**Working Title:**

**Effects of environmental contamination on diet metabarcoding data of invertebrate consumers in mesocosms and natural environments**

**Target Journal: *Molecular Ecology Resources (8000 words excluding references, unclear if this means including components after references)***

**Authors:** Ana Miller-ter Kuile, Austen Apigo, Hillary Young

**Introduction**

Biological communities and ecosystem function are shaped by the interactions between organisms (Brown et al. 2001, Hooper et al. 2005, Delmas et al. 2019, Schleuning et al. 2015). Among the many interaction types (e.g. mutualism, consumption, competition), consumptive interactions (including herbivory, predation, and parasitism) can shape the stability of biologically diverse communities (Ings et al. 2009, Delmas et al. 2019). For decades, methods for determining consumptive interactions through assessing diet contents have emerged and provided key insights into consumptive interactions across ecosystems and consumer groups (Hylsop 1980, Nielson et al. 2018). These have included many developments specific to environments or consumer groups and rely on visual observation of consumption events (live or via recordings) or recognition of diet items in diet contents, either unaided or through the use of microscopy (i.e. Baker et al. 2014, Duffy and Jackson 1986, for a review of methods best suited for different contexts and goals see Birkhofer et al. 2017).

This dependence on visual identification or observation is unfeasible or impossible for many consumer groups; specifically, for consumers that are too small for dissection and food identification, have feeding habits which render food items unidentifiable, have food items that contain few components which pass through digestion in recognizable form, or have cryptic habits that prevent them from being observed in large enough numbers for diet analyses, to name a few (Sheppard and Harwood 2005). Observation-based methods become especially limiting in community studies in which it is necessary to ascertain the diets of multiple consumers and for consumers who feed on many diet items (e.g. Polis 1991). In these cases, the most promising avenue for determining consumptive interactions has been the exploration and expansion of DNA-based diet analyses either through gut or fecal contents, initially through species-specific approaches, and now through high throughput sequencing of the DNA of all species in gut contents (Pompanon et al. 2012, Sheppard and Harwood 2005).

The use of high throughput sequencing methods for molecular gut content analysis allows for the identification of a suite of diet species at once and provides a comprehensive and efficient method for determining intraspecific, intra-population, and interspecific diets (Lucas et al. 2018, Soininen et al. 2015, Quemere et al. 2013). These methods have already illuminated interesting new interactions and ecological trends in a variety of environments (e.g. host-parasitoid: Wirta et al. 2014; plant-herbivore: Kartzinel; host-parasite: Schnell et al. 2012, predator-prey: Toju et al. 2018). As these methods have continued to advance, however, they need to be validated so that the ecological inference made from them is robust. Specifically, for organisms where small body size has limited other diet analysis methods, DNA diet analyses often necessitate the extraction of diet data from full organisms, and so the possibility of surface contamination altering the detection and species composition of presumed diet items is an important consideration.

Surface contamination could alter the results of diet metabarcoding via altering the detection and richness of presumed diet items, either through falsely inflating diet estimates (i.e. contaminants are potential diet items) or by depressing diet estimates (i.e. contaminants are not potential diet items). The effects of surface contamination could vary by environmental, ecological, or methodological aspects of a DNA metabarcoding study (e.g. collection method: Greenstone CITATIONS). Environmentally, in more closed ecosystems (i.e. aquatic pools, mesocosm experiments, soil environments), DNA can persist in the environment for months to years (Strickler et al. 2015, Barnes et al. 2014, Neilson et al. 2007, Carini et al. 2016). Combined with a high likelihood that organisms share contained substrates with each other in these environments, the likelihood of environmental contamination could be quite high. Environments where DNA contaminants may be lower risk are those with larger substrates with high cycling (i.e. marine environments; Collins et al. 2018) or environments where UV light and biological activity degrades DNA on surfaces (e.g. many terrestrial environments; effects in aquatic environments shown in Strickler et al. 2015). In any environment, broadly used DNA metabarcoding methods for determining the diet of multiple consumers which potentially feed on a wide variety of diet items are dominated by the DNA of the consumer (e.g. diet comprise 0.03 – 8.43 percent of all sequencing reads due to the inability to use blocking primers; Krehenwinkel et al. 2016, Piñol et al. 2014). In any of these studies, low abundance potential diet DNA could be even more hidden by high-biomass contaminant DNA (e.g. biomass and metabarcoding analysis in Elbrecht et al. 2017) and so recognizing and mitigating the risk of contamination is paramount.

While it has not been systematically used in diet metabarcoding studies, surface sterilization to reduce potential surface contamination has been utilized across disciplines in both single-interaction and high-throughput sequencing methodologies (e.g. single species: Greenstone et al. sterilization one, Linville et al. 2002; high-throughput: Zimmerman and Vitousek 2012, Burgdorf et al. 2014). While some fields (e.g. fungal endophytes) that use high throughput sequencing to describe community diversity use surface sterilization as standard practice, these methods can vary greatly (e.g. using washes of different disinfectants or sonication; Burgdorf et al. 2014) and often depend on the aspects of the tissues used (e.g. how sensitive they are to sterilization-induced degradation; Hallman et al. 1997). The field of fungal endophyte research has developed informed protocols based on decades of research into best practices and study-specific considerations (Brown et al. 2018). Conversely, the field of diet metabarcoding, particularly when determining diet from full individuals, has not developed a similarly systematic approach (e.g. ethanol washes in Doña et al. 2019, bleach washes in Anslan et al. 2016, no sterilization in Wirta et al. or Jacobsen et al. 2018). This lack of systemic surface sterilization in diet metabarcoding when using full individuals limits the ability to confidently assign DNA sequences to ingested diet items and to help discern study-specific considerations (e.g. likelihood of contamination or sensitivity of different consumer species to sterilization-induced degradation). Indeed, the same methods seem to lead to degradation in one study (Greenstone) while providing a robust dataset in another study (Linville et al 2002), suggesting that a broader analysis of the costs and benefits of surface sterilization as well as limitations and study-specific considerations needs to be undertaken by the field.

In this study, we look at the effects of surface contamination and surface sterilization on our understanding of consumer diets, specifically aiming to determine the costs and benefits of surface sterilization in DNA metabarcoding studies where the DNA of full organisms is used in analyses. We use high throughput sequencing results of the CO1 gene from the full body parts (opisthosomas) of invertebrate consumers (the spider, *Heteropoda venatoria*) from two environments – a “contained” mesocosm environment in which we offered consumers a potential diet item, and a natural field environment in which consumers could feed on available diet items. In each environment, we surface sterilized half of the consumers prior to DNA extraction using a series of washes in a 1:10 dilution of NaOCl and deionized water; we left the other half of consumers unsterilized. Specifically, we ask 1) Does surface sterilization alter the detection or diversity of potential diet items, suggesting that contaminants either hide or inflate diet data, and does this effect vary by environmental context (mesocosm versus field)? 2) Does surface sterilization lead to any negative effects suggesting that sterilization degrades the DNA of potential diet items? Our results suggest minimal to no significant impacts of surface contamination on diet detection or diversity, though in the mesocosm environment, some surface contamination may have influenced potential diet item detection. Furthermore, our results show a lack of negative effects of surface sterilization (e.g. sterilization-induced degradation) for diet DNA metabarcoding datasets. Given these results and the low cost (of time, data, or money) of many surface sterilization protocols, it may be judicious to surface sterilize full consumers prior to DNA extraction for diet metabarcoding.

**Methods (only including statistical analyses here for brevity!).**

*Hypothesis 1: Does surface sterilization alter the detection or diversity of potential diet items?*

For consumers from both mesocosm and field environments, we wanted to know whether surface contamination altered the detection of potential diet items for each consumer (either by increasing detection because of “false” diet detection or by decreasing detection because of abundance of non-diet DNA). Furthermore, for field consumers, we wanted to know whether surface contamination altered the diversity and composition of all potential diet items, thus altering ecological interpretations of diet data. We analyzed consumers from the mesocosm environment separate from the field environment because we wanted to know whether the risk of surface contamination vary by environmental differences related to how likely consumers and potential diet items are to interact without consumption via physical contact or shared surfaces (mesocosm = high risk, field = low risk).

*Detection of potential diet items*

For mesocosm consumers, we focused our detection analysis on the potential diet item we had offered the consumers in the mesocosm environment (*O. japonica*, which all consumers were observed to have killed, but not necessarily ingested). For field consumers, we examined all potential diet items (which could represent either real diet or surface contaminants). Because sequencing depth (total number of DNA sequences assigned) can vary considerably across samples in high throughput sequencing runs, we first rarefied our samples so that we were comparing samples with equal sampling effort (McKnight et al. 2018). We did this using the rrarefy() function in the vegan (version 2.5.6) package in R and we rarefied based on the sample with the lowest sequencing depth. We rarefied the mesocosm dataset separate from the field dataset since these samples had been preserved in different ways and for different times, which can have large effects on DNA extraction outcomes (Murphy et al. 2002). Following rarefying, we selected all ASVs which matched to the known potential diet item for the mesocosm consumers (species: *Oxya japonica*, genus: *Oxya*, family: Acrididae), and all potential diet items for the field collected consumers (Kingdom: Animalia; Clade: Bilateria, excluding consumer DNA). In addition, for field consumers, because BLAST and BOLD matched multiple ASVs to the same species taxonomy, we concatenated all ASVs based on shared taxonomic assignment (i.e. multiple ASVs matched to *diet species A* were combined into one *diet species A* taxonomy with cumulative read abundance). For both mesocosm and field consumers, we assessed per sample detection of potential diet (offered diet item for mesocosm consumers, all potential diet items for field collected consumers) using generalized linear models with potential diet item detection (presence-absence per sample) as the response variable, surface sterilization treatment as a fixed effect, and a binomial distribution.

*Richness and composition of potential diet items in field consumers*

In addition, for the field-collected consumers, we also wanted to know whether surface contamination altered both the species richness of DNA attributed to potential diet items as well as the species composition of these potential diet items. For per sample potential diet richness for field-collected consumers, we assessed per sample potential diet richness using generalized linear models with the number of potential diet species per sample as the response variable, surface sterilization treatment as the fixed effect and a Poisson or negative binomial distribution (to correct for overdispersion when needed). We assessed differences in potential diet species composition between surface sterilized and unsterilized consumers using a presence-absence PERMANOVA model fit with a binomial mixed effects model with surface sterilization treatment as a fixed effect, a random intercept term for potential diet species, and a random slope term for surface sterilization treatment. Incorporating a random intercept term for potential diet species combined with a random slope term for surface sterilization treatment allows the effect of surface sterilization treatment to vary by potential diet species, such that some potential diet species may increase in presence with surface sterilization (i.e. hidden by contaminants), while others may decrease in presence (i.e. potential diet item is a contaminant; Zurr, or other random slopes citation here). (We repeated the field consumer potential diet item PERMANOVA with abundance data and also conducted both presence-absence and abundance based PERMANOVA analyses on all potential diet items (including offered item) for mesocosm predators; Supplement).

*2) Does surface sterilization lead to any negative effects suggesting that sterilization degrades the DNA of potential diet items?*

In addition to determining the effects of surface contamination on the ecological interpretation of DNA metabarcoding diet data, we also wanted to determine whether surface sterilization had any negative effects. Specifically, we wanted to determine if the surface sterilization approach we took degraded DNA, which would alter the interpretation of any detection, abundance, or richness we measured in our samples. To determine whether DNA degradation occurred in both environmental contexts, we expected that

We also wanted to know whether for those consumers for which diet was detected, whether surface sterilization treatment alters the proportional abundance of diet DNA reads. For per sample diet proportional read abundance for both sets of consumers, we assessed per sample proportional abundance of diet in consumers with diet detected using generalized linear models with the number of diet DNA reads per sample as the response variable, surface sterilization treatment as a fixed effect and a Poisson or negative binomial distribution (to correct for overdispersion when needed; because we had rarefied to equal sampling depth, total read abundance is directly proportional to the proportion of total reads attributed to diet DNA). For the mesocosm consumers, we also explored the effect of possible DNA degradation on known diet DNA by also running similar models for the number of consumer sequence reads in each sample and the number of other diet reads (not the fed diet item) in each sample (which we expected to find within the timeframe of our feeding trials based on results from another study; Marcias-Hernandez et al. 2018); if DNA degradation occurs with surface sterilization, we expected to observe a decrease in consumer read abundance and other diet reads (Supplement).

*Model selection*

For all generalized linear models, we performed model selection by comparing the full model (including the fixed effect of surface sterilization treatment) to a null model without this effect. All models were called in the glmmTMB package (version 1.0.0, Brooks et al. 2017) in R (version 3.6.1) We chose the best fitting model based on size corrected AIC values (MuMIn package version 1.43.15). We assessed model fit using diagnostics in the DHARMa package (version 0.2.7), including tests for heteroskedasticity, and for count models (Poisson or negative binomial), zero inflation and overdispersion (Zurr, Bolker citations here). All raw data, data cleaning, and data analyses can be found in the data (CITE data here).