CO1 gene with general metazoan primers (Krehenwinkel et al., 2017; Leray et al., 2013; Yu et al., 2012). The CO1 gene is well-represented in the GenBank sequencing database (Porter & Hajibabaei, 2018). We performed all PCR preparation steps in a UV-sterilized biosafety cabinet. We used a standard desalted primer set tested by Krehenwinkel et al., (2017) for use in diet analyses of invertebrate predatory consumers, including spiders (Table 2). These primers included overhang adapters compatible with the Illumina indexing PCR (Illumina 2009) that immediately followed CO1 amplification.

We amplified the CO1 gene in each sample by PCR in a 25uL reaction volume that included 9uL nuclease free water, 12.5uL GoTaq Green Master Mix (Promega Corp.), 1.25 uL of each of the primers (at 10mM), and 1 uL of DNA template (at 10ng/uL). When DNA concentrations were lower than 10ng/uL, we added more DNA to the sample to equal 10ng of total template and reduced the amount of water added. Each sample was run in duplicate until after Illumina indexing PCR, and we ran a duplicated negative sample each PCR run. We ran each reaction with an initial denaturation step at 95°C for 3 minutes, and then 35 cycles of: 1) denaturation at 95°C for 30 seconds, 2) annealing at 46°C for 30 seconds, and 3) elongation at 72°C for one minute. We ended each PCR run with a final elongation step at 72°C for 5 minutes and then held samples at 4°C until placed in a 4°C refrigerator. To remove reaction dimer before attaching Illumina P5/P7 indices, we removed lower molecular weight amplicons (~200 bp) with Ampure XP beads at a 0.8x bead-to-DNA ratio. Samples were re-suspended from beads using a 10mM TRIS resuspension buffer.

We then attached Illumina index primers with an additional PCR step (Nextera XT Index Kit v2). Each total reaction volume was again 25uL, with 5uL of nuclease free water, 12.5uL GoTaq Green Master Mix, 1.25uL of each primer (at 10mM), and 5uL of PCR product. These were run in a standard PCR protocol for these primers: an initial denaturing step at 95°C for 3 minutes, followed by 10 cycles of: 1) denaturation at 95°C for 30 seconds, 2) annealing at 55°C for 30 seconds, and 3) elongation at 72°C for 30 seconds. We ended each run with a final elongation step at 72°C for 5 minutes and then held samples at 4°C until placed in a 4°C refrigerator.

We verified PCR amplification by visualizing 3-4 uL of each PCR product in a 1.5% agarose gel using GelRed (Biotium) at 100V for 30-40 minutes. Gels were visualized with a Bio-Rad Gel Doc XR+ imager using Image Lab 5.0. We kept samples for which both duplicates successfully amplified during the PCR steps and only kept samples on PCR runs in which both negative control duplicates resulted in no product detection by gel electrophoresis. For successful samples, we combined duplicates and mixed with an Ampure XP bead-to-DNA ratio of 0.7x. We determined the average length of the gene region using an Agilent TapeStation with a D1000 ScreenTape System following the standard protocol from the quick start guide. We then quantified these final PCR products using a Qubit fluorometer and a high sensitivity kit with 1uL of sample per reaction tube and diluted each sample in 10mM TRIS to a final concentration of 5nM.

Because of the sample size and the need for a large number of sequences per predator in order to detect rarer prey DNA (Krehenwinkel, miller-ter Kuile, supplementary), 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 (TABLE). We ran 19 samples of one predator species (*H. venatoria*) across all runs to quantify run-to-run variation in sequencing (Supplement repeated measures ANOVA). 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; Apigo & Oono, 2018; Clark et al., 2016; Toju et al., 2012). 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. Our full protocol from DNA extraction through submission for Illumina sequencing can be found in Miller-ter Kuile et al..

*Sequence merging, filtering, and clustering with DADA2*

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). This ASV denoising approach incorporates sequence abundance, quality, and error rates to cluster reads in high throughput sequencing data into a smaller subset of biological units. 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 CITE EDGAR), but found that DADA2 gave more high-read abundance ASVs and so chose to perform all analyses using DADA2 data (Supplementary Figure). We ran DADA2 on data 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 (Appendix Figure CITE).

*Verifying sequencing depth*

Using the output of the number of filtered and denoised sequences across ASVs from DADA2, we determined that each sample in our dataset had been sequenced with sufficient depth. To do this, we determined each sample’s sequencing depth using the iNEXT function in the iNEXT package in R (CITE). We observed that some samples had low sequencing depth (< 10,000 compared to 100,000+ in other samples) and chose to remove these samples from analyses. To determine a cutoff for removing these low-sequence samples, we determined the lower quantiles from 0.05 - 0.2 with increments of 0.01 in our dataset (using the quantile() function in the stats package) and determined the inflection point in these lower quantiles after which sequencing read abundance per sample increased the most between the two sets of samples.

*ASV taxonomic assignment with BLAST and BOLD*

From the output of the DADA2 algorithm, we created a list of unique ASVs and a matrix of ASV abundances across samples. We matched ASVs 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).

*Rarefying samples*

We rarefied samples (McKnight et al., 2019) because sequencing depth, or the total number of DNA sequences assigned per sample, can vary considerably in high throughput sequencing runs. We rarefied 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 & Chao, 2017). We rarefied using the rrarefy() function in the vegan (version 2.5.6) package in R. We rarefied to 15,954 reads per sample.

*Prey taxonomic assignment*

Following rarefying, we selected all ASVs in each predator sample which matched to all potential diet items from the BLAST and BOLD outputs (Kingdom: Animalia; Clade: Bilateria, excluding consumer DNA). In addition, for all consumers, because BLAST and BOLD matched multiple ASVs to the same taxonomies (e.g. at species, genus, or family level), we concatenated all ASVs based on shared taxonomic assignment. Most (%) ASVs received a family-level taxonomic assignment, so we chose to combine taxonomies at the family level, similar to both metabarcoding and microhistological methods in this field (e.g. Brose et al. 2019, Kartzinel – find other metabarcoding ones). We combined by taxonomies by combining ASVs which matched at the family level into one combined taxonomy with cumulative read abundance (i.e. all ASVs matched to *diet family A* were combined into one *diet family A* taxonomy with cumulative read abundance). All DNA matching any predator species on an individual run at the family or order level was also removed to account for potential sequence jumping within sequencing runs which could alter prey identity or diversity in favor of predator species on a shared run (CITE SEQUENCE JUMPING).

*Predator mass-length relationships*

We converted predator lengths to mass by determining mass-length scaling relationships for each predator species. To do this, we compiled data from the literature (Soehlstrom et al., Su et al., Yaninek et al. 1993), from another ongoing project (Palmyra citation) and from data collected by A. Miller-ter Kuile during other field seasons (Supplementary info/datasets) on body length and mass for individuals of all predator species, or of predators from tropical environments in the same families (e.g. Sohlstrom et al. ). We determined a by-species scaling relationship by fitting a mixed effect model predicting the log10 transformed predator mass from log10 transformed predator length with a random slope term for log10 predator length and a random intercept by predator species (e.g. letting the mass-length relationship vary by species; Supplementary Information).

*Prey size dataset*

Following methods for other interaction datasets of invertebrate predators of similar size (e.g. Brose et al. 2019), we determined prey mass based on average prey masses at the family level drawn from the same sources as the predator mass and length relationships (Sohlstrom et al., Su. Et al, Yananik et al. 1993, palmyra citation, Miller-ter Kuile).

*Prey Size:*

*Predator-prey body size relationships*

*Prey size range by predator size*

*Prey Species Composition:*

*Prey identity by predator identity and size*

*Nestedness of prey identity by predator size*

*Stage structure by predator size*

Results

*PCR success, sequence merging, filtering, and clustering with DADA2*

Of a total of 280 samples, we successfully extracted DNA from 99% of samples (n = 278 of 280). Amplification success across all samples was 80%, with 222 of 278 initially extracted samples successfully amplified and sequenced. The Illumina MiSeq runs yielded 3.9, 3.8, 3.8, and 3.3 x 107 unpaired reads that passed filtering and had Q30 quality scores of 87.82, 85.46, 79.34, and 78.02%, respectively. After quality filtering and denoising with DADA2, 3.02 x 107 paired-end reads corresponded to 1,738 ASVs. Following filtering and clustering, we determined that 40 samples had too-low sequencing depths (less than 11,211 reads total) and so they were removed from further analyses.

*ASV taxonomic assignment and Prey taxonomic assignment*

Sixty-seven percent (1,167 of 1,738) of ASVs matched to a taxonomic assignment at Class level or lower. Sixty-seven percent of these taxonomic assignments were assigned to potential prey taxonomies (n = 784), with sixty-seven percent (n=524) of these taxonomies assigned at the family level and so were used in analyses (30% of the total 1,738 denoised ASVs). There were two conflicting taxonomic assignments at the family level or higher between the BOLD and BLAST assignments which were removed from the final total above.

*Quantifying cross-run variation*

There was significant cross-run variation, with significant differences (pair-wise differences between runs with p-value ≤ 0.05) between run 1-4, 2-3, and 2-4. On average, samples had: A: 2.26 ± 0.15, B: 2.33 ± 0.24, C: 1.72 ± 0.23, and D: 1.44 ± 0.15 species in each sample. Because each species was run on a sequencing run with all other individuals from that species and because we did not compare species richness as a response variable across predator species in this study, we report this as the variation across sequencing runs but do not correct for it in future analyses (Supplementary Information).

*Predator mass-length relationships*

We compiled data from 2,362 individual predator sizes (13 – 1652 per predator species). These data had strong species-specific mass-length scaling relationships (R2m = 0.62, R2c = 0.95). We predicted predator mass from these relationships (Supplementary Figure).

*Prey size dataset*

We compiled data from 3,458 individual prey (1 – 1721 individuals per predator family), with 59% (2,028 out of 3,458) coming from individuals from Palmyra Atoll (58% of these, or 1,172, from data collected by A. Miller-ter Kuile during additional field seasons and 42%, or 856, from data collected for CITE PALMYRA PAPER).

*Prey Size:*

*Predator-prey body size relationships*

*Prey size range by predator size*

*Prey Species Composition:*

*Prey identity by predator identity and size*

*Nestedness of prey identity by predator size*

*Stage structure by predator size*