**Methods for surface sterilization of organisms before DNA metabarcoding diet data for species interactions in food webs**

**Introduction**

Baby: (complexity-stability theory foundation) We are trying to understand complex ecosystems (i.e. food web interactions and functions, etc) and their relationship to important ecological attributes, such as stability.

Werewolf: We have a cool tool (metabarcoding) that could give us data on interactions that we’ve never had before that could address these challenges/questions. While they are becoming more common, there is no one systematic way to treat these data, and so it is hard to use them in a food web context for these pressing challenges/questions.

Specifically: For many communities where metabarcoding could provide the most new information (ie. Invertebrate-dominated food webs where traditional diet analyses vs. gut contents is challenging/unfeasible) small size has historically been the limitation to broad-scale diet analyses. Even when metabarcoding, it is sometimes necessary to extract DNA from entire individuals rather than targeting the gut contents of these organisms. This means that DNA contaminants on the outsides of individuals may alter prey diversity by falsely inflating estimated prey diversity (if prey DNA is on the outside surfaces of predators) or by drowning out prey DNA diversity because contaminant DNA will fill up sequencing runs.

Silver Bullet:

How do we take these data and apply them to food web questions in two common contexts: experimental (i.e. mesocosm) environments and field environments?

1. Experimental: Or, organisms with a known range of diet items in which we are trying to detect known diet.
2. Does metabarcoding detect prey we expect to be there?
3. Does surface sterilization of predators from these environments alter the detection and abundance of known prey?
4. Field: Or, for organisms with unknown diet how do we create metabarcoding methods to detect the most prey which is definitely prey?
5. Big picture: Does sterilization alter this prey detection and diversity as well?
   1. ANALYSES: prey DNA detection, prey species richness, and prey species read abundance (all analyzed by sample)
6. Zooming in: What are the more specific outcomes?
   1. ANALYSES: presence-based Jaccard indices for sterilized/unsterilized.
   2. Consider treating this with both the raw data and then rarefied data, with the rarefied data maybe getting a better glimpse at core prey (ala. Core microbiome) and whether if we subset to this group, whether our inferences of differences by sterilization change. Broader context of trying to isolate signal from a lot of noise – this is common in time series analyses, mixture models, and many statistical techniques in this and other fields. (may need to consider other ways of isolating “core prey” from samples depending on outcomes)

**Methods**

*Field site and collections*

We conducted this work on Palmyra Atoll National Wildlife Refuge, Northern Line Islands (LAT LONG). We targeted an abundant, nocturnal, active hunting spider species (*Heteropoda venatoria*) for this project because this species is fairly large (up to several centimeter body length) and occurs in high abundance on the atoll and is therefore easy to collect and observe. Furthermore, this species was introduced to the island (Handler et al.) and so is not of conservation concern. We collected individuals during two field seasons. In 2017, we collected individuals which we kept in mesocosm environments in the lab (explained below). In 2015, we collected individuals in the field across several habitats on the atoll. All individuals were collected during the night from natural habitats.

*Mesocosm predator set-up and feeding*

In the 2017 field season, we conducted mesocosm feeding trials with field-collected spiders because we were interested in testing whether DNA metabarcoding would detect DNA from diet items we know the spider ingested and because mesocosms are a common method of assigning feeding interactions in food web contexts. We created feeding mesocosms out of yogurt containers with holes for air transfer. We placed an *H. venatoria* individual collected in the field in each of these mesocosms. After a 12-hour period alone in the mesocosms, all *H. venatoria* individuals were fed one individual of a common large grasshopper species (*Oxya japonica*), which is also an introduced species on the island (Handler et al. 2007). We left all mesocosms for 24 hours, after which we immediately froze *H. venatoria* individuals which had killed (and presumably ingested some or all of) an *O. japonica* individual at -20C.

*Field predator collection*

In the 2015 field season, we collected spiders in field environments because we were interested in whether DNA metabarcoding would detect DNA from the diets of predators occurring in natural environments. We froze all individuals at -80C immediately following collection until sterilization and DNA extraction in 2019.

*Both mesocosm and field spiders: surface sterilization*

Because we planned to extract DNA from entire individuals or from the opisthosoma of larger individuals (following methods from Krehenwinkel et al. 2016), we wanted to determine whether contaminant DNA on the outside of *H. venatoria* individuals altered sample (prey) diversity and detection. We used surface sterilization to remove possible contaminants from some individuals while leaving some individuals unsterilized (SAMPLE SIZES HERE). We used surface sterilization techniques common in other fields of molecular ecology (i.e. plant endophytes Schulz et al. 1993, Burgdorf et al. 2014) by submerging and stirring each sample in 10% bleach by volume for 2 minutes and then washing each sample by submerging and stirring in deionized water for 2 minutes. We surface sterilized mesocosm spiders (2017 field season) in the field lab on the atoll following freezing at -20C and then stored each sample in individual vials of 80% ethanol because no long-term refrigeration was available on the atoll at the time. Field-collected spiders (2015 field season) had been frozen at -80C since collection; these spiders were surface sterilized in a sterilized laminar flow hood in 2019 just before DNA extraction. Prior to DNA extraction, all samples from both 2015 and 2017 were allowed to dry for 1-3 hours in a sterilized laminar flow hood.

*DNA Extraction and Removal of Predator DNA with Ampure Beads*

We extracted DNA from each *H. venatoria* individual following a modified CTAB extraction protocol (Fulton et al. 1995). At least twenty-four hours following extraction, we quantified DNA using a Qubit fluorometer and the high sensitivity DNA quantification kit with 1uL of DNA per reaction (CITE OR SOMETHING). Because *H. venatoria* predates species taxonomically similar to itself, this makes choosing PCR primer sets that remove predator and not prey DNA challenging if not impossible. To address this challenge, we used methods developed by Krehenwinkel et al. (2016) to remove a proportion of predator DNA prior to PCR steps. To do this, we diluted each sample to 20ng/ul (a total of 40uL per sample), bead cleaned each sample using Ampure beads and a 0.75x bead ratio and keeping the supernatant from the bead cleaning step. Because Ampure beads preferentially bind to heavier molecules, during this step, more intact predator DNA binds to the beads. Thus, by keeping the supernatant, we aimed to work with a sample that had a larger proportion of prey DNA after removing some predator DNA that bound to beads (Krehenwinkel et al. 2016).

With the supernatant, we repeated the CTAB protocol steps for precipitating DNA pellets with isopropanol and 5M potassium acetate and cleaning DNA pellets with ethanol washes. Waiting at least another twenty-four hours, we quantified DNA again using a Qubit fluorometer and high sensitivity kit with 1uL of DNA per reaction tube 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*

The COI gene is well-represented in the GenBank sequencing database (SUPPLEMENTARY TABLE), so we chose to amplify this gene using general metazoan primers (Krehenwinkel AND CITATIONS FROM THIS PAPER). We performed all PCR amplification steps in a UV-sterilized biosafety cabinet. We used a standard desalted primer set (ordered from Integrated DNA Technologies – give specs here) tested by Krehenwinkel et al. (2016) for use in diet analyses of invertebrate predators (TABLE OF PRIMERS HERE). Our primers included overhang adapters compatible with the Illumina indexing PCR (CITE ILLUMINA HERE).

We ran reactions for PCR amplification with a 25uL reaction volume, including 9uL nuclease free water, 12.5uL GoTaq Green Master Mix (CITE), 1.25 uL of each of the primers, and 1 uL of DNA template (at 10ng/uL). When DNA concentrations were lower than 10ng/uL, we adjusted reactions by adding enough DNA template for a total of 10ng of DNA template and reducing the amount of nuclease free water per reaction. Each sample was run in duplicate until after the Illumina indexing PCR, and then only samples in which both duplicates successfully amplified were pooled for final submission. We ran a duplicated negative sample each PCR run. We ran each reaction with an initialization step at 95C for 3 minutes, and then 35 cycles of a denaturation step at 95C for 30 seconds, an annealing step at 46C for 30 seconds, an elongation step at 72C for one minute. We ended each run with a final elongation step at 72C for 5 minutes and then held samples at 4C until placed in a 4C refrigerator until attaching Illumina indices via an additional PCR step.

Prior to the second PCR step in which Illumina indices are attached to each sample so that they may be multiplexed, we performed a bead cleaning step with Ampure XP beads to remove reaction dimer. This bead cleaning step was performed following standard protocols for these beads (Supplement) and with a bead ratio of 0.8x. Samples were re-suspended from beads using a 10mM TRIS resuspension buffer.

Following bead clean up, we performed a PCR with unique sets of Illumina index primers following standard PCR conditions for these primers (CITE ILLumina). We prepared our samples for high throughput sequencing on an Illumina MiSeq platform using version 2 chemistry for paired end reads (CITE ILLUMINA).

Each total reaction volume was again 25uL, with 5uL of nuclease free water, 12.5uL GoTaq Green Master Mix (CITE), 1.25uL of each primer, and 5uL of PCR product. These were run in a PCR protocol with an initialization step at 95C for 3 minutes, followed by 10 cycles of a denaturation step at 95C for 30 seconds, an annealing step at 55C for 30 seconds, and an elongation step at 72C for 30 seconds. We ended each run with a final elongation step at 72C for 5 minutes and then held samples at 4C until placed in a 4C refrigerator.

Following this PCR step, we visualized 3-4 uL of each PCR product in a 1.5% agarose gel using GelRed (CITE) at 100V and 170mA 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. In each of these visualizations, we also inspected the negative control duplicates and only kept samples from PCR processes which resulted in no product detection in negative controls (NEED TO EXPLAIN THIS BETTER…)

For samples for which both duplicates successfully amplified, we combined both duplicates and bead cleaned these duplicates with an Ampure XP bead 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 (CITE Agilent). We then quantified these final PCR products using a Qubit fluorometer and a high sensitivity kit with 1uL of sample per reaction tube. We then diluted each sample in 10mM TRIS to a final concentration of 5nM, determined by this equation (concentration in ng/uL / (660g/mol x average library size in bp from TapeStation)) x 10^6.

We multiplexed all samples along with at least one negative following dilution. We also ran three positive controls in the sequencing run which were Illumina-indexed sequences of cloned fungal species (FIND FROM AUSTEN) prepared for another project (cite Austen) using the ITS gene region (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 for amplicon sequencing on an Illumina MiSeq platform with version 2 chemistry and 500 cycles and paired end reads. Following sequencing, samples were demultiplexed using Illumina’s bcl2fastq conversion software (v2.20) at the Core facility.

*Sequence merging, filtering, and clustering with USEARCH and DADA22*

We chose to merge, filter, and denoise (cluster) our sequences around exact sequence variants (ASVs) rather than operational taxonomic units (OTUs) using two bioinformatics algorithms: UNOISE3 and DADA2 (unoise3 in USEARCH 32-bit version 11.0.667 CITE and dada2 version 1.14.0 CITE). These two ASV denoising approaches incorporate sequence abundance, quality, and error rates to cluster all the reads in Illumina high throughput sequencing data into a smaller subset of real biological units. Dada2 is an open source program, and a 32-bit version of USEARCH is also open source.

Both dada2 and unoise3 are sensitive to the relative abundances of sequences in a sample. Because we used a general primer for the COI gene to target all metazoans, we expected that our sequence results would be dominated by predator sequences (see similar results in Krehenwinckel et al. 2016). Therefore, we expected this predator dominance could alter how or if ASVs would be assigned to rarer sequences in our samples, even if these represented diet items. Because of this, we performed each algorithm on our full dataset, and then “cleaned” our dataset of predator sequences using BBSplit (CITE) and repeated both algorithms on these cleaned datasets.

Prior to all denoising algorithms, we used cutadapt (CITE version 1.18) to remove primers from each sequence (Supplement). With trimmed sequences, we ran unoise3 in USEARCH, and ran dada2 in R. For both programs, we set a maximum error rate of 1 in the filtering step. From all of these outputs, 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 with the blastn command (CITE) 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 Citation), selecting the subtree with all likely prey items for this species (Kingdom:Anamalia, Clade: Bilateria) and the option provided to export a list of ASVs with their taxonomic assignments (count: summarized, output: readName\_to\_taxonName). For BOLD, we used the BOLD IDEngine of the COI gene with Species Level Barcode Records (accessed February 5-16, 2020; 3,825,490 Sequences, 216,704 Species, and 95,537 Interim Species in database) to match each ASV list to taxonomies. We combined taxonomic assignments from both programs, keeping the highest matching taxonomic match between databases and discarding assignments which did not match across both (following protocol in Elbrecht et al).

After we ran these algorithms on uncleaned sequences, we also prepared a new algorithm in which we “cleaned” the datasets of predator and other non-diet sequences prior to merging and denoising sequences again with dada2 and unoise3. We used BBSplit, a program which is designed to remove contaminants from high throughput sequencing datasets (CITE). We found no differences in key metrics (prey ASV number and prey read abundance) and so did not consider datasets cleaned by BBSplit in subsequent analyses (see Supplement for full Algorithm/BBSplit methods and results).

*Sequencing depth*

We assessed sequencing depth to ensure that we had captured the full diversity of samples using interpolation and extrapolation methods in the iNEXT package in R (versions). We computed a sampling effort curve of the number of species discovered in each sample compared to the number of sequence reads in that sample. We looked at the sampling coverage (SC) value to assess the sampling coverage (WHAT VALUES ARE BEST here?) for each sample. We ran iNEXT separately for DADA2 and UNOISE3 algorithms for all spiders in our dataset (mesocosm and field spiders combined).

*Standards: Sequencing Error Correcting and Specificity*

Prior to analyzing data for prey detection, diversity, and abundance, we needed to correct our data for error that arises during the sequencing process. We based sequencing error for each algorithm on the number of sequences assigned to the negative control in each algorithm. Non-zero values in a negative control either represent contamination from the PCR process or from the sequencing platform. Because we ensured the concentration of our negative control was zero (or near-zero) prior to sequencing submission, we assume that any non-zero values in the negative control likely come from the sequencing platform and represent the rate at which sequences jump from sample to sample across all samples in the run (CITE this phenomenon). Therefore, we used this value to build an error distribution and correct our raw sequencing data prior to community analyses. To do this, we fit the ASV abundances for the negative control to a poisson and negative binomial distribution and used BIC to fit the best of the two distributions to the data (CITE JERDE HERE). We then used the distribution of the best-fitting model to predict whether values of different abundances were likely due to error or not (if significant p-value, i.e. 0.001 or lower, these values correspond to real biological diversity and not to sequencing error). Any ASV in a sample with a read value which was likely due to sequencing error were removed prior to community analyses.

In addition to accounting for and correcting error from the sequencing process, we also used positive controls of fungal clones from each denoising algorithm to assess the specificity of that algorithm. Because these fungal positive controls are clones, they should be assigned to one or very few ASVs in the final dataset.

REST IS WORK IN PROGRESS, SEE MARKDOWN FILE

*Mesocosm Predators, or: do metabarcoding methods detect prey we know to be there?*

1. Known prey detection by sample (ASV presence-absence)
   * 1. Intermediate: rarefy to lowest read abundance
   1. Derivatives manipulating ‘surface sterilization’ term in models of form: glm(detection ~ surface sterilization\*algorithm, family = binomial)
      1. Detection significantly decreases with surface sterilization with the best model including surface sterilization without an interaction with algorithm (i.e. surface sterilization effect does not change by algorithm)
2. Known prey abundance by sample (read abundance)
   1. Derivatives manipulating ‘surface sterilization’ term in models of form: glm(abundance ~ surface sterilization\*algorithm, family = poisson/negative binomial
      1. Abundance decreased marginally with surface sterilization

*Field-collected Predators, or: for predators with unknown diet, how do we ensure that prey detected by metabarcoding methods is, indeed, prey, and how do we maximize the prey diversity, detection, and abundance from our dataset?*

1. Big picture
   * 1. Intermediate: rarefy by lowest sample read first
     2. Intermediate: Concatenate by unique taxonomic assignments first
   1. Prey DNA detection
      1. (glm(Presence ~ Sterilization \*algorithm, family = binomial))
   2. Prey species richness
      1. (glm(SR ~ Sterilization \*algorithm, family = poisson))
   3. Prey ASV read abundance
      1. (glm(reads ~ Sterilization \*algorithm, family = poisson))
2. Zooming in
   1. Jaccard presence-absence community analyses of prey ASVs
      1. Presence-absence PERMANOVA on raw data (glmer/glmmTMB(presence ~ sterilization+algorithm + (1+sterilization | species), family = binomial)
   2. Core prey (ala. Core microbiome) presence-absence analyses
      1. Presence-absence PERMANOVA on rarefied data (glmer/glmmTMB(presence ~ sterilization+algorithm + (1+sterilization | species), family = binomial)

**Results**

QC results for sequencing runs.

For dada2 (CL1, CL4, QC1) all mapped to 1 ASV. For UNOISE, they ampped to:

*Rarity cutoffs (Jerde) and rarifying for community analyses (depending on analyses).*

Assessing sequencing depth with iNEXT

We did an accurate job of sequencing depth across all samples.

For non-predator ASVs, other than two samples, sequencing depth SC was calculated as ≥ 0.98

**Acknowledgements**

Field work for this project was funded by the National Science Foundation (DEB #1457371), National Geographic Society, and a Faculty Research Grant from the UC Santa Barbara Academic Senate. We would like to thank field technicians Carina Motta and Michelle Lee for help collecting samples for this project and laboratory technicians Emily Lutz and Tessa Chou for helping prepare genetic samples. We would like to thank the U.S. Fish and Wildlife Service and Palmyra Atoll Research Consortium for supporting field work for this project. We would like to thank Dr. Ryoko Oono for use of her laboratory space and equipment. PALMYRA ATOLL THANKS HERE. We acknowledge the use of the Biological Nanostructures Laboratory within the California NanoSystems Institute, supported by the University of California (UC) Santa Barbara and the University of California Office of the President. We especially thank Dr. Jennifer Smith, manager of the Biological Nanostructures Laboratory for her assistance in preparing and troubleshooting our samples. We acknowledge the use of computational facilities at the Center for Scientific Computing (CSC), which was purchased with funds from the National Science Foundation (CNS-1725797) and is supported by the California NanoSystems Institute and the Materials Research Science and Engineering Center (MRSEC; NSF DMR 1720256) at UC Santa Barbara. Drs. Chris Jerde, Mike Lee, and Ricardo Ramiro provided invaluable feedback on error statistics and bioinformatics protocols. LAB MEMBERS OR WHOEVER EDITS HERE. We thank XX anonymous reviewers for help revising this manuscript. This is publication number PARC-XXX from the Palmyra Atoll Research Consortium.