Methods

1. Specimens

Light traps were installed near peanut fields in Worth County (Georgia, USA) and checked regularly. At each inspection, samples were collected in 70% ethanol and stored in a -20C freezer until specimens were sorted and burrower bugs identified with the aid of a stereo-microscope (gross morphology) and molecular methods (see below). Burrower bugs were collected in light-traps during the months of April, May, June, August, September and October in 2020, and July, August, September and October 2021. DNA was extracted from these samples using the DNEasy Blood and Tissue kit (Qiagen) and stored in a -20C freezer.

1. Molecular validation of species identification

Species identification was validated for samples that were also submitted for RAD-seq sequencing. RAD-seq data was aligned against draft genomes that we sequenced for *P. bilineatus* (XXaccessionnumberXX) and *D. lugubris* (XXaccessionnumberXX), using Bowtie2 (k = 20, Langmead & Salzberg, 2012). Using Samtools (Danecek et al., 2021) and custom scripts, results from Bowtie2 were summarized for all samples, and a species identification was assigned to samples according to the best alignment rate (>50%) to one of the reference genomes. In cases where alignment was low (<50%) to either of the reference genomes, these samples were classified as “unknown” and excluded from the dataset.

We also sequenced a portion of the mitochondrial COI gene and built a phylogenetic tree to validate our identification using the method described above (see Supplemental Information). Briefly, thirty samples collected in 2021 were selected according to their classification using the method above: 27 samples identified as *D. lugubris*, 2 samples identified as *P. bilineatus*, and 1 sample identified as “unknown” (i.e. low alignment rate to either of our reference draft genomes). In addition, six additional samples collected in 2022 and identified as *D. lugubris* by Mark Abney (University of Georgia) were also added to generate the first COI sequences ever to be added to the NCBI and BOLD databases. A 418bp long region of the COI gene was amplified using the BF3-BR2 primers, as described in Elbrecht et al. (2019), and submitted for Sanger sequencing (Eurofins Genomics LLC, USA). Sequences were quality-trimmed in Geneious Prime (version 2023.2.1, Build 2023-07-20 ; error probability for quality trimming set to 0.05), and forward and reverse sequences were then assembled. Sequences that failed to assemble were individually inspected for errors, and all salvaged sequences (assembled or not) that also aligned well against a reference COI sequence (acc. KR031248.1) were added back to the dataset. Our “unknown” species sample was discarded due to low-quality sequences, and it was replaced by sample 1930-1 that had already been sequenced, using a different primer set (primers Lep2t1-LepR1, Park et al., 2011) but still targeting a COI region that overlaps with the region targeted by BF3-BR2. For building the phylogenetic tree, we added to our dataset COI sequences representing other, related burrower bug taxa (e.g. *Sehirus*, *Cyrtomenus*, and *Tominotus*), close matches found via Blast (NCBI, nt/nr database; “Cydnidae”), and published *Pangaeus bilineatus* sequences. All sequences in the dataset were then aligned against KR031248.1, trimmed to a common overlapping region, and a phylogenetic tree was built using FastTree within Geneious Prime. Sequences were assigned an id according to the presence of the closest reference sequences within the same clade, thus validating or rejecting the identification obtained previously (via gross morphology or RAD-seq alignment rate to a reference genome). Phylogenetic tree and more detailed methodology are provided in the Supplemental Information.

1. Sampling of plant DNA

Gut content was profiled by PacBio sequencing based on the protocol published by Cooper et al. (2019). Briefly, all DNA samples were subject to amplification of the chloroplast gene *trnF* using primers B49873-e (GGT TCA AGT CCC TCT ATC CC) and A50272-F(ATT TGA ACT GGT GAC ACG AG) (Taberlet et al. 1991). Different combinations of barcoded primers (forward and reverse) were used to allow multiplexing samples for sequencing (Suppl Table X). In addition, ten negative control samples were included to account for any contamination in the reagents. All samples were sent for PacBio sequencing at the Washington State University Laboratory for Biotechnology and Bioanalysis (WA, USA).

1. *trnF* sequence analysis

4.1 Reference database

We built a custom database to define parameters to identify and classify our sequencing data. The entire RefSeq Plastid genomic database (NCBI, release 219) was downloaded, unzipped and concatenated into a single file. For the next steps we used Mothur (version 1.47; Schloss et al., 2009). We first performed an *in silico* PCR to amplify only the region targeted by our *trnF* primers B49873-e and A502720-F. Previously, we had found that most of our PacBio sequences were 400-600bp long (Cooper et al., 2019), but there were cases of sequences smaller than 100bp and greater than 1500bp. In fact, it is known that the gene *trnF* can greatly vary in size because of repeats and pseudogenes (e.g. Vijverberg et al. 1999; Yan et al., 2019 ). Based on this, we decided to take a conservative approach for quality-filtering our data, and excluded from our plastid database any sequence smaller than 100bp, or longer than 10K bp, which were likely non-target amplifications or plastids from unicellular organisms (e.g. algae). In addition, we also excluded all sequences with greater than 5 ambiguities. The median size of sequences passing these filters was 399bp, ranging from 101bp to 9119bp, for a total of 6,988 sequences. Next, we used custom scripts to add and format taxonomical rank information to the header section of each sequence, so that it can be read by the DADA2 classifier.

4.2 PacBio data

All data analyses were performed in R (version 4.2.1) and Rstudio (version 2023.06.1, Rstudio Team, 2023). PacBio data was quality-trimmed, assembled into ASV and classified using DADA2 (version 1.22.0; Callahan et al., 2016). Briefly, each sequencing batch (three total) started being processed separately as following: primer sequences were trimmed, and amplicons were removed if they presented unidentified nucleotides (“N”), and/or a quality score smaller than 3, and/or length smaller than 100bp. In addition, we set TruncQ to 2, and also excluded sequences with a value greater than 8 for maxEE, i.e. the maximum number of expected errors (EE), with EE defined as sum(10^(-Q/10)), where Q is the quality score. Next, sequences were dereplicated and we denoised sequences using DADA2 error estimation algorithm, with band size set to 32bp and the DADA2’s error model for PacBio data. All three batches of quality-trimmed, denoised PacBio sequences were merged, and chimeric sequences were identified and removed using the “removeBimeraDenovo” function in DADA2. Finally, all sequences were classified using our custom database as reference (trnF\_db11Feb22.fa), and the DADA2’s Bayesian naïve classifier algorithm.

After chimera removal and taxonomical assignment, data was transformed into a phyloseq object (McMurdie & Holmes, 2013) and decontaminated using the package “decontam” (Davis et al., 2017). We found that a number of likely contaminants were present in the control samples, such as sequences identified as *Citrus* and *Pinus*. After further investigation, we found that the water used to dilute the primer solutions was contaminated by these taxa, in addition to other plant and algae taxa. Although these are obvious contaminants, we cannot rule out the possibility that some of these contaminants may also be present as diet items, in particular taxa identified as native plants (e.g. *Carya* and *Pinus*). Therefore, we set up our filter to a moderately-high threshold of sensitivity (0.5) and removed all 15 ASV identified as contaminants from our samples. To further improve taxonomy, we also used the NCBI Blast tool against their nt database to classify the remaining ASVs without a genus ID (e-value = 1e-30, percent ID 90-100% for genus identification). Next, we removed any ASV with a classification other than Viridiplantae (Kingdom), or ASVs identified as mosses (Class Bryopsida). Therefore, only spermatophytes were left for downstream analysis.

For diversity analysis, we aggregated ASVs at the genus level. Alpha diversity (Shannon index) was calculated for each species and month of sampling. For beta-diversity, we accounted for different sampling depth by rarefying samples to the 50% quantile of sample size range (108 sequences per sample). Principal Coordinate Analysis (PCoA) was used to evaluate diet dissimilarity among samples, using Weighted-Unifrac distances. Permutational multivariate analysis of variance (PERMANOVA) was used to test differences in diet due to sampling location, species identity and/or time of sampling. All these analyses and figures were generated in R, using the packages mentioned above, in addition to: tidyverse (Wickham et al., 2019), ggplot2 (Wickham, 2016), RcolorBrewer (Neuwirth, 2022), ggside (Landis, 2022), vegan (Oksanen et al., 2022), microViz (Barnett et al., 2021), phangorn (Schliep, 2011), GUnifrac (Chen et al., 2023), cowplot (Wilke, 2024), and microbiome.

References

Barnett et al., (2021). microViz: an R package for microbiome data visualization and statistics. Journal of Open Source Software, 6(63), 3201, https://doi.org/10.21105/joss.03201

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016). “DADA2: High-resolution sample inference from Illumina amplicon data.” Nature Methods, **13**, 581-583. [doi:10.1038/nmeth.3869](https://doi.org/10.1038/nmeth.3869).

Chao Yan, Jianchang Du, Lu Gao, Ying Li, Xilin Hou. 2019 The complete chloroplast genome sequence of watercress (Nasturtium officinale R. Br.): Genome organization, adaptive evolution and phylogenetic relationships in Cardamineae. Gene, 699: 24-36 https://doi.org/10.1016/j.gene.2019.02.075.

Chen J, Zhang X, Yang L, Zhang L (2023). *GUniFrac: Generalized UniFrac Distances, Distance-Based Multivariate Methods and Feature-Based Univariate Methods for Microbiome Data Analysis*. R package version 1.8, <https://CRAN.R-project.org/package=GUniFrac>.

Davis NM, Proctor D, Holmes SP, Relman DA, Callahan BJ (2017). “Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data.” *bioRxiv*, 221499. [doi:10.1101/221499](https://doi.org/10.1101/221499).

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H, **Twelve years of SAMtools and BCFtools**, GigaScience (2021) 10(2) giab008

Elbrecht V, Braukmann TWA, Ivanova NV, Prosser SWJ, Hajibabaei M, Wright M, Zakharov EV, Hebert PDN, Steinke D. Validation of COI metabarcoding primers for terrestrial arthropods. PeerJ. 2019 Oct 7;7:e7745. doi: 10.7717/peerj.7745. PMID: 31608170; PMCID: PMC6786254.

Elbrecht V, Hebert PDN, Steinke D. Slippage of degenerate primers can cause variation in amplicon length. Sci Rep. 2018 Jul 20;8(1):10999. doi: 10.1038/s41598-018-29364-z. PMID: 30030475; PMCID: PMC6054607.

Elbrecht V, Leese F. Validation and development of freshwater invertebrate metabarcoding COI primers for Environmental Impact Assessment. Front Fre Sci. 2017;5:1–11.

Lahti, L, Shetty, S. (2017). Tools for microbiome analysis in R. Version . URL: <http://microbiome.github.com/microbiome>.

Landis J (2022). ggside: Side Grammar Graphics. R package version 0.2.2,

<https://CRAN.R-project.org/package=ggside>.

Langmead B, Salzberg S. [Fast gapped-read alignment with Bowtie 2](http://www.nature.com/nmeth/journal/v9/n4/full/nmeth.1923.html). [*Nature Methods*](http://www.nature.com/nmeth). 2012, 9:357-359.

McMurdie, PJ, Holmes, S (2013) phyloseq: [An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data](http://dx.plos.org/10.1371/journal.pone.0061217). PLoS ONE. 8(4):e61217

Neuwirth E (2022). RColorBrewer: ColorBrewer Palettes. R package version 1.1-3,

<https://CRAN.R-project.org/package=RColorBrewer>.

Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R,

Solymos P, Stevens M, Szoecs E, Wagner H, Barbour M, Bedward M, Bolker B, Borcard

D, Carvalho G, Chirico M, De Caceres M, Durand S, Evangelista H, FitzJohn R,

Friendly M, Furneaux B, Hannigan G, Hill M, Lahti L, McGlinn D, Ouellette M,

Ribeiro Cunha E, Smith T, Stier A, Ter Braak C, Weedon J (2022). vegan: Community

Ecology Package. R package version 2.6-4,

https://CRAN.R-project.org/package=vegan.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL <http://www.rstudio.com/>.

RStudio Team (2023). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL <http://www.rstudio.com/>.

Schliep K (2011). “phangorn: phylogenetic analysis in R.” Bioinformatics, 27(4), 592–593. doi:10.1093/bioinformatics/btq706.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J. and Sahl, J.W., 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology, 75(23), 7537-7541.

W Rodney Cooper, David R Horton, Mark R Wildung, Andrew S Jensen, Jenita Thinakaran, Dalila Rendon, Louis B Nottingham, Elizabeth H Beers, Carrie H Wohleb, David G Hall, Lukasz L Stelinski, Host and Non-host ‘Whistle Stops’ for Psyllids: Molecular Gut Content Analysis by High-Throughput Sequencing Reveals Landscape-Level Movements of Psylloidea (Hemiptera), Environmental Entomology, Volume 48, Issue 3, June 2019, Pages 554–566, <https://doi.org/10.1093/ee/nvz038>

Taberlet, P., L. Gielly, G. Pautou, and J. Bouvet. 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant Mol. Biol. 17: 1105–1109.

Vijverberg, K., Bachmann, K. Molecular evolution of a tandemly repeated trnF(GAA) gene in the chloroplast genomes of Microseris (Asteraceae) and the use of structural mutations in phylogenetic analyses., *Molecular Biology and Evolution*, Volume 16, Issue 10, 1 October 1999, Pages 1329–1340, <https://doi.org/10.1093/oxfordjournals.molbev.a026043>

Wickham, H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag

New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes

A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J,

Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H

(2019). “Welcome to the tidyverse.” Journal of Open Source Software, 4(43),

1686. doi:10.21105/joss.01686

Wilke C (2024). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 1.1.3, <https://wilkelab.org/cowplot/>.