

MOLECULAR DIET ANALYSES OF NORTH AMERICAN BATS

By

Devon R. O'Rourke

Bachelor of Arts, Williams College, 2006

Masters of Arts in Teaching, Boston University, 2010

DISSERTATION

Submitted to the University of New Hampshire

in Partial Fulfillment of

the Requirements for the Degree of

Doctor of Philosophy

In

Molecular and Evolutionary Systems Biology

September 2019

This thesis/dissertation was examined and approved in partial fulfillment of the requirements for the degree of Ph.D in Molecular and Evolutionary Systems Biology by:

Dissertation Director, Matthew D. MacManes, Associate Professor, UNH MCBS

David C. Plachetzki, Assistant Professor, UNH MCBS

Adrienne I. Kovach, Associate Professor, UNH NREN

Jeff R. Garnas, Assistant Professor, UNH NREN

Jeffrey T. Foster, Associate Professor, NAU PMI

On August 12, 2019

Approval signatures are on file with the University of New Hampshire Graduate School.

## DEDICATION

To my wife, my best friend and favourite teacher;

To my mother, for her unconditional love.

To my father, who I miss every night.

To my daughter, the reason I get out of bed every morning.

## ACKNOWLEDGEMENTS

Thank you to the committee: I appreciate the wisdom of Dave Plachetzki in devising experiments that have multiple stories to tell; the inspiration of Adrienne Kovach to consider conservation-minded projects; the care and critical thinking of Jeff Garnas; the compassion and unending dedication to his students of Matt MacManes; the calm stewardship of Jeff Foster. Thanks to all the UNH students who offered countless instances of assistance in making these experiments possible—in particular, special thanks to Meg Ange, Katherine Ineson, Jen Spillane, and Nick Rouillard for your willingness to share samples, bench space, and ideas. Marcus Dillon, Kenny Flynn, and Sean Buskirk were instrumental in showing me how to be an effective graduate student. Andrew Lang and Meg Hartwick were essential in maintaining a semblance of sanity; thank you for your friendship. Special thanks to Dean Cari Moorhead, Dean Jon Wraith, and the Graduate School for their financial support. Many other people went out of their way to help me on all things bat poop: Jon Palmer and Michelle Jusino for showing me the bioinformatic ropes; dozens of citizen science volunteers who collected guano samples for nothing but a smile; and the New Hampshire Agricultural Experiment Station for providing funds to pilot the entire endeavor. Thank you to Dr. Anita Klein for her willingness to share equipment and expertise, and to Kelley Thomas, Jeff Hall, and Steve Simpson for sharing their wisdom among all things sequencing. Thanks to Kevin Drees for being patient with such an inexperienced bioinformatics student, and Katy Parise for being so caring to me and every other student that depended on her in countless ways each and every day.

This research and my time at the university was possible thanks to many scientists willing to give a totally unqualified person a chance. Thank you to Vaughn Cooper for your willingness to show a high school science teacher how to a graduate student at the University of New Hampshire. I would never have applied to UNH if it wasn't for Dr. April Burch demonstrating the value using graduate training in secondary schools. None of this would be possible if not for Dr. Lara Hutson providing me the employment that allowed me to enter the United States; our work together with zebrafish opened my eyes to the kinds of projects graduate school might entail. Finally, thanks to my undergraduate thesis advisor, Dr. Betty Zimmerberg, who had no business allowing a thoroughly unqualified biology major in her lab, but through her generosity and care, provided the first of many hands that helped pull me in this direction.



## TABLE OF CONTENTS

DEDICATION .....	iii
ACKNOWLEDGEMENTS.....	iv
LIST OF TABLES.....	viii
LIST OF FIGURES .....	xi
ABSTRACT .....	xiii

CHAPTER	PAGE
INTRODUCTION .....	1
I. A TOTAL CRAPSHOOT? TOWARDS A UNIFIED APPROACH TO MOLECULAR DIET ANALYSIS OF ARTHROPOD COI AMPLICONS.....	8
Abstract .....	8
Introduction .....	9
Methods .....	15
Mock samples .....	16
Guano samples .....	16
Laboratory work .....	17
Data processing with QIIME 2 .....	18
Data filtering with R .....	18
Figures and statistical tests with R .....	20
Database curation .....	21
Classification .....	23
Results .....	25

Sequence filtering .....	25
Diversity estimates .....	28
Database construction .....	31
Classification comparisons .....	34
Discussion .....	39
Sequence filtering and diversity estimates .....	39
Database construction .....	45
Classification .....	49
II. LORD OF THE DIPTERANS: MOLECULAR DIET ANALYSES OF INDIANA BATS IN ILLINOIS	63
Abstract .....	63
Introduction .....	64
Methods .....	68
Site selection and guano collection .....	68
DNA extraction .....	69
Metabarcoding .....	70
Bioinformatics .....	72
Sequence processing .....	73
Constructing databases .....	73
Classification .....	75
Diversity estimates .....	76
Supervised learning with Random Forest classifier .....	77
Additional software .....	78
Results .....	78

Dietary breadth .....	78
Diversity estimates .....	80
Important features in Random forest classifier model .....	82
Discussion .....	84
III. EATING LOCALLY: MOLECULAR DIET ANALYSES OF NEW HAMPSHIRE BATS REVEAL VARIATION IN PREY CONSUMPTION ACROSS SPACE AND TIME .....	96
Abstract .....	96
Introduction .....	97
Methods .....	100
Sample collection .....	100
Mock community .....	101
Metabarcoding .....	102
Sequence processing .....	104
Sequence quality control and bat host identification .....	104
Diversity estimates .....	106
Pest analysis .....	108
Additional software .....	108
Results .....	109
Dietary breadth .....	109
Diversity estimates .....	110
Pest detections .....	114
Discussion .....	115
REFERENCES .....	130
APPENDIX .....	143

## LIST OF TABLES

TABLE	PAGE
TABLE 1. Dietary breadth of Indiana bats .....	90
TABLE 2. Dietary breadth of New Hampshire bats .....	124
TABLE 3. Pest genera detected in bat guano .....	125
TABLE S1. Guano samples collected per location .....	143
TABLE S2. Summary of per-library or per-Pipeline overlaps in shared ASVs .....	144
TABLE S3. Number of variants observed per filtering parameter and program .....	144
TABLE S4. Dunn's test for Hill Number 0 estimates using mock data .....	145
TABLE S5. Dunn's test for Hill Number 1 estimates using mock data .....	146
TABLE S6. Dunn's test for Hill Number 2 estimates using mock data .....	147
TABLE S7. Dunn's test for Hill Number 0 estimates using guano data .....	148
TABLE S8. Dunn's test for Hill Number 1 estimates using guano data .....	149
TABLE S9. Dunn's test for Hill Number 2 estimates using guano data .....	150
TABLE S10. PERMANOVA of mock data with Dice-Sorensen estimate .....	151
TABLE S11. PERMANOVA of mock data with Bray-Curtis estimate .....	151
TABLE S12. PERMANOVA of mock data with Morisita-Horn estimate .....	151
TABLE S13. PERMANOVA of guano data with Dice-Sorensen estimate .....	152
TABLE S14. PERMANOVA of guano data with Bray-Curtis estimate .....	152
TABLE S15. PERMANOVA of guano data with Morisita-Horn estimate .....	152
TABLE S16. Number of unique arthropod taxa within each of the three COI databases .....	153

TABLE S17. Number of unique arthropod sequences within each of the three COI databases .....	153
TABLE S18. Taxonomic missingness among COI databases .....	153
TABLE S19. Common taxonomic names assigned among classifiers .....	153
TABLE S20. Breadth and detection frequency of arthropod Orders .....	154
TABLE S21. Core sequence variant detection and sequence counts .....	155
TABLE S22. Summary of alpha diversity estimates for collection Site and Month .	156
TABLE S23. Dunn's Test for observed richness (Hill $q=0$ ) .....	157
TABLE S24. Dunn's Test for Shannon entropy (Hill $q=1$ ) .....	157
TABLE S25. Dunn's Test for Simpson's 1-D index (Hill $q=2$ ) .....	158
TABLE S26. PERMANOVA for Site and Month using Dice-Sorensen metric .....	159
TABLE S27. PERMANOVA for Site and Month using Bray-Curtis metric .....	159
TABLE S28. PERMANOVA for Site and Month using Morisita-Horn metric .....	159
TABLE S29. PERMANOVA for Site and Month using unweighted Unifrac .....	159
TABLE S30. PERMANOVA for Site and Month using weighted Unifrac .....	159
TABLE S31. Summary of data collected and analyzed for each site and year .....	160
TABLE S32. Summary of most frequently detected arthropod genera .....	161
TABLE S33. PERMANOVA summaries for Site and Month effects for New Hampshire guano sample community composition using five distance estimates	162
TABLE S34. Tukey's pairwise comparison of observed richness by site .....	163
TABLE S35. Tukey's pairwise comparison of Shannon's entropy by site .....	164
TABLE S36. Tukey's pairwise comparison of Faith's PD by site .....	165
TABLE S37. PERMANOVA summaries for select New Hampshire sites .....	166

## LIST OF FIGURES

FIGURE	PAGE
FIGURE 1. Methods and analysis workflow.....	54
FIGURE 2. Denoising program and parameter evaluations of mock data .....	55
FIGURE 3. Read abundances per sequence variant for bat guano data .....	56
FIGURE 4. Accumulation curves of rarefied mock samples .....	57
FIGURE 5. Diversity estimates for rarefied bat guano .....	58
FIGURE 6. Database comparisons .....	59
FIGURE 7. Effect of database clustering on taxonomic ambiguity .....	60
FIGURE 8. Classifier evaluations with mock data .....	61
FIGURE 9. Common taxonomic identities among classifiers for guano data .....	62
FIGURE 10. Comparing diets using detections or sequence abundances .....	91
FIGURE 11. Diversity estimates using Hill Numbers .....	92
FIGURE 12. Beta diversity estimates of Indiana bat samples .....	93
FIGURE 13. Random Forest classifier performance. ....	94
FIGURE 14. ASVs important to identify Site and Month differences using Random Forest classifier .....	95
FIGURE 15. Map of the twenty New Hampshire collection sites .....	126
FIGURE 16. Observed richness per sample among select 2016 NH sites .....	127
FIGURE 17. Relative proportions of arthropod COI shown for each site and year across sampling windows .....	128
FIGURE 18. Principal coordinates analysis biplot using weighted Unifrac .....	129

FIGURE S1. Locations for New England guano sampling .....	167
FIGURE S2. Per-library throughput among denoising programs .....	168
FIGURE S3. Total and fraction of expected reads in mock samples .....	168
FIGURE S4. Read abundances per ASV for each library of bat guano data .....	169
FIGURE S5. Bat guano data generates a sparse matrix of ASVs among samples.	170
FIGURE S6. Accumulation curves of mock samples that are not rarefied .....	171
FIGURE S7. Diversity estimates for unrarefied bat guano .....	172
FIGURE S8. NMDS ordination with guano data using Dice-Sorensen metric .....	173
FIGURE S9. NMDS ordination with guano data using Bray-Curtis metric .....	174
FIGURE S10. NMDS ordination with guano data using Morisita-Horn metric .....	175
FIGURE S11. Unique and shared taxa among three databases .....	176
FIGURE S12. Effect of clustering on taxonomic diversity by arthropod order .....	177
FIGURE S13. Number of ASVs assigned taxonomic information at Class through Species rank among classifiers for guano data .....	178
FIGURE S14. Map of Cypress Creek National Wildlife Refuge .....	179
FIGURE S15. Artificial Roost locations at Cypress Creek NWR .....	180
FIGURE S16. Sequence depths per ASV classified as bat .....	181
FIGURE S17. Distribution of sequence counts among frequently detected ASVs ..	182
FIGURE S18. Dipteran ASV abundances and detections per site and month .....	183
FIGURE S19. Random forest classifier accuracy by Month .....	184
FIGURE S20. Important ASVs to Random Classifier model by Month .....	185
FIGURE S21. Per Month ASV abundance and detections .....	185
FIGURE S22. Dipteran ASVs relevant to Random Forest classifier .....	186

FIGURE S23. Alpha diversity estimates of select 2016 New Hampshire sites .....	187
FIGURE S24. Proportions of sequences per ASV per sample .....	188
FIGURE S25. Distributions of sequence abundances per select ASVs .....	189
FIGURE S26. Detection frequency of ASVs in both or single year per site .....	190
FIGURE S27. Relative abundances of arthropod COI per site per week per year .	191
FIGURE S28. Principal correspondence analysis of dietary composition for select 2016 New Hampshire locations using weighted and unweighted dissimilarity measures .....	192



ABSTRACT

MOLECULAR DIET ANALYSES OF NORTH AMERICAN BATS

by

Devon R. O'Rourke

University of New Hampshire

A food web is a model of the feeding relationships among organisms in an environment. The fidelity of this model is limited principally by the ability to detect these interactions. Researchers who study cryptic interactions such as nocturnal insectivory in bats typically rely on fecal samples to identify trophic connections. Historically these diet analyses were limited to morphological inspection of arthropod fragments, however modern metabarcoding techniques have improved the richness and specificity of consumed prey: rather than bats foraging for a few arthropod orders, we observe hundreds of species among guano samples. Animal metabarcoding is not without bias; nevertheless, a decade of improvements upon such biases have focused largely on molecular portions while bioinformatic considerations remain unresolved. When researchers use distinct software to perform their analyses—tools that have not yet been compared in animal metabarcoding studies—it is unclear if distinct perspectives between two experiments represent meaningful biological differences, or if they arise because of the alternative programs and parameters deployed. We investigated three fundamental bioinformatic tasks that impact a metabarcoding experiment: sequence processing, database construction, and classification (Chapter I). These comparisons

offer guidance regarding which steps are most sensitive to parameterization and are therefore in need of optimizing for individual experiments, as well as highlight areas that are in need of critical improvement. We applied these bioinformatic lessons to a molecular diet analysis of Indiana bats, the first ever for this endangered species (Chapter II). While management decisions currently focus on protecting roosting habitat, our molecular analyses provide evidence that site-specific data is needed to more effectively inform conservation practices. For example, while these bats forage a broad swath of the arthropod community, the molecular data suggests they rely on particular aquatic habitats that are not currently protected. Finally, we investigated the diets of New Hampshire bats by collaborating with citizen scientist volunteers throughout the state to perform an extensive sampling regime in that spanned 20 locations over 2015 and 2016, and sequenced more than 900 guano samples (Chapter III). Molecular analysis of these data suggested these bats are foraging hundreds of arthropod species, including some turf and forest pests, demonstrating that our local bats provide ecosystem services. Individual diets varied across season and site, providing evidence of highly flexible and local foraging behaviors.

## INTRODUCTION

Food webs are essential to understanding how resources are partitioned within a community and how energy flows through an ecosystem. Identifying the diets of species within an environment provides the foundation for constructing food webs. In some cases, trophic interactions are readily observable—a hummingbird feeds from a honeysuckle flower or an osprey captures a trout. In other systems, relationships are often more challenging to document because such interactions occur in environments that are difficult to directly observe; nocturnal animals, organisms in aquatic habitats, and consumers of phenotypically similar prey often require alternative strategies to describe who is eating whom. Historically, morphological analyses of fecal or gut materials have been widely applied to investigate such cryptic interactions, and include diets of herbivores <sup>1</sup>, birds <sup>2</sup>, bats <sup>3</sup>, and marine mammals <sup>4</sup>. Notably, these morphological analyses offer more than basic lists of gut contents. For instance, diets can elucidate the dynamics of native and exotic seed dispersal <sup>5</sup>; or reveal how sympatric species allocate resources among wolves <sup>6</sup>, birds <sup>7–9</sup>, or bats <sup>10–12</sup>. However, morphological techniques are often limited in several ways: teasing apart hundreds or thousands of fragments from a single sample is time-intensive; taxonomic resolution of diet contents is rarely species- (or genus-, or family-) specific; less digestible prey are more likely to be identified than soft-bodied organisms <sup>13–15</sup>. Over the past decade, researchers have increasingly relied on several alternative techniques in an attempt to

overcome these shortcomings, with molecular metabarcoding increasingly adopted as the foremost method to describe previously intractable food webs <sup>13,16–18</sup>.

Modern molecular metabarcoding analyses of animal diets are conducted by extracting DNA from a sample, amplifying a particular marker gene shared among the group of taxa of interest, sequencing these gene fragments using a high throughput instrument, and assigning taxonomic identities to the observed sequence variants using a reference library <sup>19–21</sup>. These molecular techniques are applied across the globe: from diet partitioning among African herbivores <sup>22</sup> and Jamaican frugivorous bats <sup>23</sup>, to trophic interactions in the arctic <sup>24</sup> or Texas deserts <sup>25</sup>, the high-speed, low-cost, and improved taxonomic resolution offered through metabarcoding has illuminated previously undocumented species relationships (see also: Table 1 of Deagle et al. <sup>26</sup>). However, animal metabarcoding is not without complications <sup>15,19</sup>. DNA from fecal material is typically highly fragmented <sup>27</sup> thus relatively short mitochondrial gene fragments—often Cytochrome Oxidase I (COI)—are targeted for amplification owing to the higher copy number of mitochondrial to nuclear genomes in a sample. While small mtDNA targets help capture as large a fraction of the available sequence diversity in a sample as possible, amplifying short fragments reduces the ability to distinguish among highly homologous taxa. In fact, one of the foremost resources for all COI datasets, the Barcode of Life Database (BOLD) <sup>28</sup>, generates operational taxonomic units called Barcode Index Numbers (BINs) that frequently contain multiple species within a single group, with reference sequences for each BIN (derived from Sanger sequencing) typically spanning three times the length of a typical COI amplicon (derived from a high

throughput sequencer). Thus, species-level resolution depends on the nature of the expected phylogenetic diversity of the diet components, and few comprehensive analyses of interspecies variation across broad taxonomic groups exist for COI (though see Sun et al. <sup>29</sup> for one such example in mollusks, and Pentinsaari et al. <sup>30</sup> for select arthropods). Additionally, selective amplification of particular taxa is a primary concern for dietary analyses of generalist species such as insectivorous bats and birds. Both in silico <sup>31</sup> and empirical tests <sup>32</sup> have suggested that primers used in some earlier COI metabarcoding experiments <sup>33,34</sup> failed to characterize the available arthropod diversity among samples. Nevertheless, recent modifications of COI amplification protocols through blocking oligos <sup>35</sup>, mixed primer sets <sup>36</sup>, and alternative primer design <sup>32</sup> represent a few examples of changes that improve the specificity and breadth of the diversity captured in molecular diet experiments. Despite these challenges, molecular metabarcoding remains an invaluable tool to investigate species interactions and construct food webs.

Technical reviews of animal metabarcoding procedures have investigated the myriad molecular decisions a researcher faces when conducting an experiment <sup>16,17,19,20,37</sup>: Which collection and storage practices are optimal? What is the best way to extract DNA from the sample? What primers and PCR settings are necessary? Do I need to amplify replicate samples? Does the sequencing platform used matter? However, far less attention has been given to the equivalent variety of bioinformatic considerations, though discussions of how to incorporate sequence counts in animal metabarcoding studies are emerging <sup>26</sup>. In Chapter 1, I explore three of the significant challenges in

animal metabarcoding that occur after sequence data has been generated: sequence processing, database construction, and classification. Denoising refers to any method by which representative sequences are identified and counted from raw sequence data, and this is the first step in any sequence processing pipeline. Historically, denoising involved clustering sequences into operational taxonomic units (OTUs), and while the effect of OTU clustering parameters has been investigated in COI recently <sup>19</sup>, this method is itself already outdated. Instead, programs that identify amplicon sequence variants (ASVs) offer a more precise representation of the exact sequences observed in the dataset. In addition to denoising programs themselves, sequence processing also includes decisions on how to best filter the resulting representative sequences. I used a biological mock community to explore how diversity estimates would vary depending on combinations of these filtering strategies and sequence. In addition, I compared the programs that perform the classification process, as well as evaluated differences among the databases that provide the reference libraries for classification. Collectively these insights provide animal metabarcoding researchers one example with which to guide their investigations and highlight particular bioinformatic processes that are sorely in need of improvement (e.g., hybrid classifier development, database versioned releases).

I applied these bioinformatic insights to investigate diets of endangered Indiana bats in at Cypress Creek National Wildlife Refuge (Pulaski County, Illinois USA) in Chapter 2. White-Nose Syndrome has been particularly devastating to Indiana bats in the Northeastern U.S. <sup>38–40</sup>, and Illinois hibernacula represent just one of four states where

Indiana bat populations are primarily concentrated <sup>39</sup>. Eight artificial roosts were erected on the refuge previously in 2014 in attempts to protect against continued loss of suitable roosting habitat due to expanding agricultural practices surrounding the reserve.

Ongoing monitoring efforts suggested that these efforts are indeed working as intended, as multiple Indiana bat maternity colonies have been maintained at these sites since their construction <sup>41,42</sup>. Nevertheless, the existing management plan extends limited protections to bat roosting areas and does not specify what foraging habitat to protect <sup>43</sup>. Thus I applied our molecular techniques to determine the diet contents of these bats to understand better the likely landscapes used for foraging. Earlier morphological analyses of Indiana bat diets suggested they are generalist insectivores with diets primarily consisting of beetles, flies, and moths <sup>44–46</sup>, though compositional changes in these arthropod orders, as well as other insects like caddisflies, vary across study sites <sup>47</sup>. These analyses revealed that these bats are capable of consuming an even broader diet than previously recorded, and the majority of the prey consumed are aquatic invertebrates—an observation that contrasts with previous morphological observations of Indiana bat diets in forested communities just twenty miles east of the reserve <sup>46</sup>. This work provides clear evidence that future management practices can benefit from incorporating molecular diet analyses in determining when and where bats are foraging and illustrates that habitat conservation needs to extend beyond roosting sites.

Indiana bats are one of several North American species to have suffered significant population declines due to White-Nose Syndrome <sup>48,49</sup>. While recent reports suggest some areas are experiencing stabilization in population size <sup>50,51</sup>, federal guidelines <sup>52–54</sup>

and population models <sup>55,56</sup> alike suggest that management efforts are critical to preventing future declines even among seemingly stable populations. Effective management requires a thorough understanding of the foraging habitats required by the species, yet few molecular bat diet studies among North American species affected by White-Nose Syndrome have been performed. Previous metabarcoding work documented spatial and temporal variation in Little brown <sup>57,58</sup> and Big brown <sup>59</sup> bat diets, but no such molecular work has been performed in New Hampshire—a state Little brown bat populations have declined between 70-100% (K. Ineson, pers. comm). Additionally, these initial analyses were restricted to order-level comparisons, which can vastly underestimate differences in diet composition. Furthermore, seasonal variation was defined across spring, summer, and fall; these temporal windows may be too broad to accurately capture many of the changes in prey consumed given the ephemeral life histories of many aerial arthropods. In Chapter 3, I conducted an experiment to evaluate changes in biweekly diet composition among 18 bat colonies throughout New Hampshire in 2015 and 2016. I analyzed over 900 passively collected guano samples and found these bats to have highly flexible diets that vary by season and site. In addition, I discovered that a few of the most frequently detected insects were turf and forest pests, and discuss the likely limits of bat guano as a pest surveillance method. Collectively, these data suggest that molecular analyses can be used to act as an initial screen for potential pests of concern, as well as provide a broad characterization of the species' diet. However, we find that any such depiction is suitable only at a local level—particularly among instances when this information is used to inform conservation of foraging habitat.



This dissertation aims to evaluate the tools available to a molecular ecologist interested in animal metabarcoding and apply these insights to characterize bat diets. While the specific techniques investigated will undoubtedly be rapidly replaced, I hope that this work serves to illustrate the value in comparing as many programs and parameters as possible. What is not well documented in this dissertation is the importance of community partnerships. Without citizen science volunteers I would have been unable to amass the vast collection of guano used in Chapter 3, and without partnerships with USFWS managers, there would be no molecular study of Indiana bat diets. I began characterizing bat guano in 2016 using QIIME 1 <sup>60</sup> but switched to another platform, AMPK <sup>61</sup>, because of an ability to form a personal connection with the software developer. That relationship was the first instance with which I came to understand how many assumptions can be made within a single program, and was the impetus of Chapter 1. Notably, all of the analyses presented herein ultimately were performed using QIIME 2 software <sup>62</sup>, which offers not only open-access software and extensive documentation but a community forum consisting of many users and developers. Platforms like these represent the best way for the animal metabarcoding community to discuss, contribute, and evolve best practices that provide the most accurate characterization of their particular environment.

## CHAPTER I

### A Total Crapshoot? Evaluating Bioinformatic Decisions in Animal Diet Metabarcoding Analyses

#### **ABSTRACT**

Metabarcoding studies provide a powerful method to estimate the diversity and abundance of organisms. While strategies exist for optimizing sample and sequence library preparation, best practices for bioinformatic processing of amplicon sequence data are lacking in animal diet studies. Here we evaluate how decisions made in core bioinformatic processes including sequence filtering, database design, and classification can influence animal metabarcoding results. We show that denoising methods have lower error rates compared to traditional clustering methods commonly used by animal metabarcoding studies, although these differences are largely mitigated by removing low-abundance sequence variants. We also found that available reference datasets from Genbank and BOLD for the animal marker gene cytochrome oxidase I (COI) can be complementary and discuss methods to improve upon existing databases to include versioned releases. Taxonomic classification methods can dramatically affect results.

For example, the commonly used Barcode of Life Database (BOLD) Taxonomy API assigned fewer names to samples at the class through species levels in both a mock community and bat guano samples compared to all other classifier methods (BLAST, VSEARCH, SINTAX, and Naive Bayes). The lack of consensus on bioinformatic best-practices limits comparisons among studies and may be introducing unknown or unappreciated biases. Our work suggests that biological mock communities offer a powerful tool to evaluate the myriad computational decisions impacting animal metabarcoding accuracy. Further, these comparisons highlight the need for continual evaluations as new tools are adopted to ensure that the inferences drawn reflect meaningful biology instead of digital artefacts.

## **INTRODUCTION**

Metabarcoding of animal diets has fundamentally changed our insights into what these species are eating, expanding our understanding of the extent of dietary diversity and food web complexity<sup>17-18, 20-21</sup>. Modern sequencing approaches have enabled a broad range of studies but have resulted in myriad customized molecular and bioinformatic workflows that make comparisons among studies difficult. While some differences are expected in software choices among the variety of systems being studied, there has been no systematic approach to evaluating which of these various bioinformatic differences between metabarcoding studies occur because of system-specific requirements or are simply the result of the choice of the researcher. The factors that

need to be considered for amplicon analyses are extensive, including sample collection, primer and barcode design, sequencing platform, as well as sequence data processing and taxonomic assignment (reviewed by Pompanon et al. <sup>20</sup>, Clare <sup>18</sup>, and Alberdi et al. <sup>19</sup>). A basic recommendation for diet metabarcoding workflows would include sampling as exhaustively as budget and time allow, choosing primers that appear to amplify the select taxa with as little bias as possible, and sequencing with the most accurate platform available. Yet, rarely is such general advice free of context-dependent caveats, and frequently these subtle, experiment-specific distinctions are buried in supplementary bioinformatic methods. We were motivated to shed light on these often overlooked bioinformatic steps to help illustrate which processes particularly affect interpretations of animal diets.

Molecular metabarcoding experiments are motivated by a straightforward principle: characterizing the composition of a community. Nevertheless, the particular applications practiced by researchers are complex, and we focused on three related processes that are poorly described among animal metabarcoding projects: processing raw sequences into representative sequences, classifying those representative sequences, and constructing the necessary databases to perform such classification. Establishing the best practices for sequence filtering and classification are frequently determined using mock communities—known samples with expected sequence identity and abundance. This practice that is commonplace in microbial gene marker research <sup>63</sup>, but less common for arthropod datasets <sup>37</sup>. Wet-bench experiments with mock communities can be used to assess systematic error and biases in observed sequence data <sup>64</sup>, optimize

filtering parameters <sup>65</sup>, understand tradeoffs among sequence error correction approaches <sup>66</sup>, and evaluate taxonomic classification regimes <sup>67</sup>. Systematic evaluation of animal metabarcoding studies are growing but remain far more limited in scope relative to microbial experiments; synthetic mock samples have been used to explore the potential for alternative primer use <sup>68</sup>, while biological mock samples have been used to improve quality filtering of spurious sequence variants <sup>32</sup> as well as to evaluate the utility of PCR replicates <sup>69</sup>. In addition, a few studies have used real data (i.e. actual diet samples) to offer insights into the effects of sequencing platforms <sup>70</sup> as well as on a single filtering parameter <sup>19</sup>. We build upon these sequence filtering considerations by using both real and biological mock data to illustrate how certain bioinformatic decisions impact the interpretation of community richness and composition, a common focus of diet analyses.

One of the first considerations in an amplicon study is whether to correct unique sequences through a ‘denoising’ approach or a clustering approach. Denoising programs like DADA2 <sup>71</sup> or Deblur <sup>72</sup> generate error models to address potential sequence errors, while clustering programs group sequence variants into operational taxonomic units at some user-defined similarity <sup>73</sup>. These have been explored empirically in a microbial setting <sup>74</sup>; while the observed differences were small in that single study, practical reasons such as database independence and the potential to preserve sequence diversity suggest that using amplicon sequence variants (ASVs, inferred exact biological sequences) is more advantageous than operational taxonomic units (OTUs) <sup>75</sup>. Denoising methods have yet to see wider adoption in diet

metabarcoding studies (but see <sup>76</sup>); most studies use clustering methods <sup>70,77–80</sup>). We compare the denoising and clustering-based approaches using biological mock communities and a collection of bat guano samples to highlight their effects on common diversity metric results.

As with sequence filtering considerations, the process of assigning taxonomic information to sequence variants is fraught with decisions that can significantly impact the subsequent interpretation of animal diets. In fact, classification is a pair of related issues: first, there are multiple approaches to classification including alignment-based and machine-learning methods, and second, any classification method is dependent on the database of reference sequences provided. With respect to database construction there are relatively few database resources available among conventional animal diet metabarcoding studies. In particular, the Barcode of Life Database (BOLD) <sup>28</sup> serves as the principal resource among arthropod-specific metabarcoding studies, while GenBank <sup>81</sup> is often used for non-chordate investigations. Yet despite having few reference sequence sources, most animal diet studies differ from each other because each project generates a unique reference collection on an *ad hoc* basis prior to classifying sequences. In contrast to microbial reference databases such as Greengenes <sup>82</sup> and SILVA <sup>83</sup>, BOLD and GenBank are continuously updated and lack the kind of versioned history found among microbial reference databases. Thus a pair of studies using the same resource (e.g. BOLD) in the same discipline that were conducted less than a year apart may differ by tens to hundreds of thousands of reference sequences. To make matters more challenging, the animal diet metabarcoding community lacks standards for

both curation and classification. For instance, should databases be dereplicated (i.e., identical sequences be consolidated into one representative sequence), and if so, how should identical sequences containing different taxonomic names be treated? Is clustering a reference database—as is often done in microbial systems—appropriate for animal marker genes such as cytochrome oxidase subunit I (COI), and does clustering reference sequences reduce the amount of taxonomic information equally among different kinds of organisms? The role of database composition on results is unclear and makes it challenging to understand if differences observed between studies is the product of meaningful biology or database curation. Thus, we examined how particular bioinformatic criteria applied to reference databases impacts subsequent composition of the references available for classification: reference dataset choice, filtering references for taxonomic completeness, and clustering reference sequences.

Taxonomic classification of representative sequences varies considerably among animal diet metabarcoding studies but can directly affect results. Because of the extensive number of classifiers available, the potential for further variability among studies arises. For example, one may choose to classify sequences using a local alignment approach like BLAST<sup>84</sup>, a global aligner like VSEARCH<sup>85</sup>, or a kmer-based classifier, such as SINTAX<sup>86</sup> or Naive Bayes<sup>87</sup>. BOLD even offers its own classifier although few details describing the underlying algorithm are currently available<sup>28</sup> and unlike most other classifiers no source code is publicly documented. Additionally, hybrid methods are available wherein multiple distinct classifiers converge on a best match<sup>61</sup>. We compare a few of the most commonly used classifiers and further evaluate how

certain options within those programs may introduce biases—specifically by neglecting key parameters—and suggest specific measures that are needed to more accurately determine taxonomic identities in a dataset.

Finally, there is the challenge of trying to estimate species abundance of specific taxa in a diet, e.g. what proportion of the diet is comprised of mosquitoes. While creating and classifying representative sequences are processes common to most animal diet metabarcoding projects, only very recently have researchers explored diversity assessments using relative abundances (RA) of sequence counts instead of transforming these counts into a presence-absence (PA) matrix of samples and observed sequence variants (reviewed by Deagle et al. <sup>26</sup>). Despite this importance, the debate about the appropriateness of RA versus PA transformations often overlooks a more fundamental point: you need to design experiments with at least some insight into what sequences should and should not be present before you can consider any kind of transformation. While mock communities do not represent the true complexity observed in actual diet samples, mock samples are essential for ground-truthing bioinformatic processes such as filtering parameters. Specifically, the mock samples provide a positive control of known sequence identities and therefore enable evaluation of the frequency at which low-abundance sequence artifacts are generated, and how these relatively rare sequences are uniquely filtered by each bioinformatic pipeline. Mock communities can provide an empirically derived filtering strategy and assess the likelihood and relative abundances of unexpected sequences <sup>61</sup>.



We assessed sequence processing and classification methods using four libraries of COI data generated from an ongoing bat diet study that included a biological mock community sample and hundreds of bat guano samples for each sequencing run. While mock data provide a ground truth when evaluating different sequence filtering and classification techniques, guano data can provide relative comparisons of these procedures using the more complex samples typically found in animal metabarcoding projects. In addition, we sought to understand how our interpretations of apparent diversity within and between samples are influenced by such count transformations, and how the count data are influenced by the specific filtering program. Notably, while we tested each of these three broad processes separately, the entire workflow is interconnected. Sequence processing decisions can influence the composition of the representative sequences; the set of representative sequences available following sequence processing can influence diversity estimates; diversity estimates that use phylogenetic information or sequence abundances can be influenced by the initial filtering decisions; classification methods inform these phylogenetic diversity estimates; database construction affect classification. Our aim is not to present a single best pathway for all animal metabarcoding projects, but to illustrate which of these processes appear most sensitive to program (or parameter) choices. We performed each of these analyses using the QIIME 2 framework <sup>62</sup> to allow for increased methods transparency, reproducibility, and use of open source tools within the diet metabarcoding community.

## **MATERIALS AND METHODS**

Documentation, scripts, and the majority of data referred to in this manuscript are available at our Github repo: <https://github.com/devonorourke/tidybug/tree/master/docs>.

Several large database files are hosted at an Open Science Framework repo: <https://osf.io/k3eh6/>.

An overview of the methods used and the resulting data tables and figures produced are summarized in **Figure 1**.

### *Mock samples*

Mock community samples were constructed specifically for arthropod diet analyses of COI gene fragments<sup>32</sup>. The mock community used in this experiment consists of 24 representative arthropod COI sequences derived from 23 taxa; notably, one of the taxa (*Harmonia axyridis*) generated two distinct COI amplicons. The mock sample used in this project consists of equimolar concentrations of plasmids, not post-plasmid PCR product, thus some primer bias and variability in per-taxon abundance is expected.

### *Guano samples*

Individual guano pellets were passively sampled weekly from sites throughout northeastern US (**Figure S1, Table S1**). Guano samples were obtained by using sterile forceps in which pellets were collected from clean plastic sheets and stored in microcentrifuge tubes filled with 1 mL storage buffer (3.5M ammonium sulfate, 16.7 mM sodium citrate, 13.3 mM EDTA, pH 5.2). Plastic sheets were replaced weekly to avoid

cross contamination over the season. Samples were stored at -80 °C until DNA extraction.

### *Laboratory work*

Individual fecal pellets were extracted using the Qiagen DNEasy PowerSoil Kit (Qiagen, Hilden, Germany) following manufacturer guidelines. Samples were eluted with 60 µL of elution buffer and up to eight extraction blanks were included for 96-well each plate. We used a dual-indexed primer design inspired by Kozich et al.<sup>88</sup> to amplify a 181 bp COI gene fragment. This design incorporates the Illumina adapter as well as customized barcodes and COI-targeted primer sequences into a single oligo. The COI-specific primer region of this construct are identical to that used to generate the mock community data<sup>32</sup>; see the Github repo for all primer sequence information. Additional details regarding PCR conditions, quantitation of amplicons, and pooling of libraries is available

([https://github.com/devonorourke/tidybug/blob/master/docs/wetbench\\_workflow.md](https://github.com/devonorourke/tidybug/blob/master/docs/wetbench_workflow.md)).

Four libraries were submitted to Northern Arizona University and sequenced using an Illumina MiSeq platform (Illumina, San Diego, CA, USA) using v3 chemistry with 600 cycles of 2x300 bp paired-end read lengths. Raw sequence reads available at NCBI BioProject PRJNA518082: <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA518082>.

Note that each guano sample was sequenced only once, while the same mock community was independently amplified and pooled into every library. We describe the four sequencing runs as Libraries A–D, and specify the mock samples associated with each according to the library they are derived from.

### *Data processing with QIIME 2*

We imported raw sequence reads into a QIIME 2 v2018.11 environment <sup>62</sup> and trimmed unpaired reads with Cutadapt <sup>89</sup>. Trimmed reads were then filtered with one of three pipelines in QIIME 2: OTU clustering with VSEARCH <sup>85</sup> or denoising with DADA2 <sup>71</sup> or Deblur <sup>72</sup>. Full details are available at the ``sequence_filtering.md`` document in the tidybug/docs directory of the GitHub repo. Each of the four libraries were individually processed through the respective pipelines. The OTU clustering approach with VSEARCH mirrored the parameters outlined at the VSEARCH Wiki GitHub page (<https://github.com/torognes/vsearch/wiki/VSEARCH-pipeline>). For the Deblur pipeline we altered two parameters from their default: ``--p-min-reads 2`` and ``--p-min-size 1`` ensured that only singleton reads were discarded and all singleton ASVs were retained. This matches the default behavior in DADA2 to ensure that filtering behaviors were uniform across methods. Bat-associated COI sequences were identified from filtered reads and removed, then representative sequence (fasta-like) files and frequency tables of sequence counts (OTU table-like) were merged for all libraries. Sequence filtering documentation is available:

[https://github.com/devonorourke/tidybug/blob/master/docs/sequence\\_filtering.md](https://github.com/devonorourke/tidybug/blob/master/docs/sequence_filtering.md).

### *Data filtering with R*

In addition to comparing the default outputs from the three sequence processing pipelines, we explored the potential effects of additional filtering parameters. Filtering rare sequence variants frequently occurs in diet metabarcoding studies <sup>70,90</sup>. While

recommendations for removing samples with overall low read abundances are described in the microbiome literature <sup>65,91,92</sup>, best practices have not been determined for animal diet metabarcoding projects. We applied two simple filters to the default outputs of the pipelines: first, a “standard” filter required (1) dropping any sequence variant observed in just one sample across the entire dataset, and (2) retaining only samples with  $\geq 5000$  total filtered reads. Second, an “extra” filter incorporated the “standard” filters, and subtracted a single, fixed integer from each element of the feature table. This second filter removed sequence variants with very low read counts while scaling with library throughput, given that increasing number of artifacts are likely related to sequencing depth <sup>26</sup>. The integer used in the “extra” filter is obtained on a per-library basis and was defined as the maximum count value observed of an unexpected sequence variant in the (library-specific) mock sample. We identified these unexpected sequence variants by aligning all sequence variants present in a mock sample to a reference set of expected mock representative sequences to determine which ASVs were exact matches (100% identity), partial matches (97–99.9% identity), or unlikely matches ( $< 97\%$  identity). The maximum sequence count observed among unlikely match variants was what defined the “extra” filtering integer value on a per-library, per-filtering method basis. These analyses were performed in R <sup>93</sup> version 3.5.1 by importing with the QIIME2R package <sup>94</sup> and processed with Tidyverse <sup>95</sup>, Reshape2 <sup>96</sup>, Phyloseq <sup>97</sup>, and Vegan <sup>98</sup> packages. A custom R script was used to process these sequences:

[https://github.com/devonorourke/tidybug/blob/master/scripts/R\\_scripts/1\\_sequence\\_filtering.R](https://github.com/devonorourke/tidybug/blob/master/scripts/R_scripts/1_sequence_filtering.R).

### *Figures and statistical tests with R*

Following all filtering steps, we assessed alpha diversity estimates using iNEXT<sup>99,100</sup> and Vegan for each combination of filtering methods (VSEARCH, DADA2, or Deblur) and parameters (none, standard, extended). Parametric statistical tests were performed with base R “stats” function ‘aov’, and nonparametric pairwise comparisons were performed with the Matrix<sup>101</sup>, FSA<sup>102</sup>, and Dunn.test packages<sup>103</sup>.

To study the effects of filtering parameters and denoising methods on variation in community composition between samples, we conducted a permutational multivariate analysis of variance (PERMANOVA) using the ADONIS function in Vegan. We investigated these effects using three distance measures: an incidence-based measure (Dice-Sorensen (referred to as “Dice”)), and two quantitative metrics (Bray-Curtis (“Bray”) and Morisita-Horn (“Morisita”)); datasets were rarefied to a depth of 5000 sequences per sample. We performed these tests first for mock community samples, investigating the effects of denoising method and filtering parameters. In addition, we tested for the effects of denoising methods and filtering parameters on a subset of guano samples collected at a single location (Fox State Forest, Hillsborough NH) from April–October 2016, and added the date of collection (“MonthStart”) to the model (Distance ~ Method \* Filt + MonthStart). We visually evaluated community composition of these select guano data through non-metric multidimensional scaling (NMDS) for each distance measure using the metaMDS function in Vegan. Additional information linking figure and data tables to specific R scripts for diversity analyses is available: [https://github.com/devonorourke/tidybug/blob/master/docs/diversity\\_analyses.md](https://github.com/devonorourke/tidybug/blob/master/docs/diversity_analyses.md).

We relied on a series of additional R packages to create the figures used herein, including `cowplot` <sup>104</sup>, `ggpubr` <sup>105</sup>, `ggrepel` <sup>106</sup>, `ggridges` <sup>107</sup>, `stringi` <sup>108</sup>, `scales` <sup>109</sup>, and `viridis` <sup>110</sup>.

### *Database curation*

We compared three databases to assess the effects of database selection and curation on COI profiling results. Two databases were curated by other researchers. First, the AMPTK program <sup>61</sup> contains scripts to access a precompiled COI database containing both arthropod and chordate records derived exclusively from BOLD—this dataset is herein referred to as “Palmer”. Additional details regarding how this database was constructed are available (<https://amptk.readthedocs.io>). Second, a dataset containing Eukaryote-wide COI references derived exclusively from GenBank <sup>111</sup> was downloaded using the v3.2 reference sequences and is referred to as “Porter” herein. We restricted our comparisons across datasets to dereplicated arthropod records, which necessitated further filtering both of these datasets. The Palmer sequences required removing all chordate data, while the Porter dataset required the removal all non-arthropod records, applying the LCA algorithm to the remaining arthropod records, and then dereplicating.

A third database was curated following our own methods as described below—this dataset is herein referred to as “tidybug”. We accessed arthropod records from the BOLD database using the `bold` R package <sup>112</sup> on February 24, 2019. We then applied a custom script to retain only those records that contained the “COI-5P” marker code, and

removed all records that failed to include at least family-level taxonomic information—see ``bold_datapull.R`` for complete details. While the “bold” library package provided an invaluable resource to access data, it is limited in its capacity to thoroughly filter the records because the BOLD API itself restricts which columns of metadata can be accessed during downloading. For instance, we discovered that there are two “marker gene” fields that contain COI information among the specimen records, and only one of the two is used to filter records through BOLD directly. As a result, additional non-COI records, or those records that lack any information of that criteria, can pass through the R bold package’s “marker” filter. In addition, we discovered that the arthropod-wide dataset cannot be downloaded in a single request to the BOLD server. As a result we divided up our queries into smaller batches using the accompanying R ‘bold’ package. However, this approach can generate rare instances in which the NCBI formatted taxa do not match the BOLD terms, resulting in zero records in which such a discrepancy exists (see here for example: <https://github.com/ropensci/bold/issues/60>). We therefore applied our own method of data scraping to ensure all records in BOLD would be appropriately queried. Prior to dereplication, we adapted methods used to format the SILVA database that incorporated a least common ancestor (LCA) approach to retain only taxonomic information where redundant sequences contain disparate classifications. We then dereplicated with VSEARCH and applied the updated LCA-classified taxonomies to the remaining sequences using a consensus approach. Finally, to compare how clustering databases can impact the composition and completeness of information of available records, we applied the same LCA approach to cluster our curated database at 99%, 97% and 95% identities. All code used to process and curate



reference databases can be found in the `database\_construction.md` file in the project GitHub repo

([https://github.com/devonorourke/tidybug/blob/master/docs/database\\_construction.md](https://github.com/devonorourke/tidybug/blob/master/docs/database_construction.md)).

### *Classification*

We tested classifier accuracy using a mock community, and consensus between classifiers using bat guano data. Three of these classifiers are implemented in the QIIME 2 plugin q2-feature-classifier<sup>67</sup>: BLAST-<sup>84</sup> and VSEARCH-based<sup>85</sup> alignment followed by least common ancestor consensus taxonomy assignment; and a taxonomy classifier utilizing the kmer-based machine learning Naive Bayes classifier implemented in scikit-learn<sup>113</sup>. We added a second kmer-based method using the VSEARCH implementation of the SINTAX<sup>86</sup> algorithm. Collectively, these four classifiers shared the same database information—the tidybug database described above. We also explored how the BOLD classification engine would compare to other classifiers, but this comparison was limited because the specific parameters used in the classification regime are not publicly documented, nor is the specific database used for classification defined (i.e. there is no single file to download that represents the BOLD database at the time in which their taxonomy API is queried). Parameters for classifiers were modified from defaults to reflect more conservative standards where appropriate. For example, we increased the query coverage and percent identity for alignment-based classifiers, and increased the SINTAX probability cutoff. We also tried to keep parameter values similar across classifiers where possible (e.g. between BLAST and VSEARCH). We applied a custom R script to the BOLD API output to mirror the default

parameters present in BLAST and VSEARCH: first, to retain only matches with greater than 97% identity, and second, to apply an LCA process on remaining taxa in instances where multiple distinct taxa records exist. Pipelines for classification and the specific R scripts used to access the BOLD API are documented here:

[https://github.com/devonorourke/tidybug/blob/master/docs/classification\\_analyses.md](https://github.com/devonorourke/tidybug/blob/master/docs/classification_analyses.md).

Our classification comparisons focused on two DADA2-filtered datasets: the expected 24 mock community sequences, and the entirety of the “basic” filtered guano data. Mock data contain not only a ground truth of expected sequences, but also carry an expectation of taxonomic identity. We therefore assigned an expected taxonomy to each of the 24 mock samples from class through species level and compared the proportion of true positive, false positive, and false negative classification assignments from each of the five classifiers. We quantified precision as Taxonomic Accuracy Rate (TAR), and recall as Taxonomic Discovery Rate (TDR) following the conventions used in a previous microbiome classifier benchmarking study<sup>67</sup>. The F measure was calculated as the harmonic mean of TAR and TDR, and reflects a balance between precision and recall.

Unfortunately these mock samples were not obtained as vouchered specimen, and although the researchers who developed the mock community consisted of scientists with extensive experience identifying many of the samples selected, Sanger sequences of the full 650 bp COI amplicon for each specimen were used to determine the taxonomic identities of specimens using BLAST at NCBI. The resulting expected mock

names range in varying degrees of taxonomic completeness, and while most samples contain species identity, some are only described at the genus, family, or order level. The guano data contain no such ground truth, but do contain vastly more ASVs to classify. Our intention in comparing classifiers with actual guano data was to assess the instances in which classifiers agree or disagree with respect to a given taxonomic name at a particular level (from class through species).

## RESULTS

We evaluated sequence filtering and classification regimes using both bat guano and biological mock community sequence data. Biological mock community data provide an important means to compare how observed outcomes deviate from expected results, while bat guano samples provide a more realistic evaluation of how certain bioinformatic decisions impact an analysis.

### *Sequence filtering*

Representative sequences were identified using one of three amplicon processing programs (hereafter termed *denoising methods*). Note that VSEARCH is not technically a denoiser in the traditional sense of building an error model to identify and act upon sequence error, but we use the term to refer to all three programs in the general sense of attempting to collapse the entirety of a dataset into representative sequence variants, whether they be OTUs or ASVs. The resulting abundances of filtered reads varied with

respect to denoising method among bat guano samples and mock replicates with Deblur retaining fewer filtered reads than DADA2 or VSEARCH (**Figure 2**).

We aligned ASVs observed in each mock sample to the expected mock community reference sequences to determine the frequency of occurrence and relative read abundances of expected and unexpected sequence variants. For “basic” (default) parameters VSEARCH retained fewer expected sequence variants than DADA2 or Deblur, while VSEARCH produced more unexpected sequence variants than either DADA2 or Deblur (**Figure 2**). Likewise, DADA2 and Deblur retained similarly higher fractions of expected reads despite having substantially different total numbers of reads (**Figure S3**) among all replicate mock samples. Indeed, the number of observed exact matches was nearly identical among DADA2 and Deblur mock samples, although the proportion of unexpected sequences was greater with Deblur (**Table S2**). Applying “standard” and “extra” filtering parameters to each dataset reduced the number of unexpected sequences among mock samples for all denoising methods, although the impact of these filters was most evident for VSEARCH.

Because the bat guano samples did not have any known set of ASVs, we could not apply a similar comparison of denoising methods as was performed with mock samples. We instead evaluated the distribution of sequences abundances to understand how denoising methods and filtering parameters perform with guano samples and a diversity of sequences. As was observed with mock data, the application of filtering parameters tends to eliminate observed differences among denoising methods among all guano

data collectively (**Figure 3**) and on a per-library basis (**Figure S4**). Notably, read abundances and frequency of occurrence for a particular ASV were dramatically different between mock and real datasets. Mock sequence variants are characterized either with high sequence counts that were detected repeatedly among samples (i.e., the expected mock sequences) or low sequence counts detected nearly always as singletons (i.e., sequencing error) (**Figures 2, S3**). Guano data, however, consisted mainly of infrequently observed low abundance ASVs (**Figures 3, S4**). Thus, similar to mock data, the application of “standard” or “extra” filtering parameters to guano data can result in fewer detections of any particular ASV by reducing these low-abundance detections, while retaining the majority of sequences in the entire dataset. Unlike mock data, it is not possible to determine whether these bat guano ASVs being removed are due to sequencing error, but our mock dataset suggests that several low abundance ASVs are indeed produced during sequencing, and that these filtering parameters can be effective at removing unexpected sequence variants.

Overall, we found that the profiles of read distributions reflect the parameters of their respective pipeline for both mock and real datasets. For instance, because only DADA2 functions to identify errors and correct them, the choice of a specific minimum abundance threshold to discard observations (e.g., per ASV or per sample) is pipeline dependent. For example, the effect of the “extra” filter is much less noticeable for DADA2 than either VSEARCH or Deblur, because many of the low abundance reads that were discarded with that filter had already been corrected by DADA2. Choosing optimal parameters to reduce sequence error without sacrificing loss of meaningful

biological data is especially critical for animal metabarcoding studies that use diversity metrics transform read abundances into presence-absence observations.

### *Diversity estimates*

We separately evaluated the effects of denoising method and filtering parameters on estimates of inter-sample diversity among our mock community data and bat guano samples. However, because differences in sampling depth are known to influence diversity metrics among high-throughput datasets <sup>114</sup>, we also investigated the impact of rarefying—randomly subsampling data without replacement—on diversity measures. We chose this method because of its widespread use, and because alternative normalization techniques have outcomes that are likely to be context dependent <sup>115</sup>. Our motivation in comparing unrarefied to rarefied data was to highlight that additional normalization is yet another bioinformatic consideration that can impact diet analyses.

We first generated accumulation curves of sequence diversity (analogous to species richness) for mock individual mock samples using both abundance information and transformed presence-absence data. We found that denoising method and filtering parameters both impact the asymptotic estimate of diversity for rarefied presence-absence data (**Figure 4a**), but have negligible effects when abundance information is incorporated (**Figure 4b**). A more extreme version of the same trend was observed for unrarefied mock data (**Figure S6**), with both Deblur and VSEARCH making larger overestimates of richness. Nevertheless, as with rarefied data, incorporating abundance information in unrarefied data leads to similar estimates of diversity among all denoising

methods across each filtering parameter. Guano samples similarly demonstrated that incidence-based diversity estimates are higher than estimates using abundances for both rarefied data (**Figure 5**) and unrarefied data (**Figure S7**). The reduction in sequence variant equivalents using Hill values of 1 or 2 indicate that relatively few distinct sequences contribute to the majority of the sequence information in a given guano sample. Unlike the idealized mock sample where DADA2 using “extra” filtering had similar diversity estimates for abundance and presence-absence data, these guano data illustrate the variability in sequence abundances among ASVs and samples inherent in real data. Indeed, the inclusion of abundances provides a measure of robustness to denoising method and parameter combinations. In addition, rarefying data had a similar effect to the applied filtering parameters in that both reduce the number of observed sequence variants by removing low abundant samples and rare ASVs from the analysis.

We applied a Kruskal-Wallis test to determine if diversity estimates for each Hill number among rarefied mock samples were different between groups (where a group was represented by the combination of a denoising method and filtering parameter). There were significant differences among groups for each Hill number:  $q=0$  ( $H(8) = 26.36$ ,  $p \leq 0.001$ );  $q = 1$  ( $H(8) = 30.55$ ,  $p \leq 0.001$ ), and  $q = 2$  ( $H(8) = 26.70$ ,  $p < 0.001$ ). A Dunn’s Test comparing diversity estimates per Hill number was performed for each pairwise combination of denoising method and filtering parameter (**Tables S4-S6**) among mock data. For presence-absence diversity estimates (Hill number  $q=0$ ) each denoising method with “basic” filtering differed from each denoising data with “extra” filtering

parameters. Thus, for binary data types the choice of filtering method is important. However, when incorporating abundance information for diversity estimates (using Hill numbers 1 or 2), choice of denoising method matters. Pairwise comparisons of “basic” filtered DADA2 or Deblur data were different from VSEARCH, and that “standard” and “extra” filtering DADA2 estimates differed from VSEARCH estimates. However, these differences in diversity estimates for  $q=1$  and  $q=2$  were minor among denoising methods, and the observed differences associated with VSEARCH reflect the fact that there were fewer than the number of expected sequence variants because this method clustered a few distinct expected sequence variants of a single OTU. We also applied the same nonparametric tests to rarefied bat guano data. There were significant differences among groups for each Hill number for Kruskal-Wallis tests:  $q=0$  ( $H(8) = 1121.1$ ,  $p \leq 0.001$ );  $q = 1$  ( $H(8) = 211.2$ ,  $p \leq 0.001$ ), and  $q = 2$  ( $H(8) = 138.5$ ,  $p < 0.001$ ). Subsequent Dunn’s Tests among guano data broadly suggested that there are fewer significant differences between groups with increasing Hill number, indicating that the inclusion of abundance information reduces the main effects of denoising method and filtering parameter (**Tables S7-S9**).

To explore the effects of denoising method and filtering parameters on community composition between rarefied samples we used three distance estimates: Dice-Sorensen uses unweighted (presence-absence) data, while Bray-Curtis and Morisita-Horn measures incorporate weighted (abundance) data. Among mock samples, the main effects for denoising method ( $F_{(2,27)} = 1.973$ ,  $MSE = 0.024$ ,  $p = 0.001$ ), filtering parameters ( $F_{(2,27)} = 2.704$ ,  $MSE = 0.033$ ,  $p = 0.011$ ) and their interactions ( $F_{(4,27)} =$



3.276, MSE = 0.041,  $p = 0.001$ ) were significant for the Dice-Sorensen distance measure (**Table S10**). Among abundance-based distance metrics, significant main effects were observed only for filtering parameter for Bray-Curtis ( $F_{(2,27)} = 3.050$ , MSE = 0.349,  $p = 0.04$ ) and Morisita-Horn ( $F_{(2,27)} = 3.054$ , MSE = 0.02,  $p = 0.036$ ) diversity estimates (**Tables S11-S12**). We chose a subset of bat guano samples collected at the same location across several months, and therefore added the collection date as an additional main effect to the model estimating beta diversity. We found that distance estimates for real data were robust to the various bioinformatic criteria. Main effects of filtering parameter and ( $F_{(2,710)} = 1.582$ , MSE = 0.570,  $p = 0.006$ ) collection date ( $F_{(3,710)} = 40.149$ , MSE = 14.464,  $p = 0.001$ ) were significant for the Dice-Sorensen incidence distance measure (**Table S13**). Only the main effect of collection date was significant for both Bray-Curtis ( $F_{(3,710)} = 22.013$ , MSE = 9.382,  $p = 0.001$ ) or Morisita-Horn ( $F_{(3,710)} = 23.325$ , MSE = 9.822,  $p = 0.001$ ) distance measures (**Tables S14-S15**). Ordination by NMDS of the select guano data illustrate that all three distance measures produce similar clustering of data according to the month in which samples were collected (**Figures S8-S10**). As with previous studies examining the effects of quality filtering on subsequent diversity measures of microbial communities<sup>65</sup>, our data indicate that beta diversity estimates are less sensitive to filtering criteria than equivalent alpha diversity estimates.

### *Database Construction*

Assigning taxonomic identity to a dataset of ASVs is limited by the collection of available references. We compared how the composition of arthropod COI records varied among

three databases: two created from BOLD records (Palmer and tidybug) and one from GenBank records (Porter). While the Porter (1,280,577 total COI records; 515,780 arthropod-specific COI records) and Palmer databases (1,617,885 total COI records; 1,565,831 arthropod COI records) contained arthropod as well as non-arthropod COI records, the tidybug database was constructed exclusively with arthropod COI records yet contained the largest number of distinct sequences overall (1,841,956 arthropod COI records). The quality of a database is not only a function of how many records it contains, but also by how complete the taxonomies are for the references. For instance, references may or may not contain a name at the family, genus, or species rank. With respect to the number of unique taxa and unique sequences we found the tidybug database to contain more distinct records from species through order levels (**Figure 6, Tables S16-S17**). In addition, we evaluated the number of shared taxa (found in two or more databases) and unique taxa (found in only one database) at the family, genus, or species rank to determine how distinct the records were among the three databases (**Figure S11**). While all three databases shared many species records in common (39,068), it was in fact the tidybug database contained the most unique species records (49,822). The Porter dataset (20,542 species records) and Palmer dataset (4,845 species records) contained fewer distinct records overall. There were also more distinct genus and family taxa, although these differences were largely confined to comparing databases created using BOLD versus GenBank.

Database composition can be further impacted by whether or not a set of references are clustered. We clustered our tidybug database at three levels (99%, 97%, and 95%) to

understand the effects of clustering on taxonomic composition (**Figure 7**). Even a minor amount of clustering dramatically reduces the original number of dereplicated sequences from 1,841,946 to 407,356 in the 99% identity-clustered datasets. Further clustering at 97% (265,885 records) and 95% (215,055 records) results in additional reductions in the number of representative sequences. While clustering can speed up the process of classifying a dataset, it can result in loss of meaningful biological information in a database. In situations where two or more sequences that share disparate taxonomic information are clustered, a least common ancestor (LCA) process is invoked to determine the new reference cluster taxonomy. Consequently, taxonomic information is lost should that LCA process apply a consensus approach (instead of a majority approach). For example, if a cluster is derived from two sequences that differ at the species rank, the resulting taxonomy for that clustered representative sequences would be 'ambiguous' at the species rank even when the two ancestral sequences are known. Our tests indicate that clustering both reduces sequence diversity and taxonomic diversity, and that as the percent identity for a given cluster is reduced, the proportion of ambiguous taxa (i.e., unassigned taxa) in a dataset is increased.

Clustering at a fixed threshold can also skew the resulting sequence diversity and taxonomic information for groups of taxa with distinct evolutionary rates (**Figure S12**). For instance, the dipteran order contains more than twice as many unique sequences (740,201) in the dereplicated tidybug database compared to the lepidopteran records (349,346). Yet, after clustering at even 99% identity, the Lepidoptera contains more unique records (114,285) than Diptera (97,708), and this disparity increases as the

clustering radius is reduced. While clustering generally reduces the amount of taxonomic information retained for all groups of organisms, some arthropod orders are more sensitive than others to its effects. For example, 65% of psocodean references clustered at 100% identity contained species names, while just 40% of the psocodean records clustered at 95% identity contained species information. However, among tricopteran records the difference between 100% identity and 95% identity clustering results in a minor change from 82% to 75% of records containing species names. Database construction that demands further clustering would likely benefit from investigating flexible clustering radii that reflect the evolutionary diversity of taxa.

### *Classification Comparisons*

We used Taxonomic Accuracy Rate (TAR) and Taxonomic Discovery Rate (TDR) as modified metrics of precision and recall, respectively. These measures had been previously described to quantify rates of precision and recall at the community level, rather than the per-sequence level <sup>67</sup>. TAR measures the ratio of true positives to all (true and false) positive results, and therefore reflects the proportion of observed taxa that are expected. We found that all classifiers agreed at class and order levels: each classifier scored a value of 1, indicating no false positives were detected. The BOLD API scored the highest for TAR at family (0.94), genus (0.94), and species (0.92) levels (**Figure 8**). However, this score is potentially biased by virtue of the challenge in accurately defining the expected taxonomies. There were three instances where the expected taxonomy is undefined for a mock taxa name at family through species levels while the non-BOLD classifiers (BLAST, VSEARCH, Naive Bayes, SINTAX) proposed

identical taxonomic names at each level. Thus, these false positives may reflect an unnecessary penalty: for instances where a taxonomic name is lacking because they were not documented by the researcher, classifiers that assign taxonomic information are not necessarily incorrect, but they are penalized as overclassifying. Because we do not know exactly what the family, genus, or species names are for this mock specimen on these occasions (because the specimen could not be accurately classified by experts or full length COI Sanger sequences), it is impossible to determine with this mock community whether these classifiers are misclassifying those taxonomic ranks, or if these classifiers are assigning the proper taxonomy for those species even though a false-positive is registered.

We used TDR to measure the proportion of expected sequences that were observed. TDR measures the ratio of true positives to the sum of the true positives and false negatives, therefore scores closest to 1 reflect classifiers that tend to assign the correct expected taxonomies while reducing the instances in which expected assignments are left unassigned. The BOLD classifier scored lowest at class (0.96), order (0.83), and family (0.77) levels for TDR. In fact, BOLD was the only classifier to have any false negatives at class and order levels (scores for all other classifiers at these levels equal 1), and indicates that the BOLD classifier assigned fewer taxonomic names for mock samples compared to alternative classifiers. Naive Bayes scored highest for TDR at the genus level (0.9), while BLAST and VSEARCH had the highest TDR scores at the species level (0.87), indicating that no single classifier is superior at all taxonomic levels using standard parameters. Further optimization of the classifiers was beyond the scope

of this study, and is an area that warrants further attention for animal metabarcoding experiments. For example, lowering the SINTAX bootstrap confidence cutoff value from 0.9 to 0.8 would retain more expected Species names in the mock dataset, however, this may also lead to an increase in the number of false positives. This tradeoff in precision and recall is expressed as the F measure, and indicates that the QIIME 2 classifiers perform best from class through family levels (**Figure 8**). As mentioned previously, the highest TAR scores were consistently attained by the BOLD API but may reflect a bias of over-classifying by non-BOLD classifiers in the few instances where an incompletely described mock community does not depict the expected ground truth. Further tests using vouchered specimen with completely described taxonomic identities are clearly needed to better evaluate the strengths and weakness of available classifiers.

Classifier performance was also compared on a relative basis using bat guano data. We were interested in understanding two phenomena: first, the frequency with which ASVs were assigned a taxonomic name from class through species levels, and second, the degree with which one or more classifiers assigned the same taxonomic name to that ASV. While there is no ground truth to these particular assessments, this comparison allows for the potential discovery of patterns among the classifiers. For instance, do classifier types agree more with each other (e.g. kmer-based vs. alignment-based)? We found that fewer sequence variants were assigned class or order names by alignment-based VSEARCH (61% class; 61% order) or BLAST (62% class; 62% order) compared to BOLD (99% class; 83% order) or kmer-based SINTAX (91% class; 77% order) or

Naive Bayes (96% class; 81% order) classifiers (**Table S19, Figure S13**). The number of named ASVs are reduced from family through species levels, though no one classifier was assigned the most taxa at each level: Naive Bayes assigned the most taxa with family names (68%), BOLD assigned the most taxa with genus names (61%), and SINTAX contained the greatest number of taxa with named species (42%). The relatively larger fraction of taxa with assigned family, genus or species names may reflect differences in classification processes (i.e. kmer vs. alignment based), but at least part of the reduction in named taxa for VSEARCH or BLAST is explained by the fact that these alignment-based classifiers apply an LCA process after identifying potential matches while kmer-based approaches do not. More expansive mock communities as well as in silico tests will be useful for identifying which classifiers are best served to balance the desire to classify as many taxa at as many levels as possible without introducing false indications of certainty where taxonomic ambiguity is more appropriate. Finally, the BOLD classifier reported the lowest number of named species (35%) after we applied an LCA process. Because their API required using a closed source reference database, it is unclear whether the lower number of records is a function of the classifier itself or the database queried. The BOLD API has not published the parameters of the classification process, thus it is not clear what parts of that algorithm are likely behind the fewer classified records.

We assessed 31 distinct ‘sets’ to compare the various ways the five classifiers may converge upon a proposed taxonomic assignment for a particular ASV at the class, family, or species levels: one set in which all five classifiers contain the same

information; five sets in which four of the five classifiers agree and one differs; 10 sets in which three of the five classifiers agree and one or more differ; 10 sets in which two classifiers agree and one or more differ; five sets in which a classifier uniquely ascribes a taxon in which all other classifiers fail to provide information (**Figure 9**). The largest fraction of ASVs converged on a similar classification at all taxonomic levels: 59% of all class names, 57% of orders, 46% of families, 43% of genera, and 32% of species (in the case of family, genus, and species, more ASVs were unnamed than named). However, the degree with which differences existed between classifiers increased from class to family to species levels. In particular, we found that the four classifiers using the tidybug database assigned many of the same species (21% of named species) and family-level names (23% of named families) when BOLD failed to provide any taxonomic name at that level. This was in contrast with just 4% of named taxa being common to the non-BOLD classifiers when BOLD did not classify the same ASV. These results indicate that differences between BOLD and non-BOLD classifiers are partly attributed to potential differences in database composition. However, note that our database was derived from BOLD references, but because no information is available on the particular records the BOLD API uses, we could not determine whether these differences were due to the classifier or the references. Nevertheless, classifier-type trends also emerged among those using our common tidybug database. Kmer-based classifiers assigned more information where alignment-based classifiers did not: this effect was most pronounced at the class level with SINTAX and Naive Bayes classifiers classifying 2,875 ASVs that BOLD, VSEARCH, or BLAST did not (22% of all unique taxa). Likewise, alignment-based classifiers converged on common species-level



names that BOLD or kmer-based classifiers either differed in assignment (114 ASVs, 2% of all taxa) or failed to assign any name (272 ASVs, 4% of all taxa). Species names unique to a classifier were most frequently assigned by Naive Bayes (698 ASVs, 10% of dataset).

These findings suggest that classifiers using the same database can differ substantially with respect to the degree of taxonomic completeness an ASV is assigned. Kmer-based approaches produce more species-level assignments than alignment-based classifiers. However, the reference database used may also impact taxonomic identities. The second largest set of shared taxa names consisted of all non-BOLD classifiers. For example, 20% of all named family-level ASVs were common to all non-BOLD classifiers but not to taxa classified by the BOLD API. These results suggest that either the BOLD API is fundamentally different than these other classifiers, or, that the database accessed by the API is distinct from the tidybug database the other non-BOLD classifiers used. Public documentation of the BOLD Taxonomy API would be required to make such a distinction.

## **DISCUSSION**

### *Sequence filtering and diversity estimates*

As with wet-bench work undertaken in metabarcoding studies, there are many bioinformatic decisions that can impact estimates of diversity as well as the specific taxa

identified among detected sequence variants. Moreover, the particular molecular steps of an experiment are inextricably linked to the bioinformatic processes; poor decisions made in either realm reduce accurate scientific inferences. In addition, the expanding availability of computational tools coupled with the lack of a community standard for their usage makes comparing studies within even highly related disciplines challenging. Further complicating matters is the fact that many animal diet analyses transform read abundances into presence-absence frameworks. Our work was motivated to better understand how several of these common bioinformatic decisions affect interpretations of diversity and classification. Indeed, even if a researcher was to execute an ideal molecular workflow, the computational parameters invoked have the potential to alter the number of sequence variants observed depending on the denoising method of choice, and even more strikingly, completely ignore particular taxa depending on the classification strategy or choice of database.

The first major bioinformatic decision that has yet to reach consensus in practice among animal diet metabarcoding studies relates to identifying unique sequence variants. Because we standardized the baseline filtering assumptions, the observed differences of expected and unexpected sequences within the mock samples reflect differences in the denoising method. As has been shown for microbial studies (for which these methods were developed), denoising methods appear advantageous over OTU clustering both for retaining expected sequence variants and reducing spurious sequence variants. With respect to recalling the expected sequence variants among our mock community data, both DADA2 and Deblur worked similarly well. DADA2 and

Deblur detected all 24 expected sequences, while VSEARCH failed to detect 1–4 expected mock members among the four sequencing runs. For DADA2 and Deblur datasets, 22 or 23 of these were exact matches, with the remaining matches identified as “partial” alignment. The “partial” designation more likely indicates incorrectly assigned reference sequences than sequencing error itself for two reasons: first, these particular partial matches were the same sequence across all four sequencing runs, and second, they generated a similar abundance of reads as the “exact” mock sequences. The missing “exact” matches among the VSEARCH data highlights a problem inherent with clustering: loss of information due to shared sequence similarity. Two of our mock sequences were variants of the same species; one each of the exact variants were clustered together in the VSEARCH library. Because the most abundant sequence acts as the centroid when clustering, other missing “exact” matches were replaced with more abundant erroneous sequences as “partial” matches. For example, the “exact” sequence for mock sample IM44 was detected in libraries B and C, yet not present in libraries A and D. However, another sequence with over 99% identity to that exact match was present in libraries A and D as a “partial” match. Because the variant of this sequence present in libraries A and D was never identified in either DADA2 or Deblur, it suggests that this is a sequencing error and not the result of our mock community containing multiple distinct variant templates to amplify (as in the case of *H. axyridis*). Without the capacity to model sequencing error, a clustering approach such as VSEARCH is inadequate for two reasons: first, it generates more false positives, and second, it reduces the number of expected sequences when a sequence error variant is more abundant than the actual sequence.

Unlike microbial studies, relatively few animal metabarcoding studies incorporate read abundance information in their diversity estimates because of uncertainty in associating biomass with sequence counts <sup>19,26</sup>. This is particularly meaningful given that our observed differences among denoising methods performance with “basic” and “standard” filtering is entirely attributed to low-abundance sequences. Thus, the choice of denoising method as well as the filtering parameters invoked can dramatically change interpretations of richness for researchers using a presence-absence framework. For instance, 539 of the 604 unexpected ASVs contained 10 or fewer total sequences per ASV in just a single mock sample (“libD”) when processed with “basic” filtering using VSEARCH. Deblur follows a similar pattern, retaining many sequence variants with extremely low abundance: the same “basic” filter dataset from that same sample contains 525 total variants, yet all 364 “miss” variants contain 11 or fewer sequences per ASV. Yet only after applying the “extra” filter were all unexpected sequences removed from each sample, and every denoising method required had at least some unexpected ASVs removed by this filter. Discarding observed sequence variants with low abundances on a per-sample basis or on a per variant basis operates with the assumption that rarity is more indicative of sequencing error than true biological variation. For experiments that do not include positive controls with their true samples, one is left with the challenge of arbitrarily assigning what minimum read count to use. Our mock data suggest that DADA2 is most effective at eliminating low abundance errors and would be particularly useful in situations whereby no additional filtering threshold can be empirically derived. Suggestions to apply some fixed value have been

previously validated for clustering approaches <sup>65</sup> but are not necessarily encouraged for denoised data. We also found that COI datasets processed with a clustering approach absolutely required further filtering. However, removing low frequency sequences invalidates the use of some classic (e.g., Chao1 <sup>116</sup>) and newer modeling approaches that leverage singleton and low-abundances counts for estimating richness <sup>117</sup> and other diversity measures <sup>118</sup>. Given the widespread availability and documentation of denoising tools, we see no reason to cluster sequence data. As with most bioinformatic software currently utilized in animal metabarcoding sequence analyses, these platforms were initially designed and tested on microbial datasets. We recommend that additional animal marker datasets—particularly synthetic mock communities that more accurately reflect the diversity of the study—to understand how software parameterization affects performance.

Because of a dependence on transforming read abundances to presence-absence, reports of animal metabarcoding compositional similarity have historically been limited to richness for inter-sample similarity and one of a few indices (e.g., Dice-Sorensen or Raup-Crick) for intra-sample analyses. Incidence-based approaches are often justified as a more appropriate choice compared to using relative abundances of sequences due to the challenges of associating counts to biomass <sup>18,26</sup>. That is, the number of reads for a sample often is not a reliable measure of the abundance of that specific arthropod. Yet as Deagle <sup>26</sup> observes, *“to accept the notion that relative sequence counts provide no meaningful information would mean that, within one sample, a few DNA sequences from one food taxon are equivalent to 10,000 sequences from another”*. Our analyses of

bat guano data suggest that read abundances are indeed completely different than the proportions of sequences in our mock samples. In contrast, real metabarcoding data is better defined as a sparse matrix, with ASV abundances that can vary substantially among samples and a distribution of taxa that is highly uneven. Researchers do not have to choose between including or excluding abundance information in diversity estimates; both articulate related, yet distinct aspects of the data. Unweighted estimates of observed richness, for example, can protect overestimating the importance of taxa that may be easier to amplify or extract, while abundance information can help reveal which ASVs are common or dominant in a dataset. Thus both detections and relative abundances provide important insights when making inferences about the diversity of samples. As reviewed extensively by Jost <sup>119</sup> and Chao <sup>99</sup>, Hill Numbers provide a mechanism to unify measurements of richness, evenness, and diversity. Such a mechanism enables researchers to estimate the effective number of species necessary to produce the observed diversity on a continuum whereby read abundances are increasingly more relevant to the diversity estimate itself. This framework does not resolve the challenge of relating biomass to read abundances, but it does allow the researcher to investigate the degree with which read abundances contribute to their observed diversity estimates.

Alpha diversity estimates were sensitive to both denoising method and filtering parameters for presence-absence data but generally robust to these parameters when read abundances were included in the model. Among rarefied mock community samples, DADA2 performed well for all pipeline settings while Deblur required additional

filtering to accurately estimate the expected number of observed sequence variants. Diversity estimates weighted with abundance information performed similarly for all combinations of parameters, with the exception that VSEARCH failed to detect a few expected sequence variants. In contrast, estimates of beta diversity for mock samples and the real bat guano dataset were generally robust to the denoising method used regardless of whether rarefied read abundance information was used in the model or not. Similar findings have been reported for microbial studies <sup>65</sup>. We also compared alpha diversity estimates using unrarefied data and found that relatively more sequence counts were needed to obtain reliable richness estimates for Deblur and VSEARCH than the equivalent sample processed by DADA2. However increasing the sampling depth for real data can result in dropping many more samples from the dataset, thus most researchers invoke some level of filtering. Applying the “standard” filtering parameter is highly effective at reducing this sampling depth concern for Deblur data, however among VSEARCH mock samples the asymptotic estimate of diversity appears to monotonically increase. This suggests that when low abundance samples are removed but no minimal threshold of per-sample read abundance filtering is applied, richness estimates are confounded by the sequencing depth due to the positive relationship between throughput and sequence error. Thus, while presence-absence data may alleviate complications with how to interpret counts, the process itself can be highly sensitive to the bioinformatic pipeline used unless per sample read abundances are normalized in some fashion.

### *Database construction*

Assigning taxonomic information to a set of sequence variants requires a reference database, yet there is no single curated and versioned resource that is widely adopted among animal diet metabarcoding studies. This makes comparing the interpretations among experiments that use different database resources to classify representative sequences extremely difficult. In contrast to the way that BOLD is currently managed, microbial databases for Greengenes <sup>82</sup>, SILVA <sup>83</sup>, and UNITE <sup>120</sup> are versioned. Thus when a researcher desires to evaluate multiple studies completed at different dates using different versions of a database, the older experiments can be updated with newer databases for more equitable comparisons. While BOLD had previously presented packaged versioned releases, this practice ended in 2015. To complicate matters the BOLD API does not provide any mechanism to filter records by the date of collection, though many records lack this information altogether. Motivated by the fact that these same complications persist for users relying on NCBI resources, Porter and Hajibabaei <sup>121</sup> created a pipeline that makes versioned releases manageable. However, their database construction choices may reflect the needs of their experiments, and may not be best suited to other COI projects. For instance, their requirement that all records contain species-rank names prioritizes taxonomic information over sequence diversity. Other projects may benefit from a more diverse database that contains records that lack species or even genus-rank names. While no single database will be sufficient for all animal metabarcoding projects, versioned resources are essential for ensuring that unique properties between experiments reflect differences in biology and not the reference databases.



Versioned databases must also thoroughly document the decisions made regarding their construction. We focused on three major decisions: filtering references with taxonomic ambiguity, dereplicating sequences, and clustering related sequences. Taxonomic ambiguity differed between the Porter references versus the Palmer and tidybug databases. The Porter database consists nearly exclusively of full taxonomic identities—references that include species-rank names. This likely explains why there are almost half the total number of unique sequence records available between the Porter database compared to either the Palmer or tidybug databases, both which allow varying levels of taxonomic ambiguity in their records. In fact, the Porter dataset was constructed initially to contain only records with named species, however, by failing to dereplicate their records, there are over ten thousand instances where identical sequences contain distinct taxonomic identities. In this case, the LCA algorithm we applied to their database reduces these records to a common shared taxonomic level, eliminating species-level information. This highlights that dereplicating will in general reduce the total number of available records when a consensus LCA process is applied. For example, our ‘original’ BOLD arthropod records contained over 3.1 million sequences, while just 1.8 million of these references remained in our dereplicated dataset. Dereplication is essential to database construction, and we opted for a consensus LCA approach instead of a majority method to avoid potentially over-classifying the reference records.

Clustering reference sequences is routine among microbial databases and was introduced to COI references with the BOLD barcode index number (BIN) <sup>122</sup>. The

process is performed to alleviate the computational burden of classification by reducing the number of highly similar reference sequences. Yet a tradeoff occurs between computational burden and compositional representation. Clustering by decreasing percent identities creates fewer groups of distinct sequences, thus reducing the number of records needing to be searched. This makes searching a database faster, but the number of potentially distinct records—distinct in terms of both sequence identity and taxonomy—are fewer. An additional problem with clustering arises from the fact that groups of taxa may have distinct evolutionary rates and thus are differentially impacted by applying a single value when related sequences are merged. For example, there is greater variation in COI sequence in Coleoptera than Lepidoptera<sup>30</sup>. However the degree with which this variation exists among the millions of COI sequences and its effects on taxonomic classification is not well understood. We found that clustering reduced the number of the most abundant arthropod orders in a non-uniform manner. For example, while there are twice as many unique dipteran sequences in the tidybug database as the next most abundant order (Lepidoptera), clustering at 99% identity resulted in fewer overall dipteran representative sequences than lepidopteran. If clustering is a necessity, the dynamic clustering approach used by UNITE<sup>120</sup> may be preferred to the fixed binning approach currently, but it remains unresolved exactly what clustering radii are appropriate for each taxonomic level.

Overall we found that the relative amount of taxonomic information retained at a given rank did not change with clustering, but clustering did increase the overall fraction of ambiguous taxa in a dataset. While some ambiguity is unavoidable when dereplicating,

the reduction in sequence and taxonomic diversity following clustering do not appear worth the benefit of the added time to classify a dataset. We favor filtering modest taxonomic ambiguity (removing all references that lack family-level information) and applying a consensus LCA approach prior to dereplication. This retains substantially more sequence and taxonomic diversity than the species-level requirements of the Porter dataset, while the family-level filtering protects against rare instances where retaining multiple best hits with one reference lacking taxonomic information eliminates the a more information-rich alternative reference sequences. For example, if one best match reference has all taxonomic levels named, while the other best match contains only class-level information, the consensus LCA process produces a final reference that contains only class-level information. No single resource is likely sufficient to represent the various metabarcoding markers, while a combination of resources advocated by Porter et al.<sup>111,121</sup> ensures as diverse a set of reference sequences as possible. Unfortunately the status quo is such that researchers either must possess the computational background to create the exact set of references needed, or default to accessing databases that are either outdated or possibly ill-suited for their project. It would be particularly useful if the versioned releases were themselves managed by the curators of the reference information, with users spanning the various animal metabarcoding disciplines providing input regarding database management.

### *Classification*

Assigning taxonomic identities to representative sequences is often the primary goal in animal diet metabarcoding studies. Yet as with sequence processing, the classification

method adopted and parameters defined therein can alter the biological inferences obtained. These choices are extensive, and begin with the broad classification type employed (e.g. Hidden Markov Models, alignment-based, k-mer based). Other fundamental decisions on how taxa are filtered are equally important though often less well documented. For example, some researchers apply an LCA process <sup>69</sup>, though many do not; some researchers filter acceptable taxa outside of expected geographic boundaries <sup>70,123</sup>, though this decision may preclude many undiscovered taxa; others may choose to drop any species with an ambiguous name like “*sp.*”. Moreover, once a classifier is chosen, optimizing parameters for an experiment requires ground truthing, yet to our knowledge comparing various classifiers and parameters using biological samples with known taxonomic identities exists only for microbial amplicon data <sup>67,124</sup> and microbial <sup>125</sup> and viral <sup>126</sup> metagenomes. Thus, we used COI amplicons derived from both biological mock data as well as bat guano samples compare classifier performance of a commonly used animal marker gene.

We found broad agreement among classifiers for the mock data, although the non-BOLD classifiers tended to classify more samples at each taxonomic level than the BOLD API. Non-BOLD classifiers performed remarkably similar with respect to precision, while the SINTAX algorithm was less likely to specify genus or species-level names to mock samples. Because several mock specimen did not contain complete taxonomic names (9 lacked species names, 4 lacked genus names, and 2 lacked family names), we were not able to discern whether the BOLD classifier was under-classifying taxa or if the non-BOLD classifiers were over-classifying certain mock sequences. Even

among trained taxonomists, arthropod mock communities often are unable to be identified to species level and supporting Sanger data can improve taxonomic resolution<sup>37</sup>. However, we did not want to label specific taxonomic names unless the sequence data was unambiguous. To complicate matters, the BOLD Taxonomy API uses a database of unknown composition, and a classifier of unknown parameters. Thus the under-classification may be part of an internal classification difference, or may be simply because their particular database used during classification lacks that record. Notably, the tidybug database that served as the reference for all non-BOLD classifiers was derived exclusively from BOLD records. Therefore we expect the BOLD API to contain the same information, but because the database that the BOLD Taxonomy API accesses is undocumented, we could not use the same database to compare that classifier to resolve the issue of potential database differences. It would be of great benefit to those who use the BOLD Taxonomy API samples to have more thorough documentation on the exact model parameters of the process as well as a versioned and public release of the specific references within the database that are queried during classification.

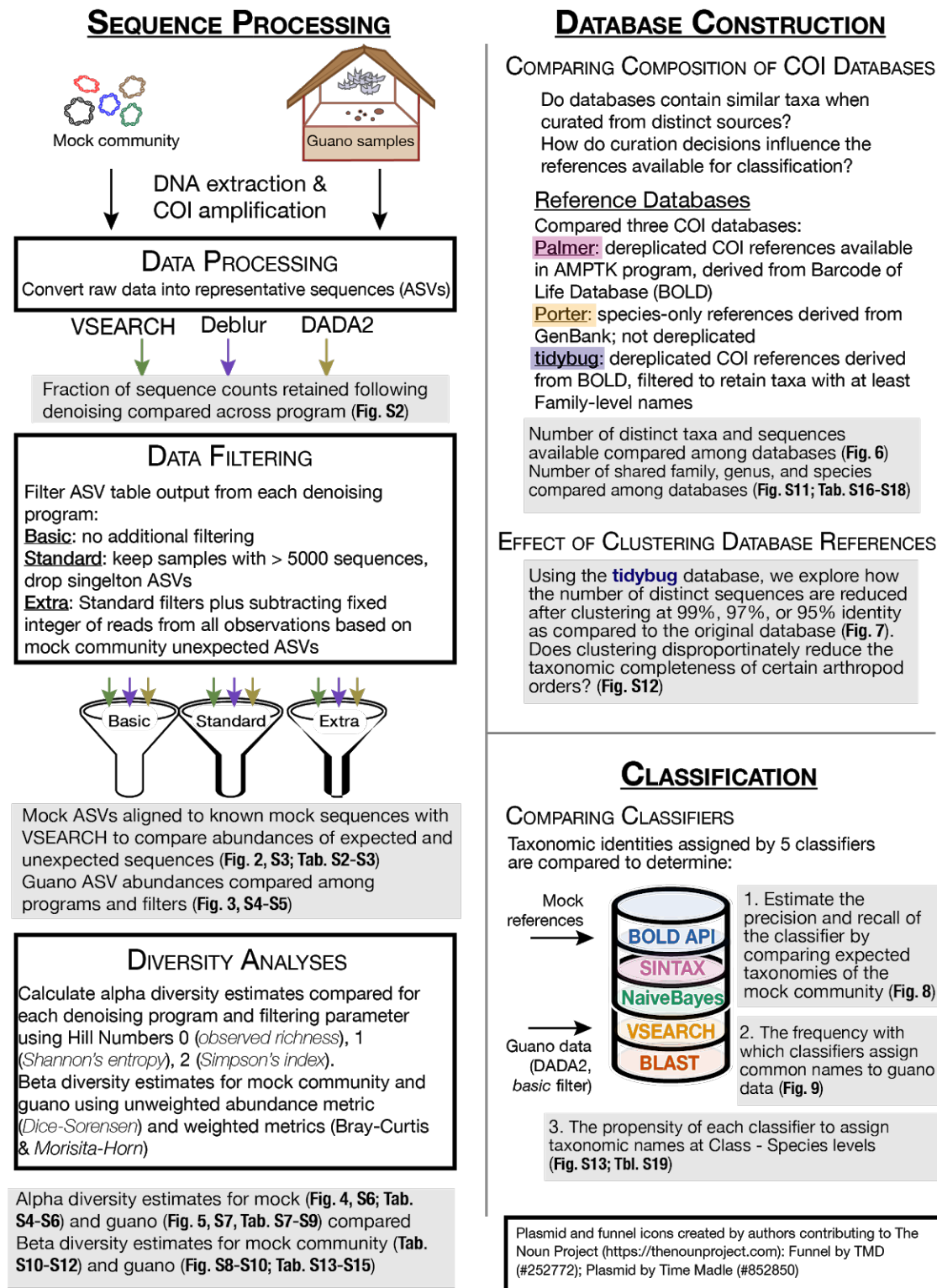
Future classifier comparisons would benefit from requiring that only vouchered specimen with distinct species names are included. Further, given the potential for variation in COI evolutionary rates among taxonomic orders<sup>30</sup>, mock community designs should incorporate both a breadth of diversity as well as substantial overlap within groups to better optimize parameters. For example, if a sliding scale of percent identities is adopted to retain taxonomic information at specific levels (e.g., > 98% for

species, > 97% for genus, > 95% for family), it is imperative to understand whether these fixed values are consistent across the potential taxa in the dataset. In our work with bat guano data we routinely see over 15 arthropod orders, making it unlikely that single fixed value could define all species or genus. While it is unlikely that species-level information can be consistently resolved with short-read amplicons, a growing body of full-length sequences using third-generation technologies <sup>127,128</sup> will improve this resolution, and therefore necessitates a more detailed understanding of whether fixed percent identity boundaries are sufficient for classifying to a particular taxonomic level.

Kmer-based classifiers offer an orthogonal method to classify samples without demanding a fixed percent identity. Indeed, the default method for some amplicon pipelines like AMPTK use a combination of global alignment and kmer-based classifiers <sup>61</sup>. We found that kmer-based classifiers nearly always assigned class and order names to our bat guano data, while alignment-based approaches often returned equivalent ASVs as undefined. Lowering the percent identity threshold from our conservative value of 0.97 to something less would undoubtedly retain more of these undefined taxa. However, because we apply an LCA process to all hits retained, there is an inherent tradeoff between precision and recall. For an alignment approach, lowering the percent identity may result in fewer undefined ASVs (increasing recall), but the proportion of ASVs with species-level information will also be reduced (decreasing precision). Thus, kmer-based approaches offer an alternative to this problem in situations in which far less of a full sequence is needed to assign some degree of taxonomic identity. However, they can also be problematic in that a small section of that sequence may be

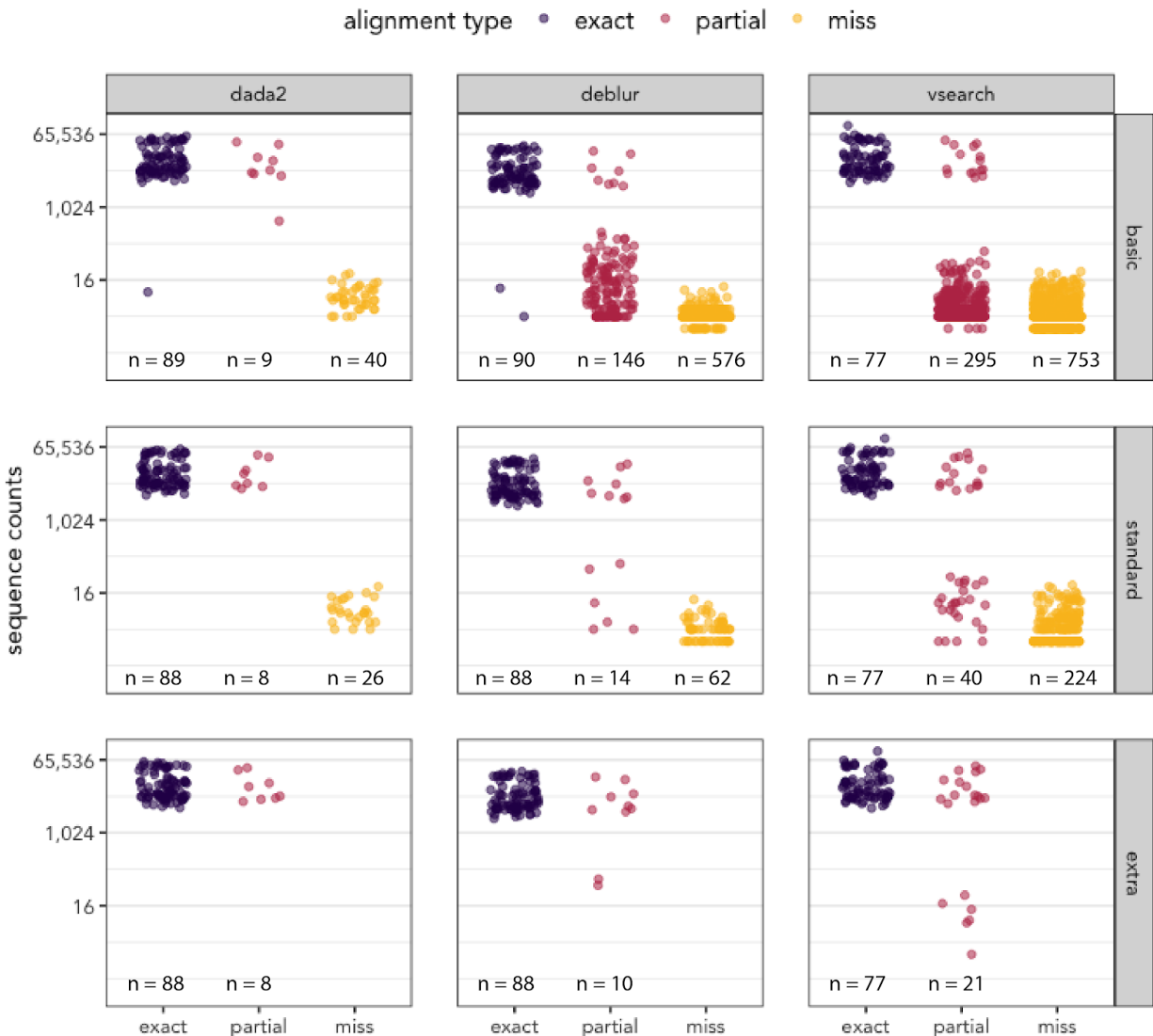
sufficient to classify a chimera or artifact sequence. Overall our comparison of classifiers with bat guano data suggest that substantial differences exist between the BOLD Taxonomy API and all other classifier approaches, particularly at family through species levels. Given the central role of BOLD in many animal metabarcoding studies, we hope future researchers will provide additional comparisons with both biological mock data and other animal diet samples. Among non-BOLD classifiers we find that a kmer-based approach can be especially useful at retaining class and order information that alignment-based classifiers would otherwise discard, but caution that specific sequence pre-processing for amplicon length, quality, and chimera filtering are necessary.

## FIGURES

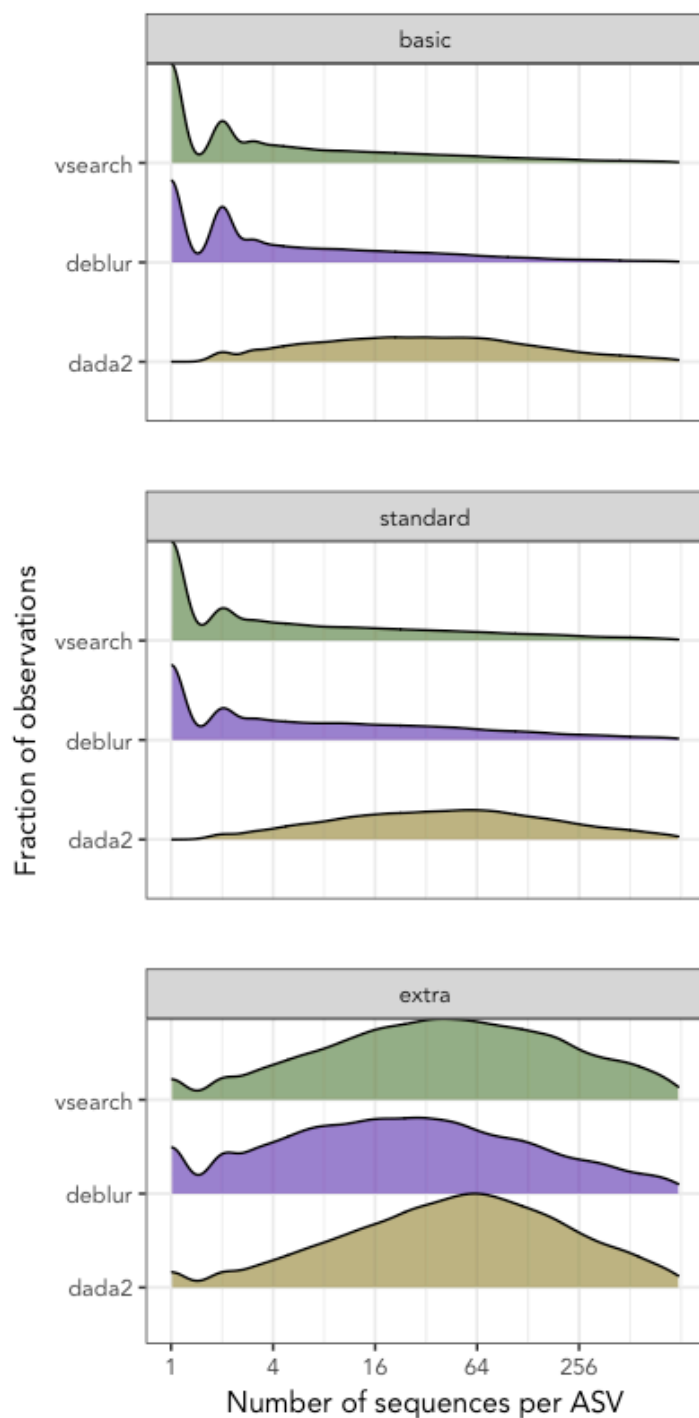


**Figure 1.** Methods and analysis workflow. See Appendix for supplementary figures and tables.

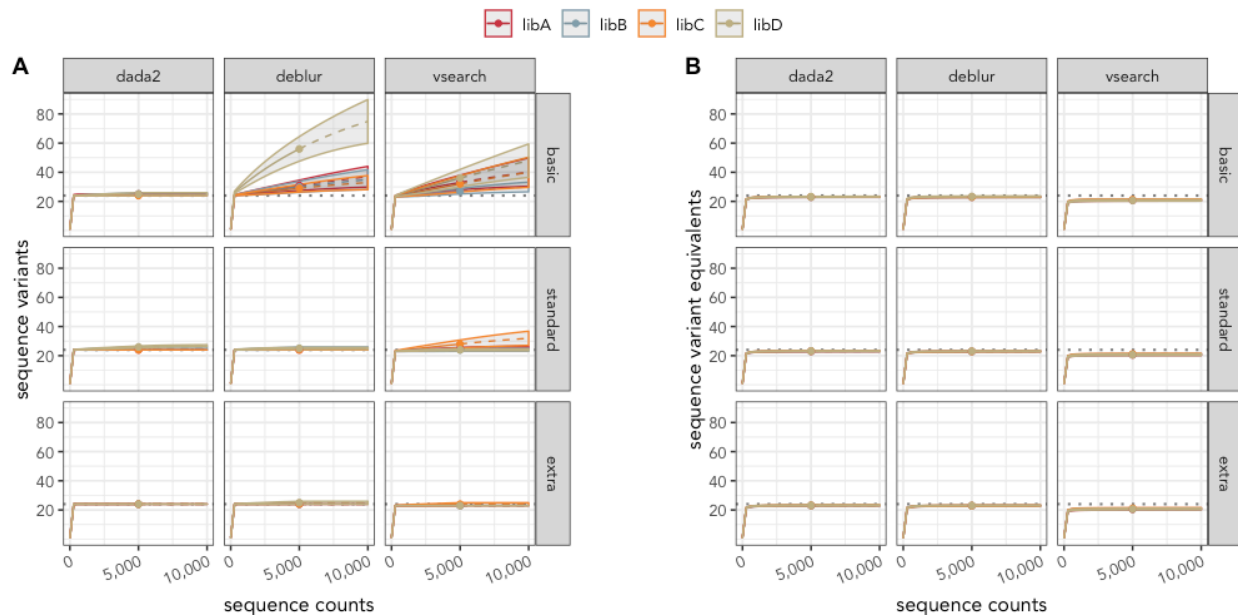




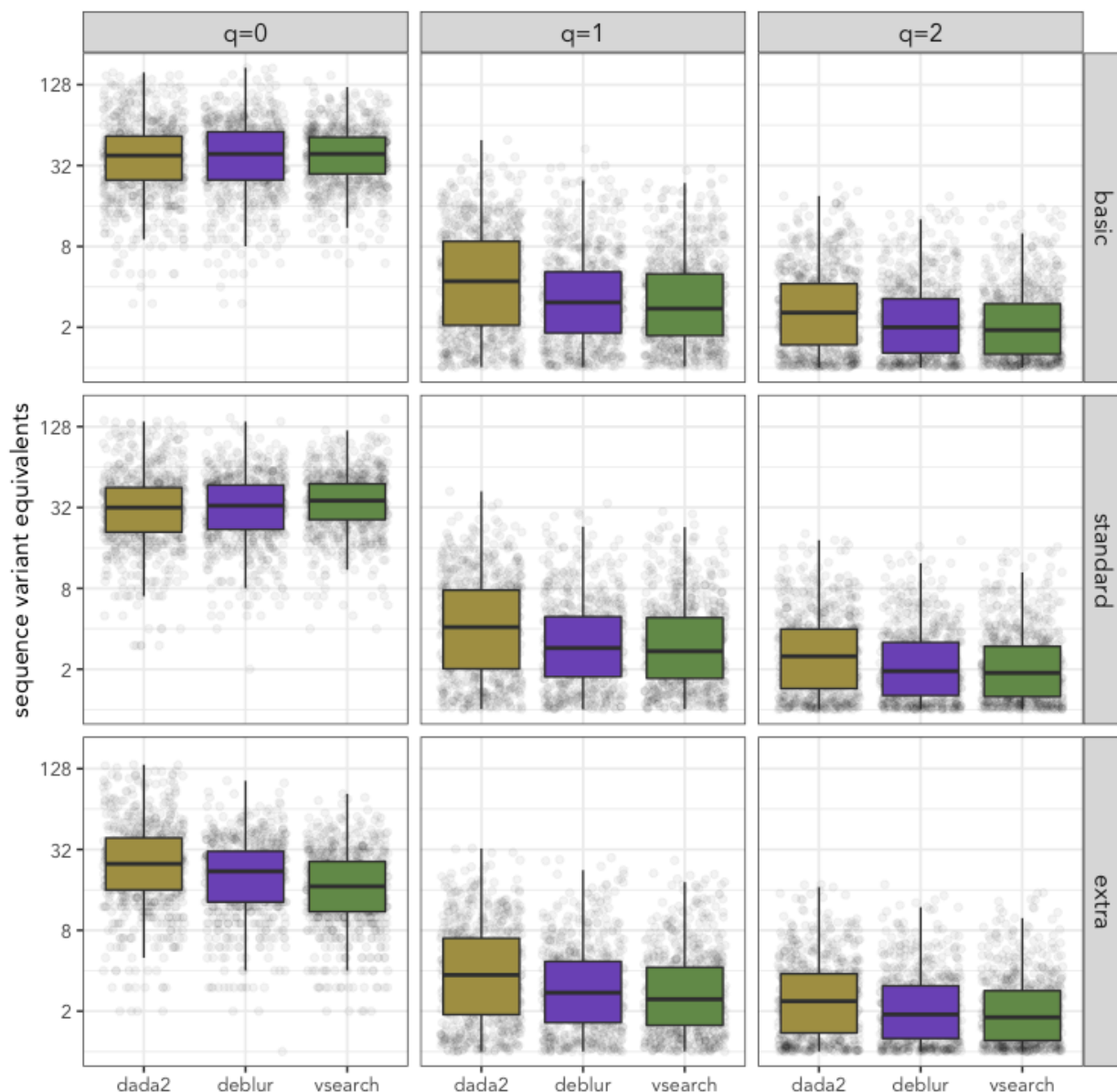
**Figure 2.** Denoising program and parameter evaluations of mock data. Denoising pipeline (vertical facets) and filtering parameters (horizontal facets) result in distinct number of inexact matches among expected mock community sequences. Values beneath each dot cloud reflect the number of ASVs among all four mock replicates in that group. ‘Exact’ matches reflect 100% alignment identity between ASV detected in mock community and a known mock sequence, ‘partial’ reflects between 97–99.9% identity, and ‘miss’ represents an ASV with less than 97% identity to known mock sequences. All queries must span at least 97% of the reference target. VSEARCH produced the fewest exact matches across all four replicate mock samples and the highest number of unexpected (partial and miss) ASVs. Both VSEARCH and Deblur produced more unexpected matches than DADA2 for unfiltered output. The “standard” filter reduced the number of unexpected sequences, yet only after applying an “extra” filter are all unexpected ASVs removed.



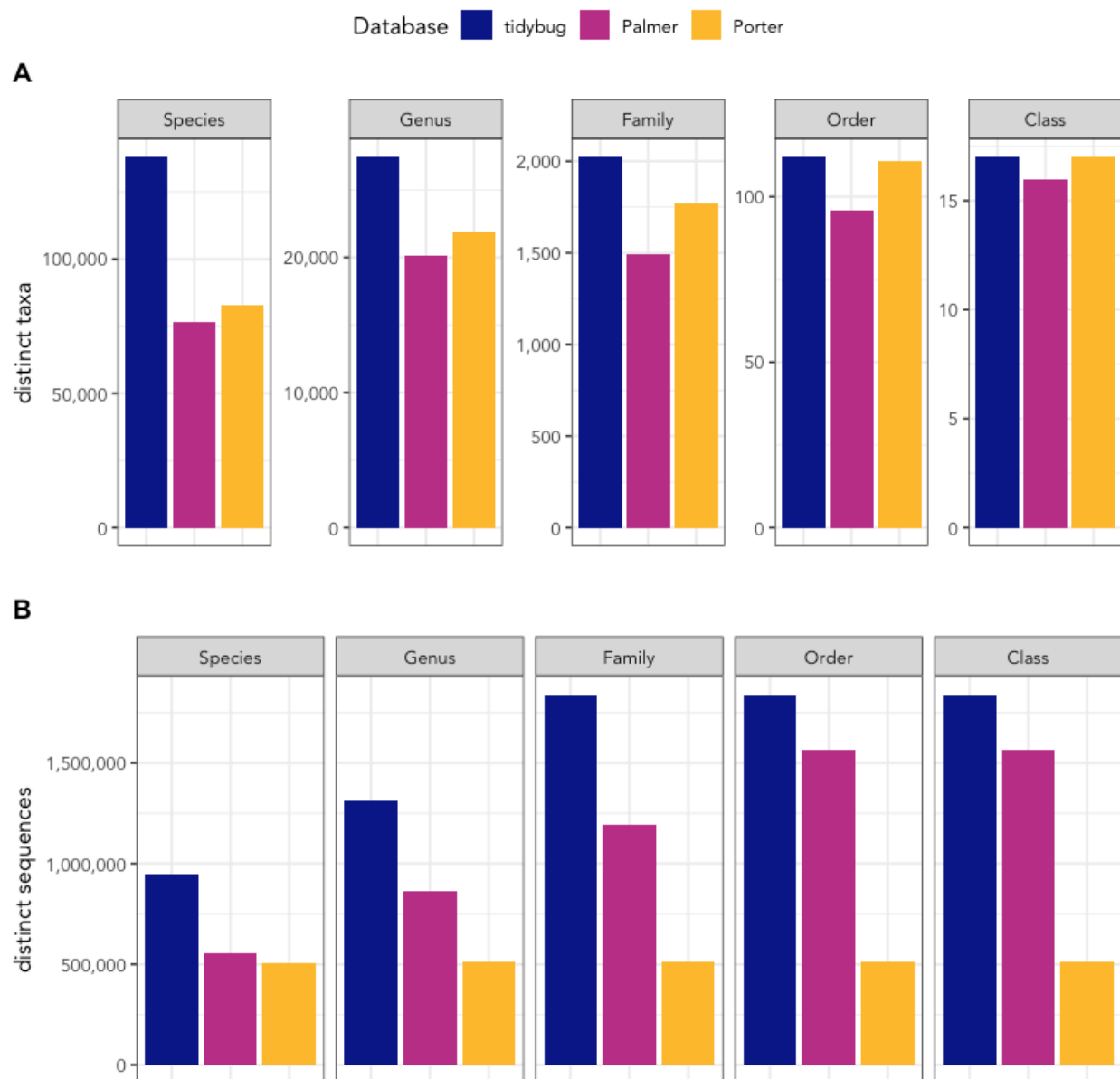
**Figure 3.** Read abundances per sequence variant for bat guano data. Unlike mock samples the more complex communities of guano data are dominated by ASVs with low sequence abundance. DADA2 retains more sequence variants with larger abundances than Deblur or VSEARCH for default outputs and “standard” filter. DADA2 is less sensitive than Deblur and VSEARCH datasets to the removal of low-abundance ASVs by the “extra” filtering parameter. Note some distributions are non-normal because this plot displays the read abundances for four separate sequencing libraries. Only sequence variants per sample with  $\leq 1000$  filtered reads shown, representing between 87–98% of all observations.



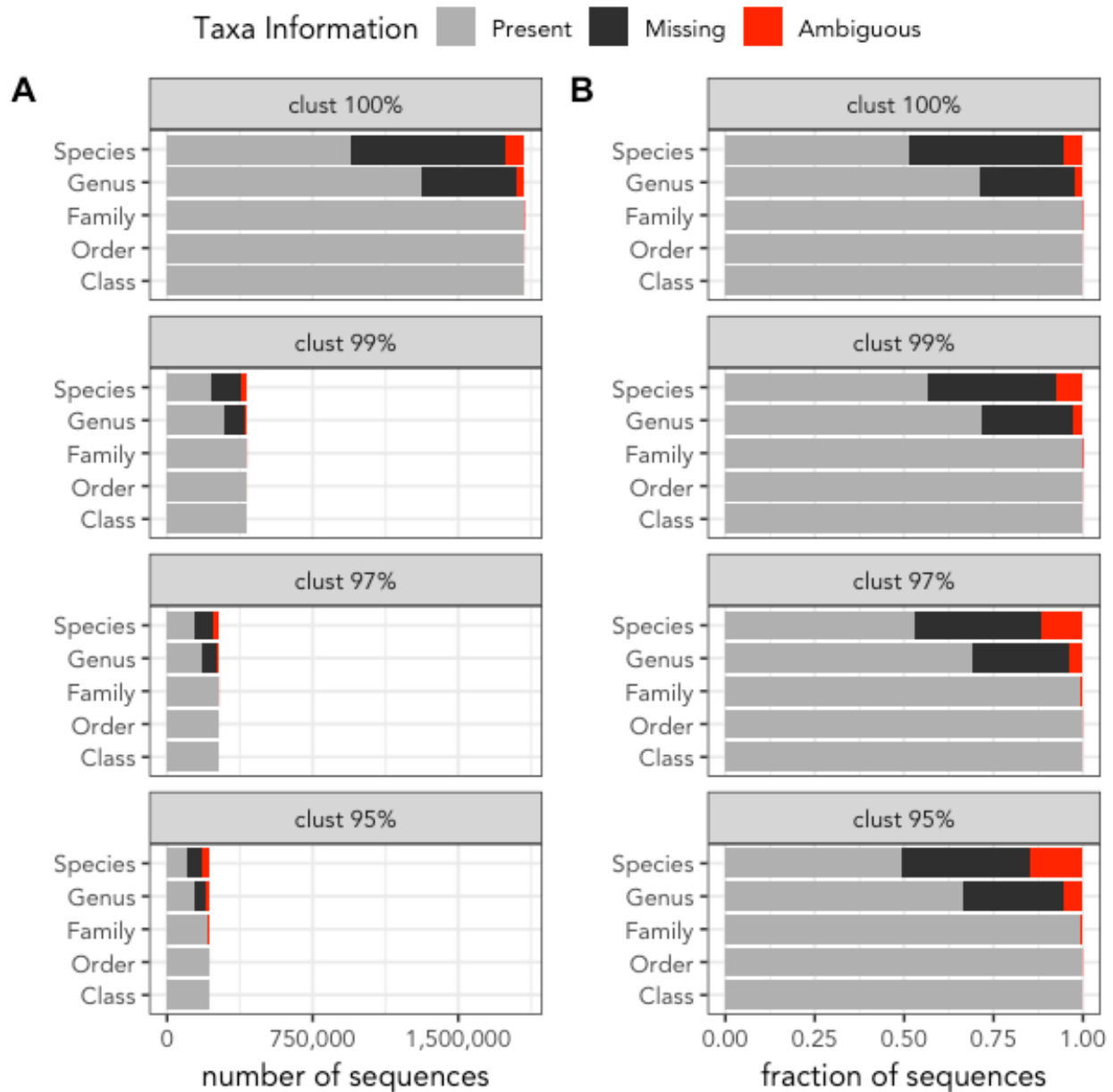
**Figure 4.** Accumulation curves of rarefied mock samples. Interpolated (solid line) and extrapolated (dashed line) diversity estimates are shown for each mock sample relative to expected diversity (dotted line) for each denoising method (vertical facets) and filtering parameter (horizontal facets). **(A)** Estimated diversity is calculated using Hill Number  $q=0$  (richness, or number of observed sequence variants). DADA2 has fewer unexpected sequence variants compared to Deblur or VSEARCH for default output, while imposing “standard” and “extra” filters reduces the differences among denoising methods. **(B)** Estimated diversity is calculated using Hill Number  $q=2$  (similar to Simpson’s 1-D). Differences among denoising methods are negligible among diversity estimates that incorporate abundance information because all the unexpected ASVs are of low abundance.



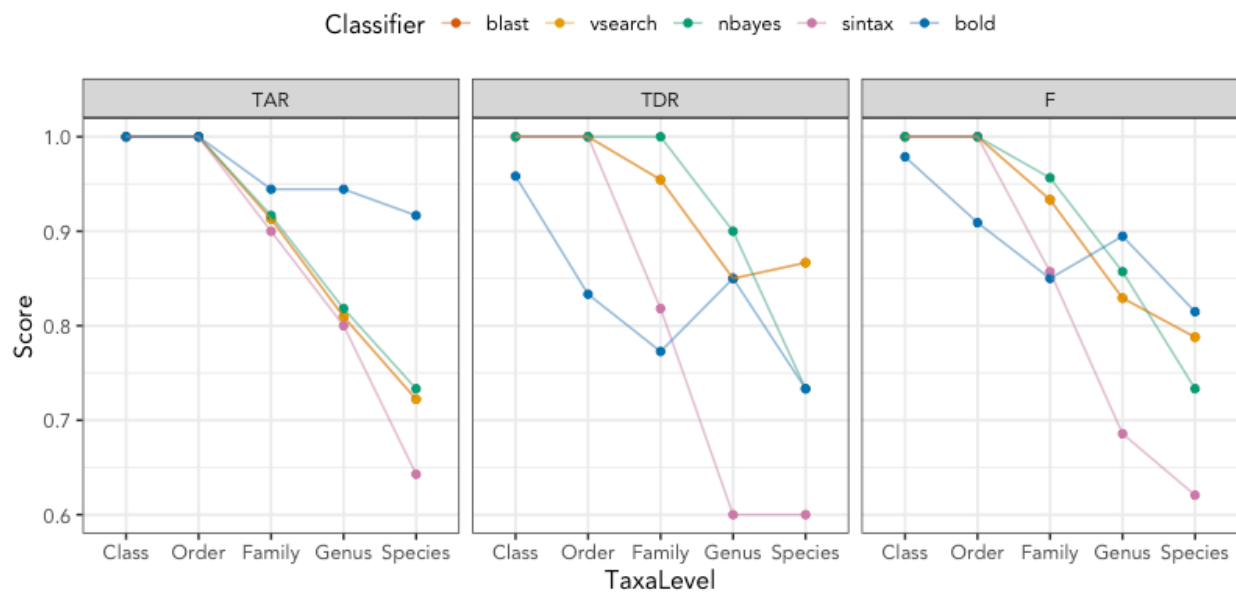
**Figure 5.** Diversity estimates for rarefied bat guano. Denoising methods are compared within each subplot according to the respective Hill value (vertical facet) and filtering parameter (horizontal facet). Sequence diversity is reduced with increasing Hill number. Filtering parameters reduce diversity estimates at a Hill a value of 0 (equivalent to observed OTUs) but are negligible for diversity estimates that incorporate abundance information (Hill values 1 or 2). Sequence variant equivalents for  $q=1$  and  $q=2$  are reduced compared to  $q=0$  estimates, indicating that relatively few distinct sequences encompass the majority of the sequence information in a given sample.



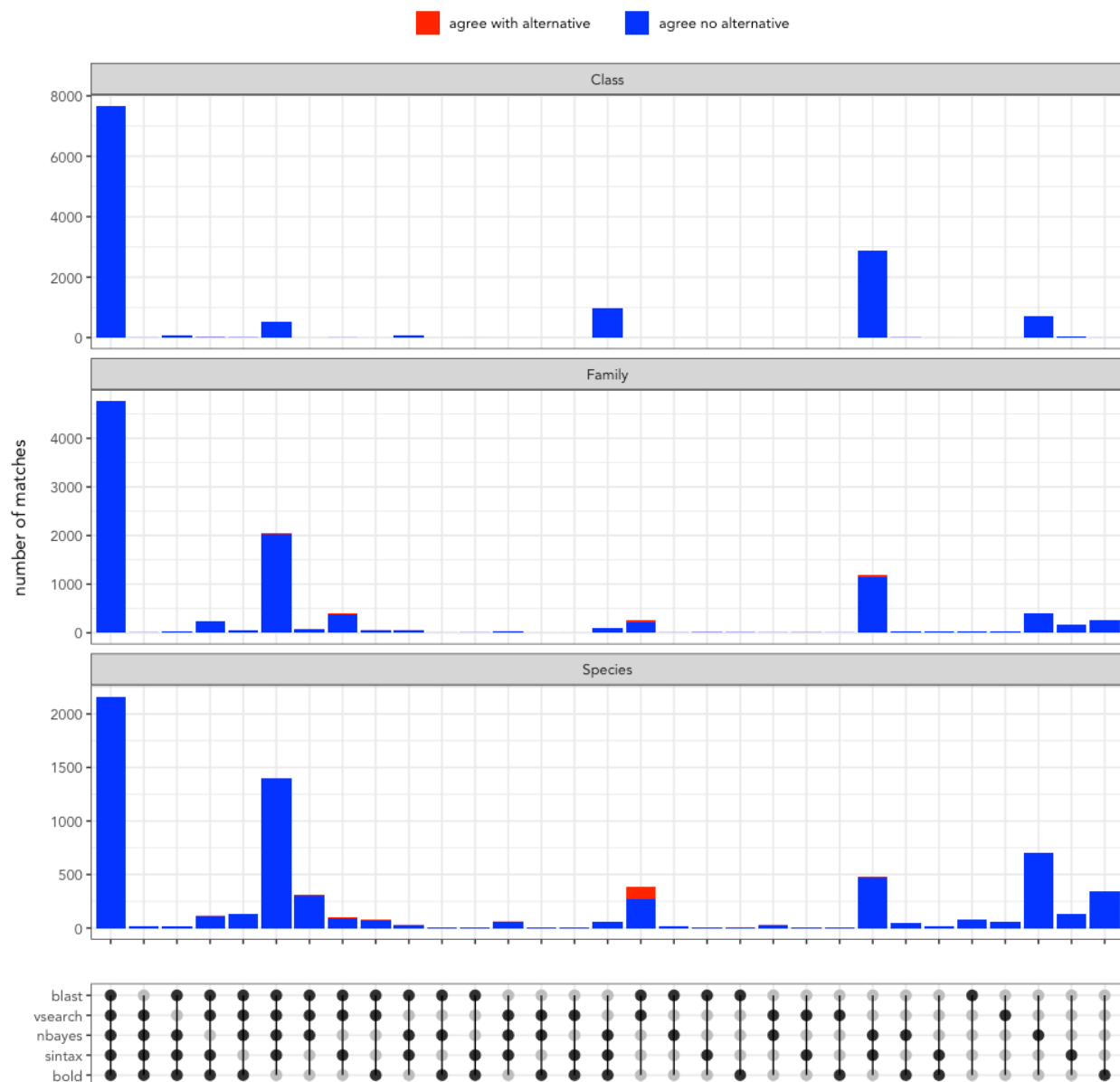
**Figure 6.** Database comparisons. Number of unique taxa and sequences at various taxonomic levels vary by database. **(A)** Porter and Palmer databases contain fewer unique arthropod taxa records than tidybug from species through order ranks. **(B)** Number of unique sequences are much lower in Porter database than Palmer or tidybug databases, though this reflects their decision of database design to include only taxa with known species names.



**Figure 7.** Effect of database clustering on taxonomic ambiguity. Clustering reduces the total number of distinct sequences and increases the proportion of ambiguous taxa. Our tidybug database consists of arthropod COI sequences that either contain taxonomic information (Present) or lack information (Missing) at taxonomic ranks class through species. Ambiguous taxa information is created via dereplication (clust 100%) or clustering (clust 99% - clust95%) when grouped sequences have discordant taxonomic description for a given rank. **(A)** Total number of distinct sequences are reduced by clustering, with increasing loss of sequence diversity as the clustering percent identity is lowered. **(B)** Clustering increases the relative fraction of ambiguous taxa in a database, with greatest amounts of ambiguity at the genus and species levels.



**Figure 8.** Classifier evaluations with mock data. Classifier performance of mock community data as measured by TAR, TDR and F. Taxon Accuracy Rate (TAR) measures the fraction of observed taxa that were expected, while Taxon Discovery Rate (TDR) measures the fraction of expected taxa that were observed. The F measure reflects the harmonic mean of TAR and TDR, and represents a balance between the precision and accuracy that TAR and TDR represent. Among classifiers (colored points) the BOLD API classifier retains the highest score across family through species levels for TAR. BLAST, VSEARCH and Naive Bayes classifiers retain higher TDR scores than BOLD or SINTAX indicating higher accuracy from class through genus level. VSEARCH values are identical to BLAST across all plots.



**Figure 9.** Common taxonomic identities among classifiers for guano data. Vertical bars represent the number of taxonomic names matching to a given set of classifiers. The particular classifiers included in a set is indicated by the connected and shaded dots in the bottom panel. For instance, the leftmost set (five shaded dots connected by a vertical line) represents instances in which all five classifiers agree for a particular taxonomic name, while the right-most sets (single shaded dots) represent instances in which a single classifier contains a unique taxonomic name. Among vertical bars, red represents instances in which a set is assigned a taxonomic name that is different from one or more non-set members (i.e. a non-ambiguous alternative exists), while blue represents the case in which the set agree on a name but the other non-set members do not have any information for that name (i.e. no alternative exists for that ASV).



## CHAPTER II

### Lord of the Dipterans: Molecular Diet Analyses of Indiana Bats in Illinois

#### ABSTRACT

Effective management of threatened wildlife requires an understanding of the foraging habitats used by those populations. Molecular diet analysis of fecal samples offers a cost-effective and non-invasive method to investigate the dietary breadth and temporal trends within wild populations. As opposed to traditional morphological identification, molecular analyses provide a rapid and more taxonomically detailed understanding of host diets. For the federally endangered Indiana bat (*Myotis sodalis*), documenting its preferred food sources can provide critical information to promote its effective conservation. Using sequence data from Cytochrome Oxidase I (COI) amplicons from bat guano data collected at Cypress Creek National Wildlife Refuge, we found that the dietary breadth of the Indiana bat is extensive, with 14 arthropod orders detected in at least 10 percent of samples and 190 unique families identified among those orders. However, the foraging activity inferred from the sequence abundance data suggests that dipteran taxa (i.e. flies) comprise the majority of the diet, indicating that this bat species consumes a broad spectrum of prey while repeatedly targeting specific taxa.

We applied machine-learning classification methods to investigate dietary preferences by month of collection and identified several taxa with strong seasonal variability. Collectively, these data depict the Indiana bat both as flexible consumers whose diet mirrors the expected life histories of available food resources, yet also aquatic invertebrate specialists making the most of the resources in the riparian habitat of Cypress Creek National Wildlife Refuge. These data can better inform policy holders faced with decisions of how to best allocate time and money to manage species habitat and improve conservation efforts.

## INTRODUCTION

The Indiana bat, *Myotis sodalis*, has the dubious distinction of being the first North American bat listed under the Endangered Species Preservation Act <sup>129</sup>. The historically broad distribution of Indiana bats once spanned much of the eastern United States <sup>130,131</sup>, however, populations were dramatically reduced through decades of anthropogenic effects and required regional and national efforts to mitigate declines <sup>131–133</sup>. Indiana bat populations appeared stable from the 1980's through the early 2000's <sup>39,40</sup>, but the emergence of White-Nose Syndrome (WNS)—an infectious disease caused by a fungal pathogen <sup>134,135</sup>—has decimated several bat species, resulting in near complete loss of some bat species at particular sites <sup>49,136</sup>. WNS has been particularly devastating to Indiana bats in the Northeastern U.S. <sup>39,40,137</sup>, and populations are currently concentrated primarily in just four states; populations in Kentucky,

Missouri, Indiana, and Illinois constitute over 95% of all Indiana bats detected in winter 2019 <sup>39</sup>.

Effective bat conservation requires protecting critical resources such as winter and summer habitats <sup>52,132</sup>. Radio telemetry has identified foraging preferences of Indiana bats for forested areas over agricultural <sup>138,139</sup> or urban <sup>44</sup> landscapes, although the species occupies a range of riparian and upland environments <sup>140–142</sup>. Understanding the particular habitats used by bats from maternity colony roosts, for example, has led to refined strategies by policy-holders to engage with land managers <sup>52</sup>. However, because Indiana bats occupy distinct territories within a landscape and often travel several kilometers between foraging habitats and roost sites <sup>47,143</sup>, research that identifies preferences about roost site selection (for example, see Jachowski et al. <sup>144</sup>) does not fully convey the habitat needs of the species.

Understanding food preferences may identify unique and additional habitat in need of protection. Yet there are limited accounts of Indiana bat foraging preferences, in part because the initial diet identification methods were time-consuming and required substantial taxonomic expertise to classify the arthropod contents. While these early studies suggest that Indiana bats are frequent consumers of dipterans (flies), coleopterans (beetles), and lepidopterans (moths and butterflies) <sup>44,45</sup>, as well as trichopterans (caddisflies) in certain conditions <sup>47</sup>, the lack of precise taxonomic identification of food items makes it challenging to translate observations into detailed management strategies. For instance, detecting a codling moth can indicate that bats

are foraging over an agricultural landscape while detecting a tent caterpillar would indicate a proclivity for forested habitat. Applying a molecular approach to diet analysis can provide the necessary taxonomic resolution to detail the breadth and specificity of Indiana bat foraging behaviors, and therefore give a more complete understanding of the habitat needs of the species.

DNA metabarcoding provides a cost-effective method to rapidly generate datasets rich with taxonomic information <sup>20-21</sup>. Molecular diet analyses have been widely applied to a range of systems and organisms, although the methodology is not without challenges and biases <sup>13,15</sup>. Early bat diet studies using a molecular approach described greater breadth and specificity of prey items consumed compared to traditional microscopy <sup>33,145</sup>. While both *in silico* <sup>31</sup> and empirical <sup>146</sup> studies have identified potential taxa that may be missed due to PCR biases, recent modifications of primer sequences have resolved many of the amplification issues for certain taxa <sup>32</sup>. Subsequent applications using this molecular method have revealed key features of bat foraging in several *Myotis* species that indicated prey specificity for *Myotis septentrionalis* <sup>147</sup> and *M. daubentonii* <sup>123</sup>; protections for the habitats that sustain these prey items would ensure these bats have available food resources. Studies of *M. lucifugus* indicate that core dietary components can vary both by location <sup>57</sup> and season <sup>58</sup>, suggesting that incorporating diet information into conservation efforts may require factoring in regional and temporal variation into management considerations. However, informed policy requires more than simply generating a list of prey items detected in a batch of guano—understanding which taxa constitute the largest portion of a bat's diet is critical when

using a molecular approach that can often identify hundreds of amplicon sequence variants (ASV's, or if clustered, similarly known as Operational Taxonomic Units).

While metabarcoding has improved our understanding of the depth and specificity of bat diet contents, linking the sequence data (i.e. counts of amplicons) to species abundances remains unresolved <sup>15,26</sup>. How these data are treated and analyzed is particularly relevant because the taxa that are identified as comprising the majority of the diet are potentially quite different depending on whether or not relative abundances of sequence counts are considered. Supervised learning (SL) tools (a type of machine learning) provide a potential method that leverages both occurrence and abundance information. A Random Forest classifier is one such SL tool, and has recently been applied to a range of 16S rRNA and ITS amplicon studies including identifying origins of ballast water <sup>148</sup>, predicting taxonomic signatures of host fecal microbiomes <sup>149</sup>, understanding maternal microbiome patterns associated with preterm delivery <sup>150</sup>, and predicting wine metabolomes <sup>151</sup>. Rather than summarizing the unique sequence variants of the data directly (e.g. through ordination), important features are identified in Random Forest classifiers by quantifying their relative contribution to the predictive accuracy of a model <sup>87,152</sup>. The motivation in using a Random Forest classifier in amplicon analyses is that it can assist in isolating the taxa that differentiate samples by some class variable. For example, determining which taxa are most predictive of location or seasonality can assist in identifying which are critical foraging areas or the seasonality of when those arthropod resources are most needed.

Located in between the Ohio and Mississippi Rivers in Southern Illinois, Cypress Creek National Wildlife Refuge contains ideal summer roosting habitat in addition to being within 8 km of a large Indiana Bat hibernaculum <sup>43</sup>. However, concerns of habitat loss and limited roost availability served as impetus to evaluate if artificial roost structures installed at two areas within the refuge (Egner and Hickory Bottom sites) would expand roosting use of areas that were otherwise not suitable for maternity colonies <sup>42</sup>. Guano collected as part of this study afforded an opportunity to provide the first molecular analysis of *M. sodalis* diet. Using a novel reference collection that combined Barcode of Life Database (BOLD) and GenBank sequence records, we highlight the breadth of prey consumed and document the seasonal changes in arthropod diet composition. In addition, we used Random Forest SL models to identify relevant taxa associated with Indiana bat foraging patterns across spatial and temporal dimensions. Indiana bats are one of several threatened or endangered species in need of significant protections. Identifying trends in foraging complements ongoing efforts to identify optimal habitat to preserve. We hope the methods described herein offer one such means with which improved species protections can be attained, and documented the bioinformatic work at the project GitHub repo: <https://github.com/devonorourke/mysosoup>.

## **MATERIALS AND METHODS**

### *Site selection and guano collection*

Installation of Brandenbark™ artificial roost structures <sup>153</sup> was completed in 2014 in two areas of Cypress Creek National Wildlife Refuge: Egner and Hickory Bottoms. These

areas consist of agricultural land mixed with mature bottomland forests containing live and standing dead trees with snags and crevices suitable for Indiana bat roosts. Both locations have access to riparian habitat, with Egner roosts abutting the Cache River, and Hickory Bottoms sites abutting Cypress Creek. Four artificial roosts were erected within each area (**Figures S14-S15**). Use of these structures by Indiana bats was confirmed through fieldwork conducted in July and August 2016 at the refuge using mist-netting, radio-telemetry, and acoustic surveys <sup>41</sup>.

Guano was collected at each of the eight roosts June 21, July 27, and September 15, 2017. Plastic sheets were placed at the base of each roost the night prior to collection and replaced with new sheets before the next collection date. Up to ten guano pellets were obtained at each roost at each date using sterile forceps and were stored individually in microcentrifuge tubes. Following individual pellet collection, remaining guano was pooled into 50 mL polypropylene conical tubes as a batch sample. All guano was sent to the University of New Hampshire and stored at -80 °C until DNA extraction.

#### *DNA Extraction*

Individual guano pellets were extracted using the Qiagen DNeasy PowerSoil kits (Qiagen, Hilden, Germany) following manufacturer guidelines. Two 96-well plates were used to process the initial samples, and the remaining samples were processed with single tube extractions using the same kit chemistry. All samples were eluted with 100 µL of elution buffer and up to eight extraction blanks were included for each 96-well plate. Batch samples were extracted using the same kit reagents with a modified

protocol. In these extractions, guano was thawed at 4 °C and up to 8 g per batch was transferred to a 50 mL conical vial. We added 15 mL of PowerBead solution and 1.2 mL of C1 solution along with five 5/32" 500C stainless steel grinding balls (OPS Diagnostics, Lebanon, NJ, USA) and homogenized at 1500 rpm for 10 min using a 1600 MiniG mixer (SPEX SamplePrep, Metuchen, NJ, USA). Then 400 µL of Proteinase K was added to the conical vial, briefly vortexed, and incubated for 1 h at 56 °C. Samples were pulse vortexed, then spun at 3220 xg for 5 minutes. Aliquots of 500 µL of supernatant were transferred to 2 mL microcentrifuge tubes, after which solution C2 was added and manufacturer guidelines were followed.

### *Metabarcoding*

Concentrations of guano extract DNA were estimated with a Nanodrop spectrophotometer (Thermo Fisher, Waltham, MA, USA) to guide the appropriate volumes of sample to add for subsequent normalization with SequalPrep plates following manufacturer guidelines (Applied Biosystems, Foster City, CA, USA). Highly concentrated samples were diluted so that samples were standardized to approximately 2 ng/µL prior to normalization. Normalized DNA was used as input for our overlap extension PCR method that targets arthropod COI fragments. Arthropod COI gene fragments are targeted for amplification using primers detailed in Jusino et al.<sup>32</sup>. We modified the original primer sequences to preserve the COI-specific regions, but added 5' extensions of 17 and 19 bp respectively. The constructs below illustrate these additional tails (bold underlined bases) as part of the modified oligos using the original Jusino sequences (not underlined):



UT-ANML-LCO1490: 5' -ACCCAACTGAATGGAGCGGTCAACAAATCATAAAGATATTGG-3'

UT-ANML-CO1-CFMRa: 5' -ACGCACTTGACTTGTCTTCGGWACTAATCAATTTCCAAATCC-3'

Samples were amplified in 15 µL reactions, with 3 µL of normalized guano DNA extract added to 12 µL of solution containing 0.2 µM of the forward and reverse primers, 0.16 µg/µL BSA, 0.03 U/µL Platinum Taq, 0.2 mM dNTPs, 1.5 mM MgCl<sub>2</sub>, and 1.5 µL of 10X buffer (Invitrogen, Carlsbad, CA, USA). Thermal cycler settings for the reaction consisted of an initial 5 min denaturation at 94 °C, followed by 5 cycles of 60 s at 94 °C, 90 s at 45 °C, and 90 s at 72 °C; an additional 35 cycles of 60 s at 94 °C, 90 s at 50 °C, and 60 s at 72 °C and finally a 10 min extension at 72 °C.

PCR reactions were subject to a 1X Ampure XP bead cleanup (Agilent Technologies, Santa Clara, CA, USA) and 10 µL of the concentrated solution was normalized in SequelPrep plates (Applied Biosystems, Foster City, CA, USA). These normalized PCR products were then subject to a second amplification using custom oligos that contained the requisite Illumina adapters, a distinct 8mer barcode, and the complementary sequence to overlap with the 5' terminus of the amplicon. The example below illustrates an example of these constructs, where the underlined portion represents an 8mer barcode, with the Illumina adapters upstream of the barcode, and the complementary overlap downstream from the barcode (in bold) to facilitate polymerase extension of the original PCR product:

**Indexed-UT1-example\_pair1a:**

5' –AATGATACGGCGACCACCGAGATCTACACCACACAAAGCTGGTCATCGTACCCAACTGAATGGAGC–3'

**Indexed-UT1-example\_pair1b:**

5' –CAAGCAGAAGACGGCATAACGAGATTTTGTGTGAGTCAGTCAGCCACGCACTTGACTTGTCTTC–3'

We added 2 µL of normalized PCR products (from the initial amplification) with 0.4 µM of each index primer in 25 µL reaction volumes using KAPA HiFi HotStart ReadyMix (KAPA Biosystems, Wilmington, MA, USA). Reaction conditions consisted of a 2 min denaturation at 98 °C, followed by 10 cycles of 30 s at 98 °C, 20 s at 60 °C, and 30 s at 72 °C, and a final extension for 5 min at 72 °C. These final PCR products were subject to another 1X bead cleanup and normalization following the same methods described above. We created the final library by pooling 10 µL of normalized PCR products into a single tube and concentrated to 40 µL with a 1X bead cleanup.

Library concentration was quantified by qPCR using the KAPA ROX Low Complete Kit (KAPA Biosystems, Wilmington, MA, USA). An Illumina MiSeq machine (Illumina, San Diego, CA, USA) with v3 chemistry generated 600 cycles of 2x300 bp paired-end reads. Raw sequence reads are available at NCBI BioProject PRJNA548356 (<https://www.ncbi.nlm.nih.gov/sra/PRJNA548356>).

***Bioinformatics***

Complete details including scripts and documentation of the various bioinformatic strategies employed for sequencing processing, database curation, classification, and

diversity estimates are provided in the project repo:

<https://github.com/devonorourke/mysosoup/tree/master>.

### *Sequence processing*

Demultiplexed sequences were trimmed using Cutadapt v.1.18<sup>89</sup>. Unpaired trimmed reads were imported into a QIIME 2 v2019.1 environment<sup>62</sup> and representative sequences were identified using DADA2 v1.6.0<sup>71</sup> via the q2-dada2 QIIME 2 plugin function 'qiime dada2 denoise-paired'. Of the 15 negative control samples submitted for sequencing, 12 generated raw sequences, but following denoising just 7 retained sequence data. All 289 guano samples submitted generated sequence data following denoising. The remaining control samples were included in subsequent classification and diversity estimates to determine the nature of potential contamination. Full details regarding sequence processing are outlined in the 'sequence\_processing.md' document:

[https://github.com/devonorourke/mysosoup/blob/master/docs/sequence\\_processing.md](https://github.com/devonorourke/mysosoup/blob/master/docs/sequence_processing.md).

### *Constructing databases*

While the primers used in this study were designed to target arthropod COI sequences, host (bat) DNA also potentially amplifies<sup>32</sup>. Because guano was passively collected, we first developed a host database consisting of sequences derived from all known bat species in the region. We developed this database to identify host DNA to assign the bat species for each guano sample, and separate these host sequences from the dataset to ensure diet analyses did not include bat DNA. Because our lab had other

guano-related projects happening concurrently with this experiment, we included all other known host reference sequences from those unrelated projects as a precaution for potential cross contamination (ultimately no unexpected host sequences were detected). Full details regarding host database design are documented in the 'host\_database.md' file:

[https://github.com/devonorourke/mysosoup/blob/master/docs/host\\_database.md](https://github.com/devonorourke/mysosoup/blob/master/docs/host_database.md).

We collected reference sequences and associated taxonomy information from two resources: BOLD <sup>28</sup> and a GenBank-derived dataset curated by Terri Porter <sup>111</sup>.

Reference sequences included COI records from arthropod, chordate, and other animal taxa, as well as fungal, protist, and other microeukaryote COI records. We dereplicated the initial collection of sequences, then applied a Least Common Ancestor (LCA) process using a consensus approach to classify records that shared identical sequence information but differed with respect to taxonomic information. Additional filters included discarding references with non-standard IUPAC DNA characters, removing sequences less than 100 bp, and retaining only references that contained at least family-level names. The final dataset included 2,181,331 distinct sequences. See the

'database\_construction.md' file for complete details:

[https://github.com/devonorourke/mysosoup/blob/master/docs/database\\_construction.md](https://github.com/devonorourke/mysosoup/blob/master/docs/database_construction.md).

All database files are stored in the Open Source Frameworks repo of this project:

<https://osf.io/qju3w>.

## *Classification*

Representative sequences were initially aligned to the custom host database of bat sequences using VSEARCH<sup>85</sup> to identify and separate host ASVs from non-host ASVs. Candidate matches were further queried with NCBI BLAST<sup>84</sup> to confirm host identities. We then used our full COI database as a third means with which to discriminate among host and non-host sequences, and classified all representative sequences using the classify-consensus-vsearch (which performs VSEARCH global alignment followed by LCA taxonomy assignment) and classify-sklearn (a Naive Bayes supervised learning taxonomy classifier) methods available through the QIIME 2 feature classifier<sup>87</sup>. Each method identified a common set of bat host ASVs which were removed from the original dataset as described in the 'classify\_host\_seqs.md' document:

[https://github.com/devonorourke/mysosoup/blob/master/docs/classify\\_sequences.md](https://github.com/devonorourke/mysosoup/blob/master/docs/classify_sequences.md).

Classification information derived from VSEARCH using the full COI database was used for assigning taxonomic identity for non-host ASVs. We used a custom script to remove sequence variants not assigned to the phylum Arthropoda, and discarded any ASVs that did not contain at least family-level names:

[https://github.com/devonorourke/mysosoup/blob/master/scripts/r\\_scripts/sequence\\_filtering.R](https://github.com/devonorourke/mysosoup/blob/master/scripts/r_scripts/sequence_filtering.R).

In addition, we evaluated a number of metrics to determine how to further filter the remaining non-bat sequencing variants, in part because 7 of the 15 negative controls contained some sequence data. Ultimately we discarded only those ASVs detected exclusively in negative controls samples while retaining sequence variants present in both guano and negative control samples. The 'contamination\_investigations.md' file

documents these steps:

[https://github.com/devonorourke/mysosoup/blob/master/docs/contamination\\_investigations.md](https://github.com/devonorourke/mysosoup/blob/master/docs/contamination_investigations.md).

### *Diversity estimates*

We estimated sequence diversity among guano samples using Hill Numbers in a custom R script:

[https://github.com/devonorourke/mysosoup/blob/master/scripts/r\\_scripts/Alpha-HillEstimates.R](https://github.com/devonorourke/mysosoup/blob/master/scripts/r_scripts/Alpha-HillEstimates.R).

Reviewed extensively by <sup>119</sup> and Chao <sup>99</sup>, Hill Numbers estimate the effective number of species necessary to produce the observed diversity on a continuum whereby read abundances are increasingly more relevant to the diversity estimate itself. We used three Hill values of 0, 1, and 2, corresponding to the traditional diversity estimates of richness, Shannon's entropy, and Simpson's 1-D diversity, respectively. Kruskal-Wallis and Dunn's tests were performed to determine if diversity estimates varied by month or site.

Dissimilarities in community composition (of sequence variants) were evaluated with binary (Dice-Sorensen) and abundance-based (Bray-Curtis, Morisita-Horn) non-phylogenetic metrics with Phyloseq <sup>97</sup> and Vegan <sup>98</sup> packages in a custom R script ([https://github.com/devonorourke/mysosoup/blob/master/scripts/r\\_scripts/betadiver\\_work.R](https://github.com/devonorourke/mysosoup/blob/master/scripts/r_scripts/betadiver_work.R)). We also calculated unweighted and weighted distance measures with the Unifrac implementation in QIIME 2 <sup>154–158</sup> to compare whether phylogenetic measures of binary or abundance-based estimates were more similar to non-phylogenetic distance estimates. We tested for main effects of site and month on community composition

using the Vegan ‘adonis’ function, and performed an analysis of multivariate homogeneity of group dispersions with the Vegan ‘betadisper’ function.

### *Supervised learning with Random Forest classifier*

The QIIME 2 ‘classify-samples’ pipeline (part of the q2-sample-classifier plugin <sup>87,113</sup>) provided a convenient open-source toolkit to train a Random Forest supervised learning classifier. We increased the number of decisions trees available to the model from the default (100) to 1000 estimators, with the intention of improving the predictive accuracy. In addition, we selected options to identify optimal feature selection using recursive feature elimination (--p-optimize-feature-selection) and optimize hyperparameters (--p-parameter-tuning) to automatically select the number of features considered during node splits on a given decision tree. Two classifiers were trained: the first classified samples by their respective site and month groupings, while a second classifier was trained using only month information. A recursive feature extraction quantified the relative importance of each feature (ASV) for each classifier. Classifier outputs were exported and processed in a pair of custom R scripts to produce the confusion matrices evaluating classifier accuracy:

[https://github.com/devonorourke/mysosoup/blob/master/scripts/r\\_scripts/machine\\_learn\\_heatmaps.R](https://github.com/devonorourke/mysosoup/blob/master/scripts/r_scripts/machine_learn_heatmaps.R),

and slope plots and heatmaps relating how relevant feature abundance and detections vary by month only or site and month groups:

[https://github.com/devonorourke/mysosoup/blob/master/scripts/r\\_scripts/machine\\_learn\\_analyses.R](https://github.com/devonorourke/mysosoup/blob/master/scripts/r_scripts/machine_learn_analyses.R).

All data processing measures related to diversity estimates and supervised learning analyses are described in the 'diversity\_workflow.md' document:

[https://github.com/devonorourke/mysosoup/blob/master/docs/diversity\\_workflow.md](https://github.com/devonorourke/mysosoup/blob/master/docs/diversity_workflow.md).

### *Additional software*

We relied on a series of additional R <sup>93</sup> libraries and Python packages such as Pandas <sup>159</sup> and other statistical software within the QIIME 2 environment including <sup>160</sup> and <sup>115</sup>. R libraries include tidyverse <sup>95</sup>, reshape2 <sup>96</sup>, cowplot <sup>104</sup>, ggpubr <sup>105</sup>, ggrepel <sup>106</sup>, scales <sup>109</sup>, qiime2R <sup>94</sup>, viridis <sup>110</sup>, gganimate <sup>161</sup>, xkcdcolors <sup>162</sup>, ape <sup>163</sup>, stringr <sup>164</sup>, taxize <sup>165,166</sup>, FSA <sup>102</sup>, formattable <sup>167</sup>, and rvest <sup>168</sup>.

## RESULTS

### *Dietary breadth*

We applied a metabarcoding technique to amplify arthropod COI gene fragments and generated sequence data from 285 bat guano samples collected at artificial roosts in the Cypress Creek National Wildlife Refuge during the summer of 2017. Although the primers used to amplify COI fragments were designed for arthropod sequences, other COI sequences such as host DNA often amplify as well. Thus, we first identified and separated host from non-host sequence variants. Many of our samples contained low abundances of one of three species of bat DNA: Indiana bat (*M. sodalis*), little brown bat (*M. lucifugus*), and evening bat (*Nycticeius humeralis*) (**Figure S16**). Of the 184 guano



samples with a unique bat species identified (66% of all samples sequenced), the vast majority of these were classified as the Indiana bat (177 samples). The rare detections of little brown bat (5 samples) and evening bat (2 samples) corroborate previous field observations<sup>41,42</sup> indicating that while other species transiently occupy similar roosts, the Indiana bat is the primary occupant of the colonies where guano was collected. We included all guano samples in our subsequent analyses regardless of host classification (or lack thereof), as many samples did not generate any host sequences. We acknowledge that a minor fraction of arthropod data may have come from a species other than Indiana bat.

The breadth of arthropod taxa detected among samples was substantial, with 14 of 21 unique orders identified in at least 10% of all samples (**Table S20**). Among these frequently observed orders we identified 190 unique families, although just five orders represented over 75% of all distinct families: Lepidoptera (39), Coleoptera (34), Diptera (34), Araneae (18), and Hemiptera (18) (**Table 1**). Although the Lepidoptera contained the greatest diversity of uniquely named families, genera, and species observed, the dipteran order contained more than twice as many unique ASVs (1190) as the next most frequent order (Lepidoptera, 567) (**Table 1**). However, despite the extensive diversity of the collective dataset, most ASVs were rarely observed in multiple samples. Of the 2,575 arthropod ASVs, just 63 of these were observed in more than 10% of all samples. Among those taxa observed in at least 10% of samples, most were classified as dipterans (44 unique ASVs among 6 distinct dipteran families). However, the ASV that was observed in the most samples throughout the study was a spider from the

genus *Eustala*, detected in over 70% of all samples (**Table S21**). Overall, our molecular techniques appear robust to amplify a wide range of arthropod taxa, with examples of high per-sample read abundances for several orders (**Figure S17**). Despite the extensive taxonomic variation of ASVs observed, our interpretations of diet preferences are fundamentally different if sequence abundances are considered (**Figure 10**). While occurrence information shows a plethora of arthropod orders being detected at different frequencies across collection sites and months (**Figure 10A**), the relative abundances of arthropod orders is decidedly dipteran (**Figure 10B**). We found that Indiana bats forage on taxonomically diverse assortment of arthropods, yet the high frequency of detections across samples and disproportionate fraction of sequences dedicated to a particular subset of ASVs within samples—particularly dipteran taxa—indicate that specific prey are consistently consumed throughout the season.

### *Diversity estimates*

We compared diversity of ASVs in guano samples using Hill Numbers 0, 1, and 2 (**Figure 11, Table S22**). A similar pattern emerged from each diversity estimate: samples collected in June and July had more sequence variants (or equivalents) than from September. A Kruskal-Wallis test was conducted to determine if diversity estimates for each Hill number varied among groups (group are defined by distinct combinations of site and month). A highly significant difference was observed for Hill Number 0 ( $H(5) = 21.98, p = < 0.001$ ). However, group differences were less pronounced for both Hill Number 1 ( $H(5) = 10.95, p = < 0.052$ ) and Hill Number 2 ( $H(5) = 11.14, p = 0.049$ ), suggesting less variation exists between groups for estimates incorporating sequence

abundances. Post-hoc Dunn's Tests for pairwise significance were also performed for each Hill Number. We found that the differences among group observed richness (Hill Number 0) varied between September and June, or September and July, across each site, but not between June and July (**Table S23**). Both Hill Numbers 1 and 2 (**Tables S24-S25**) identified a single pairwise difference (Hickory Bottoms-July, and Egner-September). Collectively we observed that the richness of ASVs within samples (Hill Number 0) was substantially larger than the 'common' or 'abundant' ASVs (as represented by Hill Numbers 1 and 2). Differences in diversity among groups are largely confined to datasets analyzed in a presence-absence context, with variation attributed by month, but not by site of collection. These data suggest that while bats may be foraging across a broad swath of the arthropod community, relatively few of these sequence variants dominate the overall abundance of reads generated in a single sample. Clearly, interpretations of bat foraging breadth are sensitive to whether or not abundance information is used as part of the analysis.

We next explored how community composition varied among groups using a multifactorial PERMANOVA (Adonis) to test for group differences in spatial median. We also used a univariate ANOVAs (betadisper) to test for dispersion differences for each group separately. Significant differences among group spatial medians were detected for each of the distance metrics applied, indicating that dietary components vary by site and month (**Tables S26-S30**). However, the PERMDISP tests also proved significant; because we used a balanced design, these results suggest that month and site variability in community composition occurs both because of dispersion and spatial

group median differences. We found a larger proportion of variation was explained by month than either site or the interaction term for every metric. Indeed, the PCoA ordinations suggested that samples were partitioned largely by month but not site (**Figure 12**), particularly for unweighted estimates (**Figures 12A and 12D**). However, the proportion of variation explained by the first two principal components is generally small for most distance measures, likely due to the fact that many of the prey items that bats consume are present throughout the entire sampling period of the study, thus the overall impact of month or site differences are likely minor. Notably, the weighted Unifrac measure had the largest proportion of variation (32.6% for weighted Unifrac, and between 11.3–17.1% for all others).

#### *Important features in Random forest classifier model*

To identify the relevant features (ASVs) of our data important to discriminating among site and month class variables, we trained a Random Forest classifier using a hold-out set consisting of 80% of the samples. We then tested the classifier accuracy using the remaining 20% of the data (not used in the training dataset) and observed a high overall accuracy rate of 0.89 (**Figure 13**). The few instances in which the classifier misassigned a sample to the wrong group were confined to either of the group factors (site or month), but never both (site and month). A recursive feature extraction was performed as part of the initial classifier evaluation, thus by selectively removing features and comparing the change in model performance, we identified the ASVs most relevant to model training, and therefore identified the features most important to discriminating among groups (**Figure 14**). We found that the majority of features

important to model accuracy belong to just three orders: Diptera (122), Lepidoptera (41), and Araneae (22) comprised the majority of the 237 features identified by the classifier. Diptera—in particular crane flies, mosquitoes, and chironomids—are not only among the most consistently detected and highly abundant prey items in our samples (**Figure 10**), but they also had the greatest variability in those measures for site and month classes (**Figure S18**).

We also trained a separate classifier to distinguish samples by month only, and found that fewer features were required to train an even more accurate model (128 ASVs compared to 237 in site and month classifier; overall accuracy = 0.91 compared to 0.89 in site and month classifier; see **Figure S19**). Each classifier (month only, and site and month) identified the same features relevant to model accuracy: all 128 ASVs identified by the month classifier as important were among the 237 identified in the site and month classifier. Moreover, the most relevant classifier features—defined here as the 75th percentile with respect to model importance—were highly similar, with 30 of 32 ASVs shared among the two classifiers. Interestingly, while most of these important ASVs were dipteran, the individual ASVs with the highest individual importance values were non-dipteran arthropods (**Figure S20**). Given the high overlap in feature selections between the two models it is therefore unsurprising that dipteran taxa were identified as the most variably consumed group of taxa, while a few select psocodean and hemipteran taxa are important foraging targets exclusively in September (**Figure S21**). However, a few of the previously important dipteran ASVs were no longer identified by the month-only classifier. In these instances it appears that bats were actively foraging

these taxa repeatedly across all collection months, but differed in sampling abundances between sites (**Figure S22**).

Overall we find that Indiana bats utilize a diverse foraging strategy spanning many arthropod orders, although our molecular tools suggest that Diptera are by far the most frequently targeted prey items. Similar taxa are typically identified among the two sites, and variability underlying feature detection was often driven by month of collection rather than collection site. However, there are a few instances in which a frequently observed ASV is not easily attributed to a particular sampling month, and such cases generally indicate a particular taxon is foraged upon throughout the entire summer (e.g. ASV-3, a crane fly of the genus *Rhipidia*). A series of animations that plot the changes in ASV abundance and detections were created to help visualize the differences observed for both the site and month and month-only classifiers. These animations clarify the dramatic turnover in detections and sequence abundances observed for many Diptera, while also highlighting a few other dynamic changes for other arthropod orders: [https://github.com/devonorourke/mysosoup/blob/master/docs/animation\\_analyses.md](https://github.com/devonorourke/mysosoup/blob/master/docs/animation_analyses.md).

## DISCUSSION

Our molecular approach to studying Indiana bat diet confirmed earlier morphological guano analyses for this bat species: many arthropod orders including Coleoptera, Diptera, and Lepidoptera were repeatedly detected. However, we observed diptera taxa

as the largest proportion of fecal content, while most of the prior morphological studies suggest Indiana bat guano consists of beetle and moth taxa (see Figure 1 in Sparks et al. <sup>44</sup> for review). This disparity was exhibited also in a recent survey conducted in Shawnee National Forest—just 20 miles east of our location—that suggested Indiana bats consume largely moths and beetles <sup>46</sup>. While it is probable that these differences are partly due to prey availability, it is also quite likely that our broader understanding of Indiana bat prey consumption is influenced by the analytical tools applied. We found that applying a molecular technique provides greater taxonomic specificity of diet components, and also expands the richness reported by classic morphological analyses. Moreover, while the list of prey targets may be long, the conventional description of the Indiana bat as a generalist insectivore—a term repeatedly applied to many North American *Myotis*—is worth reconsidering in light of these molecular diet profiles. Specifically, the relative proportions of sequences detected in guano suggest a far more targeted set of prey, principally focused on an array of Diptera and a genus of spider, *Eustala*. Therefore, while these molecular tools have confirmed and expanded our understanding of Indiana bat diets, translating these data into actionable management practices is invariably shaped by whether or not the relative abundances of sequences are considered in the analysis.

The largest fraction of sequence counts were classified as dipteran and aranean, and this disparity over other detected arthropod orders likely reflects a combination of availability and biomass rather than an artefact of experimental design. As reviewed by Deagle et al. <sup>26</sup>, incorporating abundance information into fecal analyses is challenging

for several reasons, including different digestion rates of arthropod prey or DNA extraction biases. Observed differences in sequencing depths can also be impacted by the particular molecular tools applied. For example, *in silico* analyses<sup>31</sup> and empirical tests<sup>32,37</sup> suggest that primer choice can influence observed taxonomic diversity, as can the various choices of sequencing platform and depth of coverage<sup>37</sup>. The primers we employed in this experiment were previously tested using biological mock communities and indicated only minor bias among particular arthropod orders (see Figure 1 in Jusino et al.<sup>32</sup>). Interestingly, these biases lead to marginally greater coleopteran and lepidopteran sequences rather than Diptera, making it unlikely that our frequently detected spider and fly sequences are a result of preferential template binding. Moreover, we found that most arthropod orders in this dataset generate ASVs with similar distributions of sequence counts (**Figure S17**)—the distinction is that there are fewer unique ASVs detected among non-dipteran orders. In fact, if we consider the five most deeply sequenced ASVs in every sample, we find at least one non-dipteran or non-aranean ASV among 66% of all samples. Thus, it does not appear that the relatively larger fraction of fly and spider taxa is due to any particular molecular bias.

We found that the inclusion of sequence counts often reframed our subsequent interpretations of diversity. For example, observed richness varied by month, yet abundance-weighted estimates suggest that few taxa contribute to the overall species diversity in a sample, and there is little variability in the proportion of dominant taxa across site or month groups (**Table S22, Figure 11**). The majority of highly abundant ASVs were classified as dipteran, though a few taxa in other orders were also frequently



detected and generated highly abundant sequences (**Table S21**). Likewise, our interpretation of community composition was influenced by abundance information: variation in community composition is more strongly attributed to collection month for abundance-unweighted than weighted measures (**Figure 12, Tables S26-30**). Through the lens of occurrence data, the Indiana bat appears to be a generalist insectivore consuming a changing menu of arthropods throughout the summer season as has been previously reported, yet abundance-data portrays a voracious predator largely targeting aquatic invertebrate prey based on seasonal availability.

Among all ASVs detected in our experiment, the Random Forest models identified only a fraction of the taxa as being important for classifying samples to month (**Figure S21**) or site and month (**Figure 14**) groups. Interestingly, these changes in detections and abundances for particular ASVs often matched the expected life histories of the specific taxa. For example, we detected leafhopper (ASV-7 and ASV-12) and barklice (ASV-14) ASVs largely in September, and these hemipteran and psocodean populations are known to build through the season and emerge as adults on the wing in large cohorts later in the summer (M. Jeffords, pers. comm). Indeed, the strongest signals attributed to individual ASVs were among non-dipteran taxa, despite the majority of the diet consisting of flies (**Figure S20**). In particular, the classifiers repeatedly identified dipteran species as important for discriminating samples by month (**Figure S21**) or by site and month (**Figure S14**). However, other dipterans were not selected by the classifier, likely because their frequencies were consistently high among all group classes. For example, *Culex erraticus*, among the most frequently detected mosquitoes

is often observed in July and September, while *Aedes vexans* is generally detected in June. However, other taxa including many crane flies in the genera *Erioptera* and *Rhipidia* were frequently observed in every month at each site. Because our study did not conduct insect trapping at the time of guano collection it is unclear whether differences in monthly variability of some taxa (and a pronounced lack of variability in others) is due to selective foraging or prey availability. Future foraging studies will benefit from pairing traps with molecular data to better understand the relationships of particular prey availability and observed diet.

Interpreting the dietary preferences of these bats in either a presence-absence or abundance-based context would ultimately provide similar support for the ongoing recovery plans for the Indiana bat in the Cypress Creek National Wildlife Refuge to protect forested wetlands as critical habitat for foraging (in particular, see Chapter 4 of the Cypress Creek National Wildlife Refuge Habitat Management Plan <sup>43</sup>).

Nevertheless, characterizing these bats as generalist predators in a presence-absence framework may overestimate their dietary resilience. Our alpha and beta diversity comparisons suggest that many low abundance and rare sequence variants support this perception of a generalist behavior. An abundance-weighted analysis of their dietary contents point to a greater dependence on a subset of aquatic insects. Therefore, future management efforts may need to focus specifically on mechanisms to conserve habitats critical to those prey taxa. Fortunately such language already exists within the broader Cypress Creek National Wildlife Refuge management plan to protect waterfowl, however what is good for the duck may not be best for the bat. For example, it is

unclear whether the prescribed water level manipulations in the upstream channels leading into wetland habitats adjacent to bat roosts will reduce the availability of dipteran prey should stream flow velocity be altered.

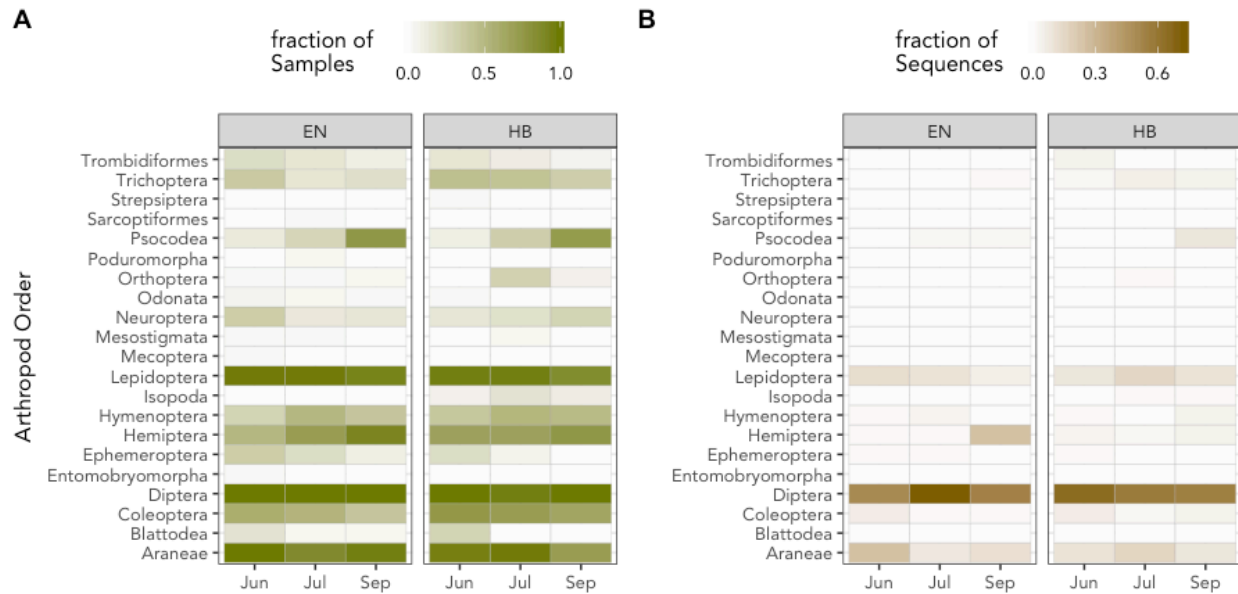
Much of existing bat conservation policy focuses on identifying and conserving winter hibernacula and summer maternity roosts. For the Indiana bat specifically, there has yet to be a framework defining mechanisms to understand the particular resources essential for foraging habitats. We found that the molecular techniques applied herein achieved a far greater taxonomic resolution of bat diets than previous morphological estimates, and do so in a rapid and cost-effective manner. Our data suggest that bat foraging strategies include both temporal and spatial variation, thus future studies are best served to include both variables in collection regimes. Our bioinformatic and molecular workflows are easily adapted for any other insectivorous bat species, and guano is cheap and easily obtained once roosts are identified. The current limitations towards a more contextual understanding of particular foraging behaviors for many threatened or endangered bats remains a human one, and we advocate more extensive investigations using these molecular methods to craft better informed, site-specific conservation management plans.

## TABLES

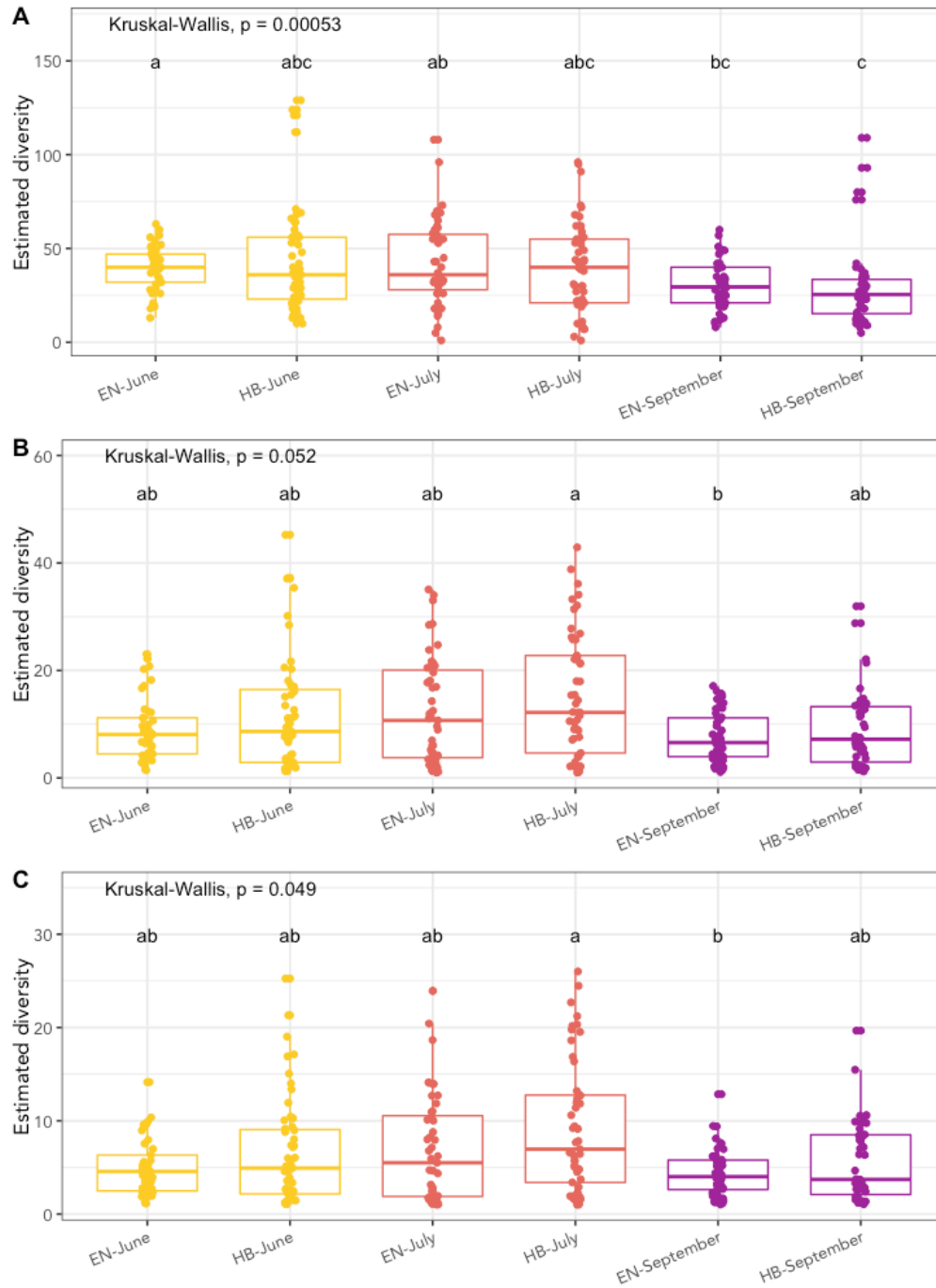
Arthropod Order	Samples	Families	Genera	Species	ASVs
Diptera	275	34	113	95	1190
Lepidoptera	262	39	173	183	567
Araneae	253	18	55	71	219
Hemiptera	196	18	50	36	162
Coleoptera	168	34	68	59	176
Hymenoptera	119	10	23	10	68
Psocodea	104	7	11	10	61
Trichoptera	90	4	8	10	22
Neuroptera	56	5	13	9	31
Ephemeroptera	43	4	5	6	13
Trombidiformes	36	9	7	2	15
Blattodea	27	2	3	5	14
Orthoptera	22	5	9	6	21
Isopoda	16	1	1	1	1

**Table 1.** Dietary breadth of Indiana bats. We detected 14 distinct arthropod orders in at least 10% of all 285 guano samples sequenced. Our molecular and bioinformatic techniques were both sensitive and specific, resulting in classifying over 83% of all unique sequence variants (ASVs) to at least Genus. At least one sequence variant was classified from the orders Diptera, Lepidoptera, Araneae, Hemiptera, or Coleoptera in the majority of samples. The number of uniquely named family, genus, and species sequence variants is greatest among Lepidoptera, however among all ASVs detected the Diptera contained more than double the number of unique (ASVs) as the next order (Lepidoptera). Indiana Bats may be eating more uniquely named moths, but the greatest sequence diversity of COI amplicons detected in guano samples were decidedly dipteran.

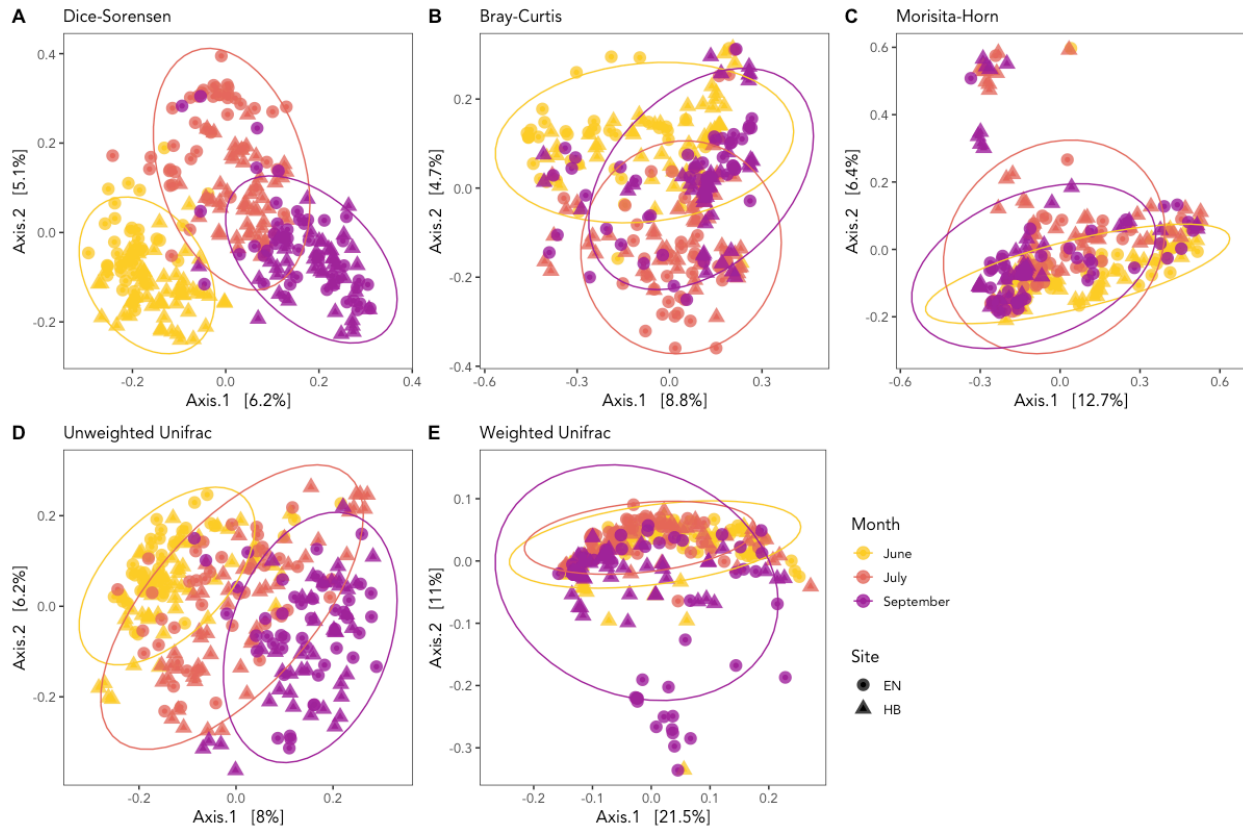
## FIGURES



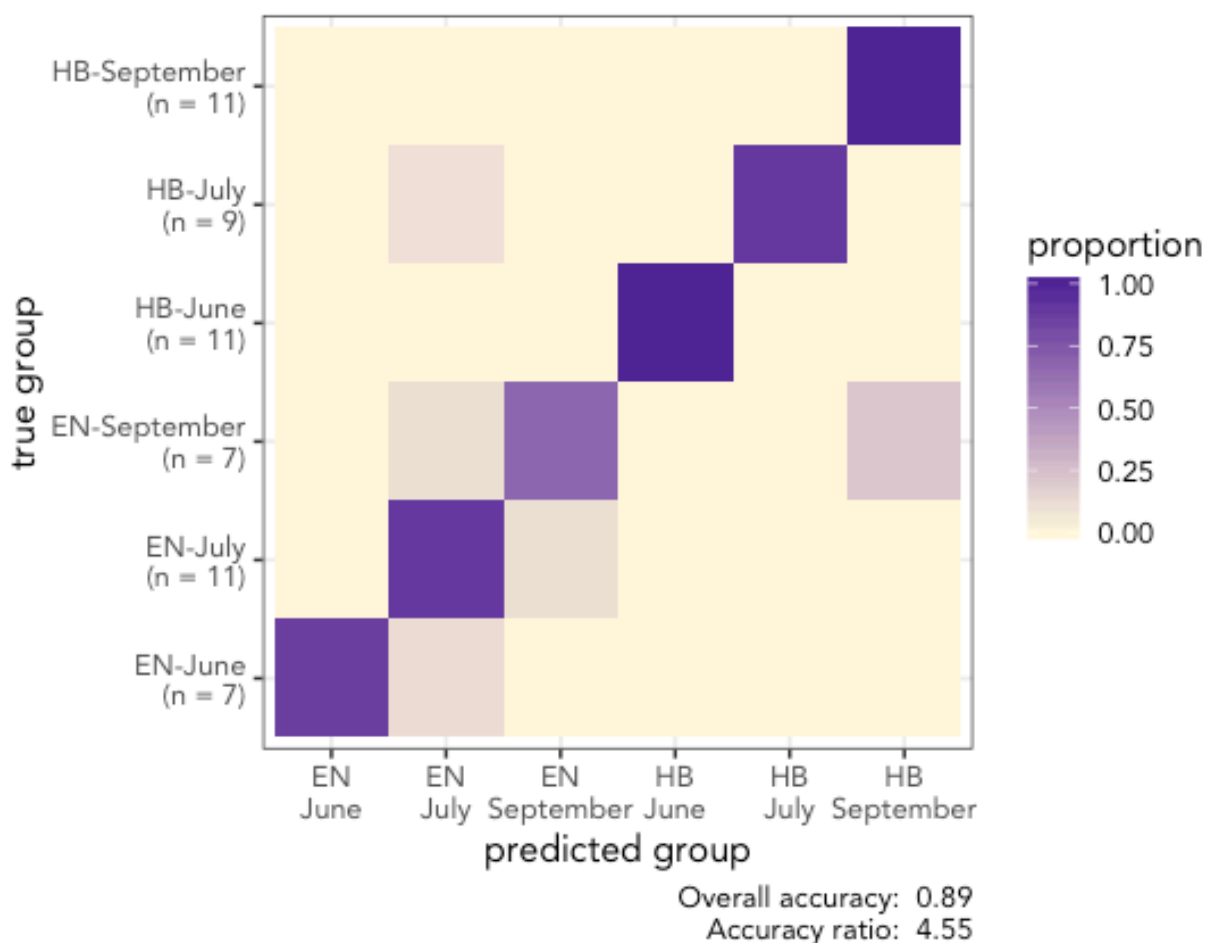
**Figure 10.** Comparing diets using detections or sequence abundances. Perspectives of bat dietary diversity are influenced by inclusion of relative abundances of sequences detected. (A) In a presence-absence context, there is a substantial variety of arthropod taxa detected among samples grouped by collection months (x axis) and sites (vertical facets: Egner, “EN”; Hickory Bottoms, “HB”). The heatmap gradient represents the fraction of samples per site and month group containing at least one ASV classified to the particular arthropod order represented (y axis). For example, the majority of samples at both EN and HB sites contain at least one psocodean ASV in September but not in June. Likewise, most samples in each group contain detections for ASVs classified as Araneae, Coleoptera, Diptera, Hemiptera, and Lepidoptera. (B) When data is represented in terms of relative abundances of reads instead of occurrence, the overall fraction of reads for any particular group is dominated by dipteran-classified ASVs for all groups. Dipteran sequence diversity is highest among all orders (Table 1), yet the per-sample variability in Dipteran ASVs is no different than many of the other highly sequenced non-Dipteran taxa (Figure S4), suggesting that the marked difference in detections and abundances is due to foraging habits rather than a molecular biases.



**Figure 11** – Diversity estimates using Hill Numbers. Observed richness (A) diversity estimates are greater than abundance-weighted estimates using Shannon's entropy (B) or Simpson's index (C), indicating that richness is driven by low abundance sequence variants. Pairwise differences between sampling site or season are indicated by different letters at the top of each boxplot, with abbreviations used to indicate Egner (EN) and Hickory Bottoms (HB) locations.

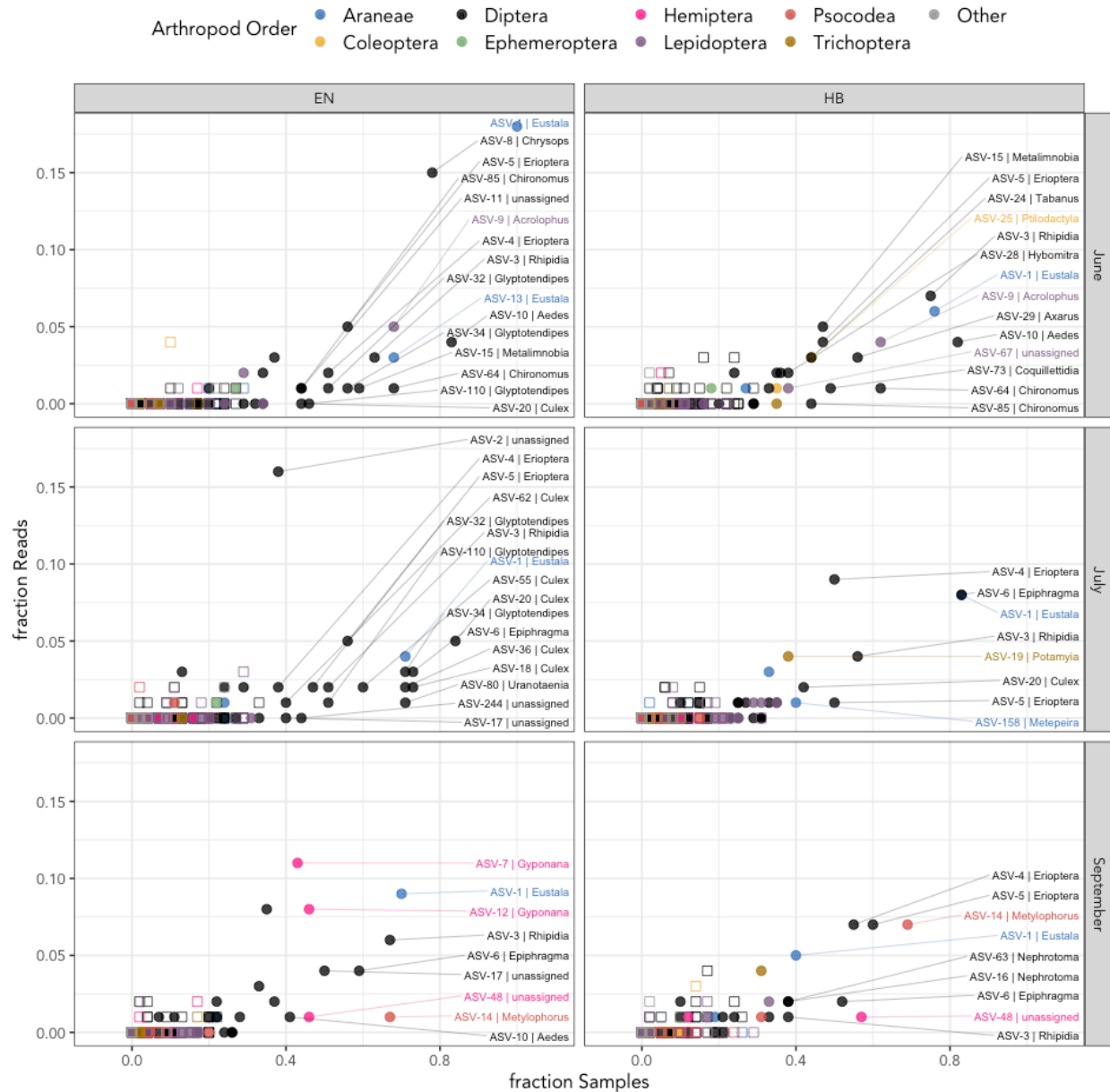


**Figure 12** – Beta diversity estimates of Indiana bat samples. Perspectives of community composition of bat diet can be shaped by including abundance and phylogenetic information. Five measures of distance estimates differed based upon how rarefied abundance information or phylogenetic information was used. Principal coordinates analysis (PCoA) from each distance estimate was ordinated with samples distinguished by collection month and site (Egner, “EN”; Hickory Bottoms, “HB”). Unweighted phylogenetic and unweighted abundance (A), unweighted phylogenetic and weighted abundance (B-C), or weighted phylogenetic and unweighted abundance (D) had little variation explained by the first two principal components, while the phylogenetic and abundance weighted distance estimate (E) contained nearly twice the variation. Both unweighted abundance measures (A,D) partition samples by distinct months, but not by distinct sites, though these distinctions become less apparent as abundance information is increasingly weighted in distance estimates for non phylogenetic (B, C) and phylogenetic (E) measures. Collectively we find that rare and low abundant taxa are likely inflating community composition differences between samples.



**Figure 13** – Random Forest classifier performance. A Random Forest model was trained to identify the features (i.e. ASVs) relevant to classifying a guano sample to a particular collection site (Egner, “EN”; Hickory Bottoms, “HB”) and month group. We trained the model using 80% of the dataset, then tested the model accuracy by classifying the remaining 20% of samples. This confusion matrix demonstrates that the model was highly successful at matching the predicted group with the true group for most samples. The recursive feature extraction performed in this process identified few features (less than 10% of all ASVs detected in the study) as important for model accuracy, and the majority of these were dipteran. Accuracy ratio = overall accuracy / baseline accuracy (the accuracy that would be achieved by assigning all samples to the most common class).





**Figure 14** – ASVs important to identify Site and Month differences using Random Forest classifier. Several dipteran taxa and other non-dipteran ASVs vary in terms of relative detections (x axis) or abundances (y axis) per site and month group. The particular ASVs identified as important to a Random Forest model discriminating samples according to site and group classes are denoted by closed circles, while all other ASVs in the study are signified by square brackets. Labels indicate the genus name for those highly detected ASVs in a particular group. Some features are highly abundant and frequently detected in all groups (e.g. ASV-1), yet these proportions vary enough for differences to be meaningful for the decision tree to discriminate among site and month classes. For example, with ASV-1, we see a decrease in read abundance at the Egner site, but a slight increase at Hickory Bottoms between June and July. Other groups tend to have monthly changes in abundance or detections that follow similar patterns for each site, but the magnitude of these differences are what help the model discriminate among site and month. For example, ASV-4 tends to increase in abundance and detections from June into July and September at both sites, however, the total number of detections and read abundances are generally larger for Hickory Bottoms than Egner. While these bat guano samples are dominated by dipteran ASV sequences, many of these features are distinct by site and month groups, indicating that bats are foraging from a shifting menu throughout space and time, consisting largely of aquatic invertebrates.

## CHAPTER III

### Eating locally: molecular diet analyses of New Hampshire bats reveal spatial and temporal variation in prey consumption

#### **ABSTRACT**

Insectivorous bats are voracious consumers of a diverse array of arthropod prey. Previous morphological analyses of arthropods in North American bat diets suggest that while these insectivorous bats may be considered generalists collectively, particular arthropod preferences can strongly vary depending on the specific foraging environment. Yet, taxonomic resolution in these studies has frequently been limited to the order-level, limiting our understanding of diversity within diets. Molecular metabarcoding provides a more taxonomically refined assessment of diet than morphological analyses and can often determine common or distinct prey items at the species level. Of the few molecular analyses of North American bat diets, substantial variation was reported between sites and across seasons and years. However, analyzing compositions between seasons may be too broad a time period to capture the extent of local variability in prey consumption, and it is unclear whether the yearly compositional changes previously observed at a single site are likely to occur across a larger area. We assessed the spatial and temporal variability in bat diets in New Hampshire, USA by sampling across 20 sites at biweekly intervals over 2 years to provide a comprehensive evaluation of the propensity for local foraging. Coleoptera,

Diptera, and Lepidoptera were the most prevalent of the 12 arthropod orders detected, although their particular compositions varied among locations and dates. Few taxa were consistently detected at most locations between sites, seasons, and years, indicating that while some prey items are repeatedly consumed, the majority of the dietary breadth of bats is ephemeral and site specific. A few of these commonly consumed taxa were known turf and forests pests, demonstrating that these bats are providing ecosystem services throughout the state. Richness and composition of prey varied across space and time in distinctly local fashions suggesting that these bats are flexible foragers adept at consuming the locally available resources.

## INTRODUCTION

North American insectivorous bats have highly flexible foraging strategies. Dietary analyses indicate consumption of a broad assortment of prey<sup>3</sup>, yet the composition of prey contents reported appears particularly sensitive to temporal or spatial factors among little brown (*Myotis lucifugus*)<sup>57,58</sup> and big brown (*Eptesicus fuscus*)<sup>59,169</sup> bats. Similar patterns of temporal variability of arthropods in bat diets occur in their European relatives<sup>76,123,146</sup>. It is thus unsurprising that many of the factors associated with intraspecies variation in little brown bat diets—a species with extensive historical records—are connected with either location or time. These factors include variation in prey abundance<sup>170</sup>, bat age<sup>171</sup>, landscape features<sup>172</sup>, and ambient temperature<sup>173</sup>. Heterogeneity in sampling location or date can alter the community composition of prey available and therefore observed, suggesting that a sampling design that surveyed

multiple populations simultaneously and repeatedly throughout the foraging season would provide a more comprehensive understanding of the niche breadth of a species.

The techniques used to describe diets also play a fundamental role in how we characterize the foraging habits of a species. Historically, morphological analyses have described bat diets at order-level compositions as some ratio of Coleoptera, Diptera, Lepidoptera, with smaller fractions of Trichoptera and Ephemeroptera and a few other taxa. For example, these methods indicate that North American bats like *E. fuscus* are beetle specialists<sup>174</sup>, while *Myotis* spp. more frequently consume flies and moths<sup>12,175,176</sup>. More recent studies using molecular methods support some of the historically observed order-level diet findings—*E. fuscus* still love beetles<sup>59,177</sup> (but not as much in southern desert communities<sup>25</sup>), and *M. lucifugus* continue to forage on flies and moths<sup>57,58</sup>. Although not without their own biases (see Nielsen et al.<sup>13</sup> for a review of various diet tracing techniques, and Alberdi et al.<sup>15</sup> for molecular metabarcoding specifically), molecular analyses generally provide higher resolution for detecting prey items relative to morphological techniques, and can reveal a much broader palate than previously described<sup>18</sup>. Rather than a small menu consisting of a few arthropod orders, molecular studies of insectivorous bats routinely describe hundreds of unique sequence variants detected with varying frequencies across numerous orders. In light of the superior taxonomic resolution that molecular metabarcoding provides, historical assessments may have significantly underestimated the niche and dietary breadth of the bat species described.

Extinction likelihood may be related to dietary diversity among insectivorous bats <sup>178</sup>, thus the detailed information provided by metabarcoding may be of particular importance for North American bat species that have been devastated by White-Nose Syndrome <sup>48,134,135,179</sup>. Fortunately there are some areas where regional stabilization has been reported <sup>50,51</sup>, and diet information can provide an opportunity to more fully characterize the local habitat resources required by these populations. Indeed, federal guidelines <sup>52–54</sup> and population models <sup>55,56</sup> alike suggest that management efforts are critical to preventing future extinction even among seemingly stable populations. However, molecular diet information is scant for a few of the species impacted by White-Nose Syndrome and entirely absent for others. At the same time, insect populations have recently exhibited dramatic declines worldwide <sup>180–182</sup>, and anthropogenic factors such as water <sup>183</sup> and land <sup>184</sup> use, as well as pesticide application <sup>185,186</sup> can reduce insect availability and therefore affect aerial insectivores such as bats.

Among the few multi-year molecular diet analyses of North American bats, local compositional variability was observed both within and between seasons at the same location in little brown <sup>57,58</sup> and big brown <sup>59</sup> bats. Because these multi-year assessments were conducted at a single site for each species, more data are needed to determine whether bats in other locations will display similarly diverse foraging patterns across sampling years, or if diets will be relatively consistent from year over year. Additionally, temporal variability was assessed in just three time intervals for both studies, reflecting phases of pregnancy (May to mid-June), lactation (mid-June to mid-

July), and post-lactation (mid-July to September) where energetic and dietary changes are known to vary <sup>187–189</sup>. Nevertheless, these relatively long sampling windows may smooth otherwise more volatile patterns in the prey consumed and may bias against the particularly ephemeral taxa with limited adult lifespans.

We were therefore motivated to build on these foundational studies and completed a two-year sampling regime that targeted multiple sites across southern and central New Hampshire, USA. Guano from primarily little brown bats was passively collected weekly at 20 locations in 2015 and 2016, seven of which were sampled in both years (although only three of these sites had at least 2 months of overlapping sampling periods). Our study had three main objectives: first, assess the dietary breadth of these bats across all of our locations; second, compare the extent with which bat diets vary along temporal and spatial gradients; and third, identify what, if any taxa consumed by these bats were pest species. We discuss how guano sampling can be used to simultaneously deepen our understanding of the dietary flexibility of a species, inform conservation efforts by identifying the foraging habitats, and identify potential forest or agricultural pests in an area.

## **METHODS**

### *Sample Collection*

Individual guano pellets were passively collected at sites throughout New Hampshire (**Figure 15**) beginning June 2015. We relied on citizen scientists to assist with collecting samples at 19 of the 20 sites. These locations consisted of a mix of forest and

agricultural landscapes typical of the region. We obtained data at a nature center (HOL), a forest research station (FOX), conservation lands (BRN, MAP, MAS, WLD), and privately-owned homes (all other sites) from colonies occupying structures such as attics, barns, garages, and bat houses. Volunteers were provided with supplies (forceps, dust masks, nitrile gloves, ethanol wipes, plastic sheets), and were instructed to collect 10 fresh guano samples by transferring individual pellets into the pre-filled microcentrifuge tubes containing 1 mL storage buffer (3.5M ammonium sulfate, 16.7 mM sodium citrate, 13.3 mM EDTA, pH 5.2). Plastic sheets were replaced weekly to avoid cross contamination over the season, and samples were shipped in batches back to our lab approximately every month. Samples were then stored at -80 °C until DNA extraction.

### *Mock Community*

In addition to sequencing bat guano, we included a biological mock sample with each library to assist in sequence quality control following Jusino et al. (2019), whereby DNA extracted from voucher arthropod specimen was amplified using ANML primers. PCR products were cloned into plasmid vectors and Sanger sequenced to match specimen taxonomy with sequence identity. Taxonomic identities were initially assigned by a trained entomologist's visual identification. More exclusive taxonomic levels that could not be distinguished by the taxonomists were added by manually aligning full length COI Sanger sequence data to NCBI's nt database. Sequences were required to have at least 98% identity and 92% coverage to be named to the species level. Any instance in which multiple best hits were available resulted in removing the species classification

(this occurred in just one instance). In addition, we removed a single instance in which a mock sample had a single best hit as "Sp. 1ES" despite being unambiguous because other references in our database and on NCBI suggested there are other "Sp." names that did not clearly identify this record as a single species.

### *Metabarcoding*

DNA was extracted from individual guano pellets using 96-well plate format of the Qiagen DNEasy PowerSoil Kit (Qiagen, Hilden, Germany) following manufacturer guidelines. Samples were eluted with 60 µL of elution buffer and up to eight extraction blanks were included per 96-well plate.

Arthropod COI gene fragments were targeted for amplification using primers detailed in Justino et al. <sup>32</sup>. We modified the original primer sequences to preserve the COI-specific regions, but integrated linker, pad, adapter, and barcode sequences into the oligo following Kozich et al. <sup>88</sup>. The construct below illustrates these modifications: underlined portions indicate the Illumina primer, the square brackets indicate the 8mer barcode, and the bold section indicates the COI-specific portion:

#### **ANMLxF:**

5' -AATGATACGGCGACCACCGAGATCTACAC [ATCGTACG] TATGGTAATTC**GGTCAACAAATCATAAAGATATTGG**-3'

#### **ANMLxR:**

5' -CAAGCAGAAGACGGCATACGAGAT [AACTCTCG] AGTCAGTCAGCC**GGWACTAATCAATTTCCAAATCC**-3'



A complete list of all primer pairs used is available:

[https://raw.githubusercontent.com/devonorourke/nhguano/master/docs/primer\\_pairs.txt](https://raw.githubusercontent.com/devonorourke/nhguano/master/docs/primer_pairs.txt).

We used 25  $\mu$ L reactions with 10  $\mu$ L of extracted DNA, 1  $\mu$ L each of 10 mM forward and reverse primer pairs, and 13  $\mu$ L of AccuStart II PCR SuperMix (Quanta BioSciences, Gaithersburg, MD, USA). Mock community samples were amplified using a distinct primer pair that was not used on any guano samples for any libraries in this study. Furthermore, mock samples were amplified in a separate reaction from guano samples to avoid any potential cross contamination. For the mock samples, 1  $\mu$ L of template DNA (10 ng/ $\mu$ L) with 9  $\mu$ L of nuclease free water was used in lieu of 10  $\mu$ L of guano extract. Reaction conditions consisted of an initial 2 min denaturation at 95 °C, followed by 30 cycles of 20 s at 95 °C, 15 s at 50 °C, and 60 s at 72 °C and finally a 10 min extension at 72 °C.

PCR products were quantified using a PicoGreen assay (Invitrogen, Carlsbad, CA, USA) with a Tecan plate reader using excitation and emission wavelengths of 480 nm and 520 nm, respectively (Tecan Group, Männedorf, Switzerland). Samples were pooled in approximately equimolar ratios; because there were hundreds of samples pooled in a single library, the exact expected equimolar volumes were rounded to the nearest bin (bin sizes set at 0.5, 0.75, 1.0, 1.25, 1.5, 2.0, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5, 8.5, 9.5, 10, 15, or 20  $\mu$ L). Extraction blanks or guano samples with detectable DNA that exceeded the maximum pool bin volume were included at a fixed volume of 20  $\mu$ L. The

initial pool volume was reduced with a vacuum concentrator to approximately 2 mL and was cleaned with a QIAquick PCR purification kit (Qiagen, Hilden, Germany); libraries were eluted in 30 µL elution buffer. Libraries were quantified with a Qubit High Sensitivity assay (Thermo Fisher Scientific, Waltham, MA, USA) and fragment sizes were analyzed using TapeStation D1000 ScreenTape (Agilent Technologies, Santa Clara, CA, USA). A single mock community sample was included in eight of the nine libraries sequenced.

Three libraries containing samples from 2015 were sequenced at the University of New Hampshire on an Illumina (Illumina, San Diego, CA, USA) HiSeq platform using v2 chemistry with 500 cycles of 2x250 bp paired-end read lengths. Six libraries containing samples collected in 2016 were sequenced on a MiSeq machine at TGen North using v3 chemistry with 600 cycles of 2x300 bp paired-end reads. Raw sequence reads are available through NCBI: see BioProjects PRJNA518082 and PRJNA560640 (e.g. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA560640>).

### *Sequence processing*

Raw demultiplexed sequences were trimmed using Cutadapt v-2.3<sup>89</sup>. We imported trimmed paired-end reads into a QIIME 2 environment<sup>62</sup> and used the 'qiime demux summarize' function to visualize per-base quality scores. Libraries consisted of similarly high quality profiles for both forward and reverse reads, except the tail of each read diminished in quality because our amplicons were shorter than the length of the Illumina

reads. We used DADA2 v1.10.0 <sup>71</sup> to denoise and join paired end data with the QIIME 2 q2-dada2 function 'qiime dada2 denoise-paired', resulting in a particular set of representative amplicon sequence variants (ASVs) for each library. These library-specific sequences and ASV tables were merged into a single dataset. Complete details regarding sequencing processing steps are described at the GitHub repo for this project: [https://github.com/devonorourke/nhguano/blob/master/docs/sequence\\_processing.md](https://github.com/devonorourke/nhguano/blob/master/docs/sequence_processing.md).

### *Sequence quality control and bat host identification*

We included non-template controls (NTCs) and positive controls (a biological mock community) in each library and performed a contamination reduction to account for potential reagent contamination or platform cross talk. A portion of this workflow included classifying the sequence variants detected in the DADA2-pipeline using VSEARCH <sup>85</sup> alignment to a custom host database including bat species known to occupy the study area (principally New Hampshire, but also other states in New England and New York; detailed below). This purpose was two-fold: first, to remove any host or suspected chimeric sequences with host DNA from subsequent diet analyses; and second, because we passively collected samples below roosts were interested in determining whether the previous visual surveys of bats matched the molecular classifications. However, because few bat sequences were detected with this method, we classified the same dataset with a broader COI database using two different methods: VSEARCH and a Naive Bayes classifier available as a QIIME plugin <sup>87</sup>. These two approaches assigned the little brown bat to 595 of the 596 samples with bat-classified sequence variants (just one sample contained a different North American bat,

*Lasiurus borealis*). Furthermore, one *M. lucifugus* ASV was among the most abundant, generating approximately 5% of all sequences among the nine libraries in our study. Visual surveys conducted in 2010 at a few of the sampling sites detected a mix of little brown and big brown bat species inhabiting these sites, and the proportion of little brown sequences was disproportionate to some of the sites, despite a relatively even sampling intensity at other sites where few or no little brown sequence data were detected. Therefore, we proceeded with our analyses but acknowledge that other bat species may have contributed a small portion of these samples at certain sites. Our molecular data, combined with our own visual surveys, suggest that our diet analyses are exclusively from little brown bats at the following locations: ALS, CHI, CNB, EPS, FOX, HOL, PEN, WLD. Other sites with robust sampling intensity—particularly BRN, MAP, MTV, and HOP—that lack conclusive molecular information may contain fecal samples from other bat species, most likely the big brown bat, although none of our sequence information identified this species. We documented the sequence quality control steps in a separate document:

[https://github.com/devonorourke/nhguano/blob/master/docs/decontam\\_workflow.md](https://github.com/devonorourke/nhguano/blob/master/docs/decontam_workflow.md).

In addition, we provide information regarding the COI database designs as well for host:

[https://github.com/devonorourke/nhguano/blob/master/docs/hostCOI\\_database\\_design.md](https://github.com/devonorourke/nhguano/blob/master/docs/hostCOI_database_design.md).

We also include information for the broad COI database:

[https://github.com/devonorourke/nhguano/blob/master/docs/broadCOI\\_database\\_design.md](https://github.com/devonorourke/nhguano/blob/master/docs/broadCOI_database_design.md).

### *Diversity estimates*

Mock and negative control samples were removed from the original merged dataset and the remaining bat guano samples were then used as input into a pair of QIIME 2

functions to determine an appropriate sampling depth for normalization. We generated a summary table with 'qiime feature-table summarize' and calculated bootstrapped estimates of richness per sample across a range of sampling depths from 500 to 5000 reads with the 'qiime diversity alpha-rarefaction' function. Samples were then rarefied to our selected sampling depth of 1000 reads, a depth that retained as many samples with as much sequence diversity as possible. Among these remaining samples, we grouped collection periods into 14-day windows and retained only those samples where there were at least two samples at a given site and window group. Following rarefying and site filtering, 18 of 20 sites had sufficient sampling depth and breadth for further analysis (data from ROL and GRN were discarded) and retained a median of 5 samples per site per window for remaining sites.

All diversity estimates were performed with QIIME 2. The remaining rarefied samples were used to calculate richness, Shannon's entropy, and Faith's phylogenetic diversity<sup>190</sup>. Analysis of variance for each diversity measure were conducted and a post hoc Tukey summary was performed with R v-3.5.3<sup>93</sup> using RStudio v-1.2.1335<sup>191</sup>. We next calculated differences in community composition for nine sites from 2016 that were the most consistently sampled locations using weighted and unweighted metrics (Dice-Sorensen, Bray-Curtis, unweighted Unifrac, weighted Unifrac. We tested group significance of site and window (collection period) main effects in R using the Vegan 'adonis' function. Principal coordinates analysis was conducted for each distance estimate. The rooted tree used in the weighted-phylogenetic diversity measures was calculated using the QIIME 2 implementation of Fast Tree<sup>192</sup> and MAFFT<sup>193</sup>, while

Unifrac was used for phylogenetically-weighted distance estimates <sup>115,155–158</sup>. Additional documentation for the QIIME functions executed in these diversity analyses are available: [https://github.com/devonorourke/nhguano/blob/master/docs/diversity\\_analyses.md](https://github.com/devonorourke/nhguano/blob/master/docs/diversity_analyses.md).

A custom R script was used in statistical analyses and to create visualizations:

[https://github.com/devonorourke/nhguano/blob/master/scripts/r\\_scripts/NH\\_diversity.R](https://github.com/devonorourke/nhguano/blob/master/scripts/r_scripts/NH_diversity.R).

This same script was also used to compare the abundances of sequence variants among sites sampled in both years. Statistical analyses for all diversity measures are included in a single directory:

<https://github.com/devonorourke/nhguano/tree/master/data/stats>.

### *Pest Analysis*

To identify whether the taxa classified in our dataset were considered forest or agricultural pests, we cross referenced lists maintained by the United States Forest Service (USFS) and the United States Department of Agriculture (USDA). We used a custom R script to perform the comparisons, which were restricted to first identifying how many sequence variants were exact species matches, then expanding the search to identify instances in which common genera were shared. This was done because there were instances in which the USDA did listed taxa as ambiguous species (e.g. *Malacosoma* sp.) thus even if we classified our taxa to a species level, an exact match was not possible. See the custom script for details:

[https://github.com/devonorourke/nhguano/blob/master/scripts/r\\_scripts/pest\\_work.R](https://github.com/devonorourke/nhguano/blob/master/scripts/r_scripts/pest_work.R).

### *Additional software*

Note that additional software within the QIIME environment such as Pandas <sup>159</sup> and BIOM <sup>160</sup> were used, as was a suite of R libraries including ape <sup>163</sup>, decontam <sup>194</sup>, formattable <sup>167</sup>, qiime2R <sup>94</sup>, geofacet <sup>195</sup>, gganimate <sup>161</sup>, ggmap <sup>196</sup>, ggpubr <sup>105</sup>, ggrepel <sup>106</sup>, htmltools <sup>197</sup>, lubridate <sup>198</sup>, phyloseq <sup>97</sup>, qiime2R <sup>94</sup>, reshape2 <sup>96</sup>, scales <sup>109</sup>, tidyverse <sup>95</sup>, vegan <sup>98</sup>, webshot <sup>199</sup>.

## RESULTS

### *Dietary breadth*

In collaboration with citizen scientists at 20 New Hampshire locations (**Figure 15**), we passively collected nearly 2,500 bat guano samples each summer, from June to August 2015, and April to October 2016. We analyzed sequence data for over 900 of these samples at 18 sites (**Table S31**) that contained at least 1000 arthropod-classified reads. The final dataset includes 12 arthropod orders identified in at least 1% of samples, although the largest fraction of sequence and taxonomic diversity was observed among just three orders: Coleoptera, Diptera, and Lepidoptera (**Table 2**). Beetles, flies, and moths accounted for over 87% of sequences and 78% of sequence variants among these 12 arthropod orders. Among all bat fecal samples we observed many distinct genera within Coleoptera (25), Diptera (25), and Lepidoptera (26) orders, while we detected fewer genera in other orders like Ephemeroptera (7), Trichoptera (5), or Hemiptera (4) (**Table S32**). Importantly, most of the taxa that were consistently prevalent in diets throughout each season were beetle, fly, and moth taxa, indicating

that bats are foraging on a diverse assortment of many genera within these three orders at most sites and most times. Overall, the extensive diversity of prey consumed by these bats—particularly among beetle, fly, and moth taxa—is neither limited to a few locations or moments in the season.

### *Diversity estimates*

While we observed thousands of distinct sequence variants assigned to one of 12 arthropod orders, individual samples generally contained relatively few observed sequence variants (median=14, mean=20.2, SD±22.5). Among the top half of the most heavily sampled sites in 2016 (n=9) elevated richness was infrequently detected at different time intervals at different sites (**Figure 16**). Species richness differed significantly by site ( $F_{(8,451)} = 8.496$ , MSE = 2191.0,  $p < 0.001$ ) and collection date ( $F_{(15,451)} = 5.083$ , MSE = 1310.8,  $p < 0.001$ ) as well as the interaction term ( $F_{(54,451)} = 1.525$ , MSE = 393.2,  $p = 0.0126$ ). Overall, most had similar levels of richness throughout their sampling periods, and a Tukey's post hoc test suggested that significant differences (adjusted for multiple comparisons) were attributed to just three of the nine sites tested: CNB, MAP, and PEN (**Figure S34**). Each of these three sites had higher richness at distinct sampling weeks compared to other sites, suggesting that prey availability is likely site specific.

We also compared within-sample diversity between sites using measures that include abundance (Shannon's entropy) and phylogenetic (Faith's PD) information (**Figure S23**). Shannon's diversity differed significantly by site ( $F_{(8,451)} = 7.125$ , MSE = 6.893,  $p$



< 0.001) and collection date ( $F_{(15,451)} = 3.053$ ,  $MSE = 2.954$ ,  $p < 0.001$ ) as well as the interaction term ( $F_{(54,451)} = 1.486$ ,  $MSE = 1.438$ ,  $p = 0.0179$ ). Most pairwise differences in Shannon's diversity between sites were not significant at the  $p < 0.05$  significance level (**Figure S35**). Faith's PD differed significantly by site ( $F_{(8,451)} = 7.125$ ,  $MSE = 6.893$ ,  $p < 0.001$ ) and collection date ( $F_{(15,451)} = 3.053$ ,  $MSE = 2.954$ ,  $p < 0.001$ ) as well as the interaction term ( $F_{(54,451)} = 1.486$ ,  $MSE = 1.438$ ,  $p = 0.0179$ ). Unlike Shannon's or observed diversity estimates, half (18 of 36) of the pairwise comparisons for Faith's PD estimates were significant at the  $p < 0.05$  alpha level (**Figure S36**).

Notably, inferences using abundances based on sequence counts may be complicated by the large proportion of sequences attributed to just one or two ASVs per sample (**Figure S24**). Even among ASVs repeatedly sequenced in many samples, we found a bimodal distribution of proportions of reads per sample (**Figure S25**). Sampling guano from single pellets can result in some instances of relatively high diversity, though it appears that most samples contain only a minor fraction of the overall richness of the entire dataset, and that distributions of sequence counts are often highly uneven.

Our repeated sampling efforts also provided an opportunity to examine variation in prey composition among sites and years. Among sites surveyed in both 2015 and 2016, the most prevalent ASVs in any single year were also detected in both years, while ASVs detected only in either 2015 or 2016 only were detected in no more than four samples in a year (**Figure S26**). Many ASVs detected in both years are beetles (e.g. ASV's 1 and 5 are present in each of the six locations both years). Other arthropods were repeatedly

observed at particular sites with varying detection frequencies. For example, trichopteran ASVs classified as *Phryganea sayi* were identified in 18 samples in one location (HOP), 7 samples at another (MTV), but only in three or fewer samples at six other locations.

We observed substantial variation among arthropod orders across sites and sampling windows whether we treated these data as relative abundances of sequences (**Figure S27**) or by transforming sequences into a presence-absence context and evaluating the proportions of order-level detections (**Figure 17**). Among all samples, beetle, moth, and fly taxa were generally prevalent, however the particular proportions of these and other arthropod orders varied in distinct ways among collection sites and dates. For example, sequences may have largely been from beetles in most samples at a site for both years (e.g. MAP), or contained a more diverse mixture of order-level compositions between years (e.g. FOX) (**Figure 17**). Some sites may have contained a larger fraction of an order not widely detected among other sites. For instance, samples from Holderness (HOL) in 2016 repeatedly contain a larger fraction of Ephemeroptera than most other sites. However, the relatively larger fraction of mayflies are more frequent earlier in the season at HOL, in contrast to the larger proportion increasing towards mid-summer in Cornish (COR). There is less seasonal variability within a given site and year when treating data as proportions of detections (**Figure 17**) rather than abundances (**Figure S27**), but order-level compositions among locations are distinct for both analyses, and highlight a trend of site-specific foraging patterns. Whether data are viewed in terms of abundances of or detections of arthropod orders, bats foraging in New Hampshire

landscapes consume a diverse diet in distinctly local fashions that change throughout the season.

While order-level comparisons reveal broad dietary patterns, they lack the resolution to detect if the same taxa within an order are repeatedly consumed spatially and temporally. We tested whether the community composition of ASVs varied among the nine most intensively sampled locations in 2016, and found significant main effects for site ( $R^2 = 0.09$ ,  $p \leq 0.001$ ) and sampling date ( $R^2 = 0.01$ ,  $p = 0.002$ ), as well as their interaction ( $R^2 = 0.05$ ,  $p \leq 0.001$ ) using weighted Unifrac distance estimates. We found similar significant main effect and interaction patterns among the three other metrics tested (**Table S33**), although weighted Unifrac explained a larger proportion of variation on the first two principal components axes (45.5%) than Dice-Sorensen (14.5%), Bray-Curtis (20.1%), or unweighted Unifrac (17.7%) estimates (**Figure S28**). These distance measures collectively suggest that there is no simple association explaining diet composition across spatial or temporal dimensions. Nevertheless, the relatively larger proportion of variation captured by the weighted Unifrac measure suggests that both sequence variant abundance and evolutionary relationships are important factors associated with shared sampling date and locations. Indeed, the resulting biplot ordination using the weighted Unifrac distances suggest that some particular sampling periods and locations are associated with particular taxa (**Figure 18**). For instance, a pair of Ephemeroptera species (ASV-15, *Stenonema femoratum*; ASV-17, *Hexagenia limbata*) are strongly associated with spring and summer samples in Holderness (HOL) and Penacook (PEN), two locations that border large water bodies (Squam Lake and

the Merrimack River, respectively). However, other samples with strong associations to particular ASVs did not share any obvious pattern of site specificity or sampling date. For example, scarab beetles of the genus *Phyllophaga* (ASVs 2, 7, 11) are prevalent among samples spanning multiple locations and dates. Overall, this major finding of variability in diet across space and time is observed in both order-level and ASV perspectives, though a molecular approach indicates an even greater variability because of instances in which distinct taxa within a common order are consumed between distinct sites or locations.

### *Pest detections*

Despite the extensive diversity among bats, a common theme emerged across locations and dates: bats consistently foraged on some pest species. In fact, the most abundant groups of arthropods detected in this study were known turf and forest beetle pests: the Asiatic garden beetle, *Maladera castanea*, and white grubs of the genus *Phyllophaga* (**Table 3**). While *M. castanea* (325 samples) and *Phyllophaga* spp. (469 samples) were repeatedly detected in this study, there were few total pest species detected overall, and all other pests detections were rare—typically less than 1% of all samples. In addition, all were known to exist in the Northeast, though several had not been recorded in New Hampshire specifically, according to records available through the Insect and Arachnid Collections at the University of New Hampshire<sup>200</sup>. We broadened our search parameters to include those genera observed in our dataset that matched the genera listed as pests by the USFS and USDA: these need not be exact species matches, and only those instances where exact species matches exist are denoted in the table. Many

of the taxa observed in bat diets that matched to some pest genera were unable to be classified to species-level (these are labeled in the table as 'Ambiguous'), illustrating one of the challenges in using short amplicon sequences. While smaller fragments are less likely to be degraded, there are fewer base pairs to discriminate among closely related taxa. In some cases, it may be evident that this ambiguous classification is not a large concern. For example, while there are five known species in the *Coleotechnites* genus in New Hampshire, the significant forest pest, *Coleotechnites milleri*, is not one of them, and the host tree does not exist on the east coast. However, in other cases such as with the genus *Dendroctonus* it is not clear whether this ambiguous sequence is derived from one of the three endemic bark beetles, or if it represents a more concerning species such as *D. mexicanus*. Collectively, these results demonstrate bats providing ecosystem services across New Hampshire, and illustrate that guano sampling may be an important tool for detecting both expected and unknown forest and agricultural pests.

## DISCUSSION

Our study reaffirmed that bats in New Hampshire are foraging predominantly on beetles, flies, and moths described in earlier studies using morphological analyses<sup>176,189</sup>. However, our molecular approach illustrates that bats consume hundreds of unique species from at least 12 orders of insects and spiders, indicating a more expansive niche-breadth than previously understood (**Table 2**). Metabarcoding

techniques can reveal previously unrecognized taxa<sup>18,20</sup>, and indeed we identified taxa among Blattodea, Psocodea, and Megaloptera that were not identified using morphological approaches, despite many of the species commonly occurring in New Hampshire. For example, cockroaches such as *Periplaneta fuliginosa* and *Parcoblatta pennsylvanica*, or fishflies such as *Chauliodes pectinicornis*, are endemic to the state and were frequently detected in our guano analyses. Our molecular approach provides strong evidence that these insectivorous bats are capable of consuming a highly diverse assortment of prey.

Bat diets are not only diverse, but our longitudinal, multi-site analysis suggests that the particular composition of prey varies depending on when and where a sample is collected, consistent with earlier studies by Clare<sup>57-59</sup>. A detection-based perspective of order-level compositions suggested that bats in most sites are consuming some proportion of flies and beetles for most sampling periods, but the particular proportions could vary distinctly among different sites across sampling weeks (**Figure 17**). For example, while samples from a site in east Canterbury (CNB) in 2016 shifted from flies to beetles in early to mid-summer, another site just 12 km west (PEN) contained a larger proportion of beetle detections than flies over the same period. Among the sites in 2016 that demonstrated a reduction in beetle detections towards late-summer and fall, we observed increased detections for other orders such as flies, caddisflies, moths, barklice, and true bugs, but the particular proportions among taxa differed among sites and sampling periods.

Compositional differences are further complicated by the fact that bat diets comprise a mixture of common and rare taxa. Some genera of prey are detected in most sites and in most sampling periods, and while many of these genera are beetles, some taxa in other orders are also quite common (**Table S32**). Because we lacked information about prey availability, it is unclear if the increased frequency of particular genera reflect dietary preferences or are simply more abundant among collection locations and dates. For examples of studies measuring prey availability, see Vesterinen et al.<sup>123</sup> and Agosta et al.<sup>174</sup>, where they found particular arthropod preferences for individual bat species. Nevertheless, some insects have relatively limited adult lifespans, and we suspect that the changes in detection frequencies for some taxa may reflect opportunistic foraging. For instance, mayflies such as *Hexagenia limbata*, *Stenonema femoratum*, and *Stenacron interpunctatum* are known to emerge in swarms as adults, and we detect these species in multiple samples in late spring at some locations (e.g. PEN and HOL), in early summer at others (e.g. HOL and CNB), and in late summer and fall at other locations (e.g. ALS, HOP). We observed that bat diets contain some shared components—most samples will have some fraction of beetles, flies, and moths detected—but their particular compositions appear to be locally and seasonally sensitive, consistent with what has been previously reported for little brown<sup>57,58</sup> and big brown<sup>59</sup> bats.

Greater diversity occurred within and among samples at the level of individual ASVs as compared to arthropod orders. There were hundreds of potential species consumed by bats throughout the entire study, for instance (**Table 2**). However, most samples

generally contained only a small fraction of these ASVs—mean ASVs per sample among sites ranged from a median of 7 to 32 ASVs (**Figure 16**). The low observed species richness per sample relative to the larger pool of ASVs detected throughout the entire study may be attributed to the fact that we chose to sequence individual pellets rather than bulk samples. Nevertheless, these richness values are similar to earlier studies of little brown <sup>58</sup> and big brown <sup>59</sup> diets that used a bulk sampling approach, as well as other North American insectivore richness values sampling from individual bats <sup>25</sup>.

We also evaluated alpha diversity using Shannon's entropy and Faith's PD to compare observed richness to other measures that include abundance and phylogenetic information. The highly uneven distribution of sequence counts among ASVs likely contributed to the increased variability in observed richness estimates among samples compared to Shannon's entropy. Interestingly, the observed richness was not necessarily indicative of dietary complexity as measured by Faith's PD values—a metric that incorporates sequence identities when evaluating the distinctiveness of the community of ASVs within a sample. For example, multiple samples from west Canterbury contained more than 30 ASVs in several sampling windows, yet the same samples generally had low Faith's PD values, because nearly all of these sequence variants were some type of beetle (**Figure 17**). In other sites, relatively low richness values were observed, but these samples contained some of the more evolutionarily diverse assortment of prey, such as in Holderness, where the few ASVs per sample observed consisted of a complex mix of mayflies, beetles, spiders, flies, moths, and



other orders. In other locations the trajectories of per-sample richness matched changes in Faith's PD values (e.g. FOX).

An ASV-level analysis can reveal changes in intrasample diversity at a fine scale that order-level summaries may otherwise obscure. Diversity, as measured by any of the metrics we explored, change across season, though the particular trajectories of these changes can vary among locations.

We found limited evidence for ASV compositions to reflect strong spatial or temporal associations. We focused on the most frequently sampled sites in 2016 and found that the main effects of site and sampling dates, as well as their interaction component, were significant for every distance estimate calculated (**Table S33**), though the effect sizes were generally small: residual  $R^2$  for each metric was  $\geq 0.85$ . These small differences reflect a challenge analyzing differences among samples using ASVs versus a more inclusive taxonomic ranks like arthropod order: while ASVs can elucidate differences between samples that are identical at an order level, just 28% of the ASVs we detected were present in more than one sample, and only 33% of ASVs were detected in multiple samples. The result is a very sparse matrix where most samples have few shared ASVs. Subsequent ordinations of distance matrices following principal components analysis illustrated that group differences are greatest among samples collected towards the beginning or end of the sampling season (**Figure S28**). Additionally, we found that including sequence abundances and phylogenetic information captured the more variation among the first two principal components than analyses using neither or

just one of the two information types, indicating that while ASV-level analyses can exacerbate differences among samples compared to order-level comparisons. Using information about abundances and evolutionary relatedness may prove helpful when evaluating compositional data. For example, the biplot created using distances calculated from a weighted Unifrac metric illustrated that some of the most repeatedly detected taxa need not be associated to a particular site or date (**Figure 18**).

Nevertheless, despite the most variation being captured by the weighted Unifrac distance metric, further research is needed to better understand how to relate biomass of prey in bat guano to the observed sequencing output—feeding studies as well as biological mock experiments are two such experimental approaches that will prove informative.

In fact, incorporating quantitative analyses in animal diets remains an area of ongoing debate, as many potential steps in a typical workflow have been shown to affect the relationship between proportions of biomass to the observed sequence counts in a sample <sup>19,20,26</sup>. However, development of novel arthropod COI primer sequences <sup>32</sup> have improved upon previously suspected biases <sup>31</sup>, reducing the likelihood that differences in sequence counts are a result of preferential annealing. In fact, we found little evidence of taxonomic bias using the primers developed by Jusino et al. <sup>32</sup>. Multiple orders have the capacity to generate abundant reads for individual ASVs (**Figure S24-S25**), and while many of the most highly amplified and frequently detected ASVs were beetles, flies, and moths, 10 of 12 orders have at least one ASV among the 95th percentile of per-ASV abundance (~5100 sequences per ASV). Additionally, among the

most highly sequenced samples (the most abundant sample in three independent weeks per site), the most abundant ASV within a given sample could be from one of seven orders. Therefore, while we were more likely to detect a highly abundant beetle, fly, or moth sequence, this likely had more to do with prey consumption than primer bias.

Interestingly, a few of the most prevalent ASVs were pests. Bats are known to provide ecosystem services <sup>204</sup>, and insectivorous bats can provide significant economic benefits <sup>205</sup> as consumers of pests in a variety of agricultural and forested environments <sup>32,169,206,207</sup>. We sampled in sites throughout New Hampshire that encompassed a mix of land cover types including beech/oak forests, deciduous and evergreen mixed forests, agricultural lands, and water bodies <sup>208</sup>. Two pests we observed repeatedly: the Asiatic garden beetle (325 samples detected) and a genus of white pine grubs (429 samples). These are known turf and forests pests, respectively, but are of limited concern and both were known to exist in New Hampshire (**Table 3**). Other exact pest species matches were generally rare and included a mix of agricultural pests such as forest tent caterpillar (22 samples) and forest pests like the White-spotted pine sawyer (23 samples). Nevertheless, while we did not observe a large number of pest species, we are likely underestimating the number of pests these bats consumed because identifying exact pest matches is impeded by the procedure used to classify the ASVs. Many ASVs could be classified only to genus or family level because multiple best alignment references contained disparate taxonomic identities. As a result, only their shared information was retained and the remaining ambiguous portion was unclassified

(see the 'Ambiguous' species detected in **Table 3**). Thus, there are likely pests that happen to share sufficient sequence homology to closely related non pests that make identifying particular pest species unlikely using these molecular techniques. Finally, while few overall species matches were obtained, we obtained a more diverse list by expanding the search to include common genus-level matches for known pests. It is uncertain if bats would consume the non-native species of their endemic relatives, but assuming they share similar life histories, it appears likely given the flexible foraging capacity exhibited by these bats. Because of the ease with which guano is collected and the relatively inexpensive manner that sequence data can be obtained, molecular analyses can be used to provide both a broad characterization of the species diet and act as an initial screen for potential pests of concern.

Collecting guano each week at multiple sites was necessary to examine the extent of local variation in bat foraging patterns, yet it would not have been possible without collaborating with many dedicated citizen scientists. There is a paucity of molecular diet information for most North American bats (and indeed, bats globally), and connecting interested volunteers with researchers can rapidly expand our basic understanding of what prey these bats are consuming at the local level. These partnerships may be essential because bats often occupy private landscapes. For example, the majority (73%) of forested land in New Hampshire is owned privately <sup>209</sup>, thus a study interested in using bat guano to broadly surveil forest pests is likely going to require the cooperation with its citizens. Our partnerships with volunteers resulted in the discovery of highly flexible foraging patterns in the little brown bat in New Hampshire. This is a

positive sign for these populations that have survived white-nose syndrome and may be on the road to recovery. Molecular diet analyses can provide valuable information to rapidly monitor the health of these and other bat populations, particularly as we gather data on how their diets may be changing over time.

## TABLES

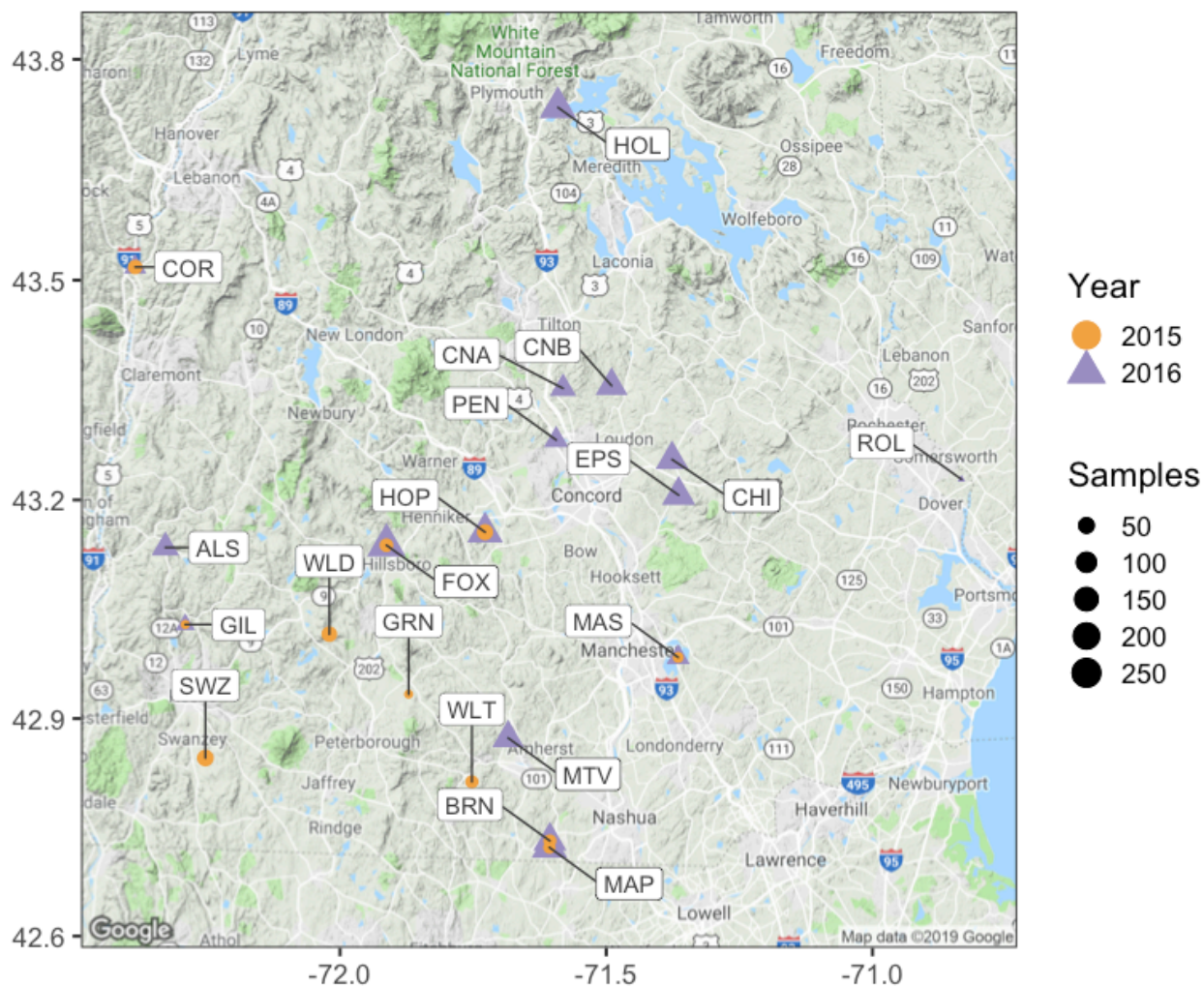
Order	Samples	Family	Genus	Species	ASV
Coleoptera	830	37	107	138	1578
Diptera	705	43	111	117	1113
Lepidoptera	644	37	165	185	728
Trichoptera	335	13	26	46	198
Ephemeroptera	299	9	20	34	182
Hemiptera	238	15	33	36	175
Hymenoptera	146	7	24	12	98
Megaloptera	111	2	2	4	85
Araneae	128	11	24	34	73
Psocodea	78	8	13	11	38
Blattodea	97	3	6	5	22
Neuroptera	29	5	7	7	13

**Table 2.** Dietary breadth of New Hampshire bats. Consumed prey among 2015 and 2016 samples collected from New Hampshire bats includes 12 distinct arthropod orders present in at least 1% of samples. Coleoptera, Diptera, and Lepidoptera orders contained the greatest proportion of taxonomic diversity, while Trichoptera, Ephemeroptera, Hemiptera, and Araneae orders each had more than twenty species detected. While megalopteran taxa contained many ASVs, these sequence variants were nearly completely classified to just three species, indicating that while this order is frequently sampled by bats, the same few species of fishflies are repeatedly targeted.

Order	Genus	Samples	Sites	Species detected
Coleoptera	<b>Phyllophaga</b>	469	18	<i>P. anxia</i> + (213), <i>crassissima</i> (1), <i>drakii</i> + (12), <i>fervida</i> (1), <i>fosteri</i> + (4), <i>foxii</i> (1), <i>hirsuta</i> (257), <i>hirticula</i> + (29), <i>ilicis</i> (1), <i>marginalis</i> + (10), <i>tristis</i> + (28), Ambiguous (390)
Coleoptera	Maladera	325	18	<b><i>M. castanea</i></b> + (325)
Coleoptera	Monochamus	45	12	<b><i>M. scutellatus</i></b> + (23), Ambiguous (22)
Lepidoptera	Malacosoma	28	13	<i>M. americanum</i> + (1), <b><i>disstria</i></b> + (22), Ambiguous (5)
Lepidoptera	Epinotia	27	12	<i>E. criddleana</i> (5), <i>medioviridana</i> + (1), <i>nisella</i> + (1), <i>solicitana</i> + (12), <i>transmissana</i> + (3), Ambiguous (7)
Lepidoptera	Synanthedon	24	10	<i>S. acerni</i> + (24)
Lepidoptera	Acleris	21	10	<i>A. chalybeana</i> + (1), <i>logiana</i> + (2), <i>maccana</i> (1), <i>schalleriana</i> + (2), <i>semipurpurana</i> + (3), Ambiguous (12)
Lepidoptera	Coleotechnites	19	8	<i>C. atrupictella</i> + (3), <i>piceaella</i> + (4), Ambiguous (15)
Coleoptera	Dendroctonus	18	9	<b><i>D. terebrans</i></b> (1), Ambiguous (17)
Lepidoptera	Choristoneura	14	10	<i>C. conflictana</i> +* (1), <i>fractivittana</i> + (1), <i>rosaceana</i> + (2), Ambiguous (10)
Lepidoptera	Coleophora	14	4	Ambiguous (14)
Hemiptera	Aphrophora	12	8	<i>A. canadensis</i> (8), <i>cribrata</i> + (4), <i>parallela</i> + (1), Ambiguous (10)

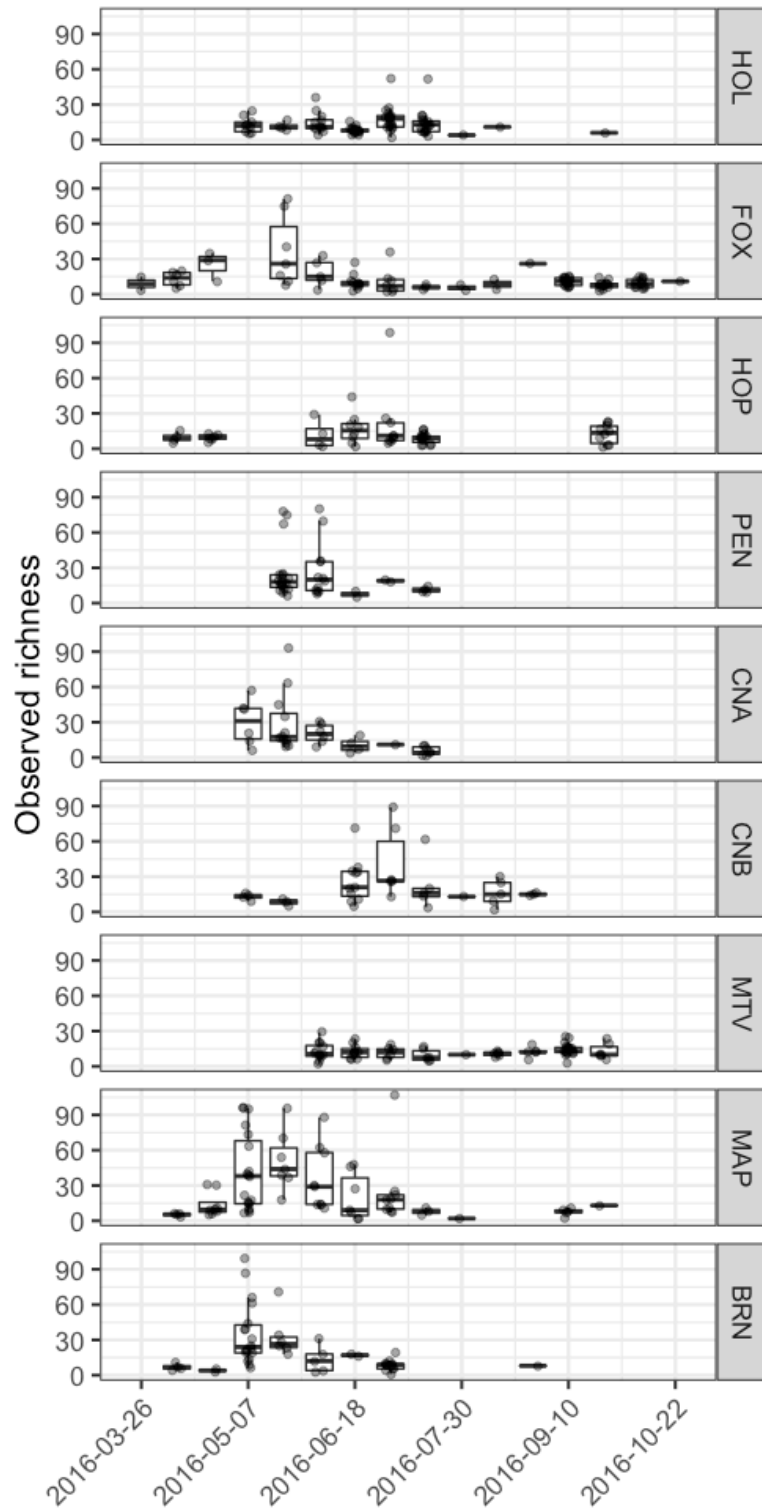
**Table 3.** Pest genera detected in bat guano. Sequence variants classified to genera that matched those listed by the US Forest Service or US Department of Agriculture are shown below. Exact species matches are highlighted in bold, and species endemic to New Hampshire are denoted (+). The most prevalent pest genera, *Phyllophaga*, is listed by the US Forest Service as a complex group, thus no single species is highlighted. Species-level classification is limited with this particular COI gene fragment, as many genera have more Ambiguous than named species. Numbers of samples detected for each species listed in parentheses. Insect pest surveillance with bat guano casts a wide taxonomic net, but additional target-specific markers are needed for species-level resolution.

## FIGURES

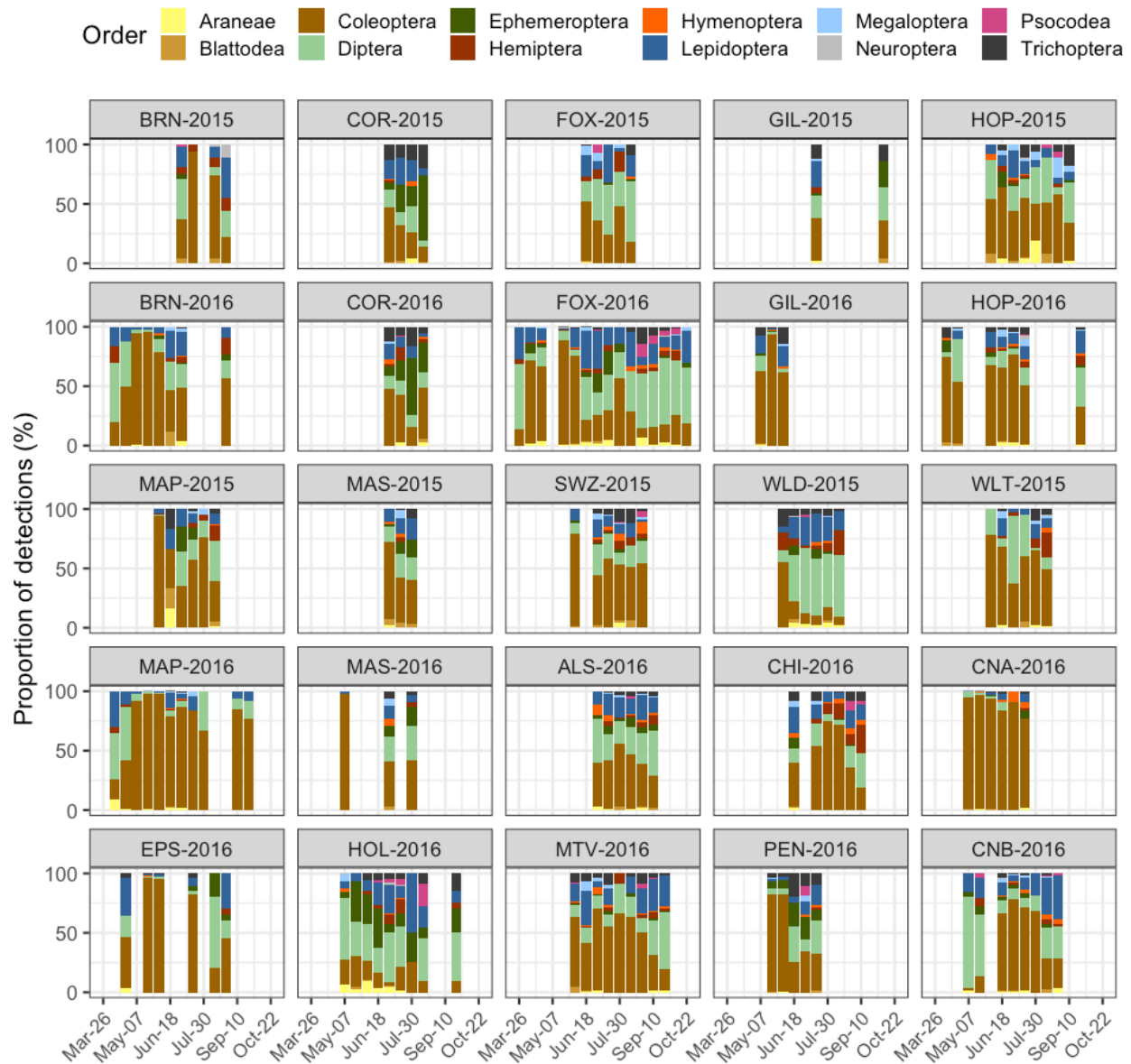


**Figure 15.** Locations for each of the 20 collection sites. Sampling years (shape and color) and number of samples collected each year (point size) are indicated at each geographic position. Guano was collected from bats inhabiting barns, garages, or bat houses. While many sites were within 1 km of some water body, three locations were directly abutting ponds or lakes: (HOL), (WLD), and (MAS). Site codes represent the following New Hampshire towns: ALS, Alsted; BRN, Brown Lane, Hollis; CHI, Chichester; CNA, Canterbury; CNB, Canterbury; COR, Cornish; EPS, Epsom; FOX, Fox State Forest, Hillsborough; GIL, Gilsum; GRN, Greenfield; HOL, Squam Science Center, Holderness; HOP, Hopkinton; MAP, Maple Hill, Hollis; MAS, Massabesic Audubon Center, Auburn; MTV, Mont Vernon; PEN, Penacook; ROL, Rollinsford; SWZ, Swanzey; WLD, Willard Pond, Antrim; WLT, Wilton.

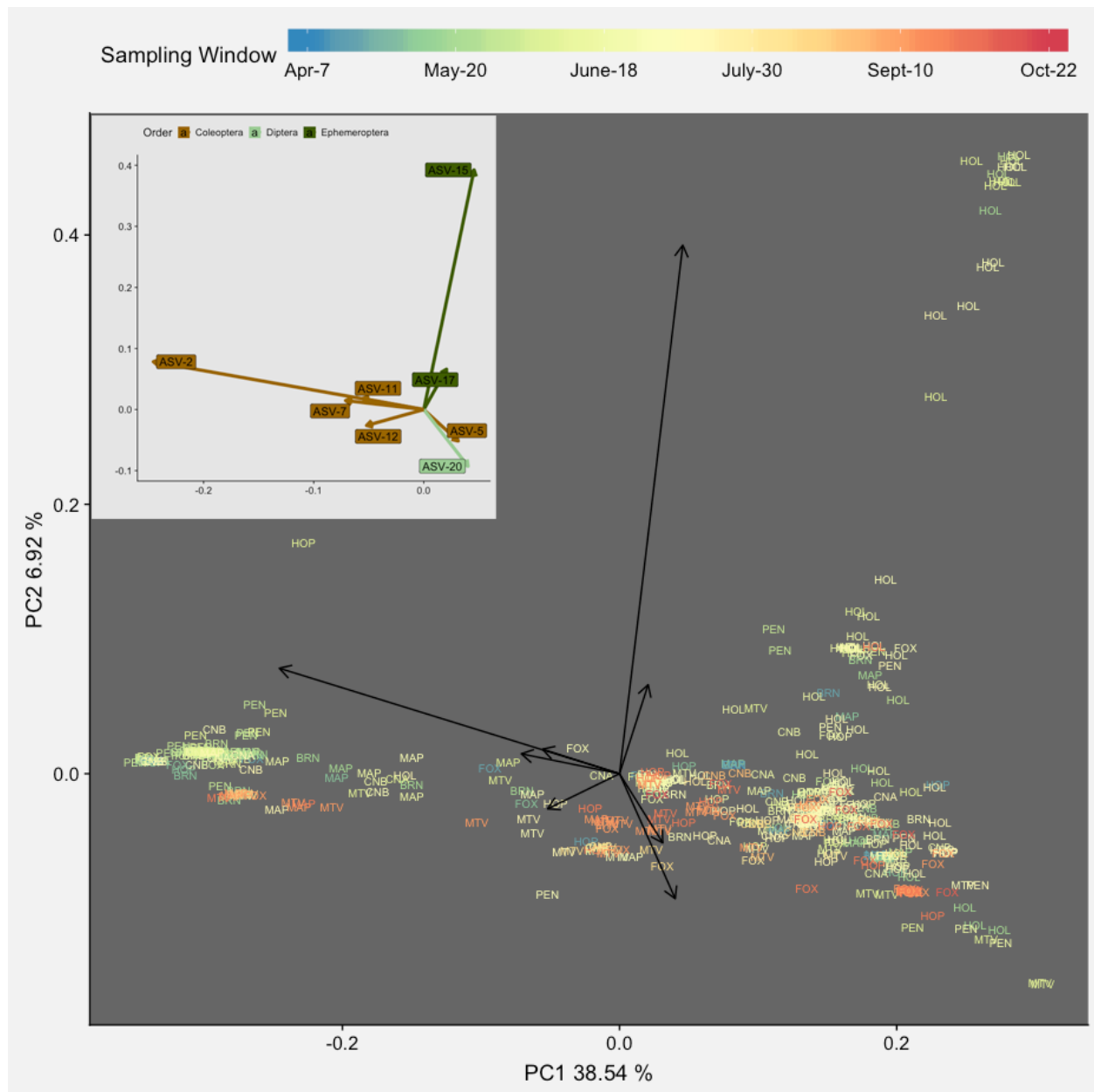




**Figure 16.** Observed richness per sample among select 2016 NH sites. Most samples have relatively low richness, though some sites contain instances with elevated richness. These can occur at distinct sampling periods among sites, likely indicating a local opportunity of increase prey availability.



**Figure 17.** Relative proportions of arthropod COI shown for each site and year across sampling windows. Sampling windows (vertical bars) are grouped in 14-day intervals; facet labels indicate sampling site and year. Read abundances are transformed into binary detections, and proportions are calculated as the number of sequence variants detected within an order relative to the total number of detections of all orders across samples in shared site and sampling window groups. Overall patterns suggest distinctly local diet foraging patterns: among 2016 sites, samples can be a balance of beetles, flies, and moths (e.g. FOX) or predominantly beetle (e.g. MAP or CNA). While the overall proportion of beetle detections in 2016 samples are reduced into late-summer and fall, distinct orders replace beetle taxa at different sites: FOX is replaced largely with flies, CNB has more moths, while CHI has more complex mix of true bugs, caddisflies, and barklice.



**Figure 18.** Principal coordinates analysis biplot using weighted Unifrac. The PCoA biplot shows the relationship of arthropod COI sequence composition for the dates in which selected 2016 samples were collected (color) at each location (abbreviated text). These nine sites examined were among the most intensively sampled 2016 locations with the most consistent overlap in sampling windows. The black arrows indicate the relative magnitude of species scores (i.e. individual ASVs); the inset figure denotes the particular sequence variant and its assigned arthropod Order. Samples collected in early spring in Holderness (HOL) for example, are most clearly associated with particular mayfly species. While the first principal component explains a large fraction of variation of the data, a PERMANOVA test of main effects for location ( $r = 0.09$ ,  $P \leq 0.005$ ), date ( $r \leq 0.01$ ,  $P \leq 0.005$ ), and their interaction ( $r = 0.09$ ,  $P \leq 0.005$ ) suggest that the diets are extremely variable throughout the season, even among sites in relatively close (< 10 km) proximity.

## REFERENCES

1. Holechek, J. L., Vavra, M. & Pieper, R. D. Botanical Composition Determination of Range Herbivore Diets: A Review. *J. Range Manage.* **35**, 309–315 (1982).
2. Rosenberg, K. V. & Cooper, R. J. Approaches to avian diet analysis. *Studies in avian biology* **13**, 80–90 (1990).
3. Whitaker, J. O., McCracken, G. F. & Björn, M. S. Food Habits Analysis of Insectivorous Bats. in *Ecological and Behavioral Methods for the Study of Bats* (eds. Kunz, T. K. & Parsons, S.) 567–592 (2011).
4. Pierce, G. J. & Boyle, P. R. A review of methods for diet analysis in piscivorous marine mammals. *Oceanogr. Mar. Biol. Annu. Rev.* **29**, 409–486 (1991).
5. Foster, J. T. & Robinson, S. K. Introduced birds and the fate of hawaiian rainforests. *Conserv. Biol.* **21**, 1248–1257 (2007).
6. Paquet, P. C. Prey Use Strategies of Sympatric Wolves and Coyotes in Riding Mountain National Park, Manitoba. *J. Mammal.* **73**, 337–343 (1992).
7. Sherry, T. W. Comparative Dietary Ecology of Sympatric, Insectivorous Neotropical Flycatchers (Tyrannidae). *Ecol. Monogr.* **54**, 313–338 (1984).
8. Orłowski, G. & Karg, J. Diet breadth and overlap in three sympatric aerial insectivorous birds at the same location. *Bird Study* **60**, 475–483 (2013).
9. Ganey, J. L. & Block, W. M. Dietary overlap between sympatric Mexican spotted and great horned owls in Arizona. *Research Paper RMRS-RP-57. Fort Collins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Research Station.* 9 p. **57**, (2005).
10. Fukui, D., Okazaki, K. & Maeda, K. Diet of three sympatric insectivorous bat species on Ishigaki Island, Japan. *Endanger. Species Res.* **8**, 117–128 (2009).
11. Lopez, J. E. & Vaughan, C. Food niche overlap among neotropical frugivorous bats in Costa Rica. *Rev. Biol. Trop.* **55**, 301–313 (2007).
12. Ober, H. K. & Hayes, J. P. Prey Selection by Bats in Forests of Western Oregon. *J. Mammal.* **89**, 1191–1200 (2008).
13. Nielsen, J. M., Clare, E. L., Hayden, B., Brett, M. T. & Kratina, P. Diet tracing in ecology: Method comparison and selection. *Methods Ecol. Evol.* **9**, 278–291 (2018).
14. Kunz, T. H. & Whitaker, J. O., Jr. An evaluation of fecal analysis for determining food habits of insectivorous bats. *Can. J. Zool.* **61**, 1317–1321 (1983).

15. Alberdi, A. *et al.* Promises and pitfalls of using high-throughput sequencing for diet analysis. *Mol. Ecol.* (2019).
16. King, R. A., Read, D. S., Traugott, M. & Symondson, W. O. C. INVITED REVIEW: Molecular analysis of predation: a review of best practice for DNA-based approaches. *Mol. Ecol.* **17**, 947–963 (2008).
17. Symondson, W. O. C. Molecular identification of prey in predator diets. *Mol. Ecol.* **11**, 627–641 (2002).
18. Clare, E. L. Molecular detection of trophic interactions: emerging trends, distinct advantages, significant considerations and conservation applications. *Evol. Appl.* **7**, 1144–1157 (2014).
19. Alberdi, A., Aizpurua, O., Gilbert, M. T. P. & Bohmann, K. Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods Ecol. Evol.* **9**, 134–147 (2018).
20. Pompanon, F. *et al.* Who is eating what: diet assessment using next generation sequencing. *Mol. Ecol.* **21**, 1931–1950 (2012).
21. Valentini, A., Pompanon, F. & Taberlet, P. DNA barcoding for ecologists. *Trends Ecol. Evol.* **24**, 110–117 (2009).
22. Kartzinel, T. R. *et al.* DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 8019–8024 (2015).
23. Emrich, M. A., Clare, E. L., Symondson, W. O. C., Koenig, S. E. & Fenton, M. B. Resource partitioning by insectivorous bats in Jamaica. *Molecular Ecology* **23**, 3648–3656 (2014).
24. Wirta, H. K. *et al.* Exposing the structure of an Arctic food web. *Ecol. Evol.* **5**, 3842–3856 (2015).
25. Gordon, R. *et al.* Molecular diet analysis finds an insectivorous desert bat community dominated by resource sharing despite diverse echolocation and foraging strategies. *Ecol. Evol.* **9**, 3117–3129 (2019).
26. Deagle, B. E. *et al.* Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data? *Mol. Ecol.* (2018). doi:10.1111/mec.14734
27. Deagle, B. E., Eveson, J. P. & Jarman, S. N. Quantification of damage in DNA recovered from highly degraded samples--a case study on DNA in faeces. *Front. Zool.* **3**, 11 (2006).
28. Ratnasingham, S. & Hebert, P. D. N. bold: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Mol. Ecol. Notes* **7**, 355–364 (2007).
29. Sun, S. 'e *et al.* DNA barcoding reveal patterns of species diversity among northwestern Pacific molluscs. *Sci. Rep.* **6**, 33367 (2016).
30. Pentinsaari, M., Salmela, H., Mutanen, M. & Roslin, T. Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life. *Sci. Rep.* **6**, 35275 (2016).
31. Clarke, L. J., Soubrier, J. & Weyrich, L. S. Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. *Mol. Ecol.* (2014).

32. Jusino, M. A. *et al.* An improved method for utilizing high-throughput amplicon sequencing to determine the diets of insectivorous animals. *Mol. Ecol. Resour.* **19**, 176–190 (2019).
33. Zeale, M. R. K., Butlin, R. K., Barker, G. L. A., Lees, D. C. & Jones, G. Taxon-specific PCR for DNA barcoding arthropod prey in bat faeces. *Mol. Ecol. Resour.* **11**, 236–244 (2011).
34. Razgour, O. *et al.* High-throughput sequencing offers insight into mechanisms of resource partitioning in cryptic bat species. *Ecol. Evol.* **1**, 556–570 (2011).
35. Robeson, M. S., 2nd *et al.* Assessing the utility of metabarcoding for diet analyses of the omnivorous wild pig (*Sus scrofa*). *Ecol. Evol.* **8**, 185–196 (2018).
36. Esnaola, A., Arrizabalaga-Escudero, A., González-Esteban, J., Elozegi, A. & Aihartza, J. Determining diet from faeces: Selection of metabarcoding primers for the insectivore Pyrenean desman (*Galemys pyrenaicus*). *PLOS ONE* **13**, e0208986 (2018).
37. Braukmann, T. W. A. *et al.* Metabarcoding a Diverse Arthropod Mock Community. *Mol. Ecol. Resour.* (2019). doi:10.1111/1755-0998.13008
38. Jachowski, D. S., Rota, C. T., Dobony, C. A., Ford, W. M. & Edwards, J. W. Correction: Seeing the Forest through the Trees: Considering Roost-Site Selection at Multiple Spatial Scales. *PLoS One* **12**, e0169815 (2017).
39. U.S. Fish and Wildlife Service. 2019 Indiana Bat (*Myotis sodalis*) Population Status Update. *U.S. Fish Wildl. Serv.* (2019).
40. Thogmartin, W. E., King, R. A., McKann, P. C., Szymanski, J. A. & Pruitt, L. Population-level impact of white-nose syndrome on the endangered Indiana bat. *J. Mammal.* **93**, 1086–1098 (2012).
41. Mangan, K. & Mangan, M. T. *2016 Indiana Bat Survey Report*. ({U.S. Fish and Wildlife Service}, 2019).
42. Mangan, K. & Mangan, M. T. *Bat Use of Artificial Roosting Structures 3rd Annual Report*. ({U.S. Fish and Wildlife Service}, 2017).
43. U.S. Fish and Wildlife Service. Cypress Creek National Wildlife Refuge Habitat Management Plan. (2014).
44. Sparks, D. W., Ritzi, C. M., Duchamp, J. E. & Whitaker, J. O. Foraging Habitat of the Indiana Bat (*Myotis sodalis*) at an Urban-Rural Interface. *J. Mammal.* **86**, 713–718 (2005).
45. Tuttle, N. M., Benson, D. P. & Sparks, D. W. Diet of the *Myotis sodalis* (Indiana bat) at an urban/rural interface. *Northeastern Naturalist* **13**, 435–443 (2006).
46. Feldhamer, G. A., Carter, T. C. & Whitaker, J. O., Jr. Prey Consumed by Eight Species of Insectivorous Bats from Southern Illinois. *Am. Midl. Nat.* **162**, 43–51 (2009).
47. Murray, S. W. & Kurta, A. Nocturnal activity of the endangered Indiana bat (*Myotis sodalis*). *J. Zool.* **262**, 197–206 (2004).

48. Frick, W. F. *et al.* Disease alters macroecological patterns of North American bats. *Global Ecology and Biogeography* **24**, 741–749 (2015).
49. Frick, W. F. *et al.* An emerging disease causes regional population collapse of a common North American bat species. *Science* **329**, 679–682 (2010).
50. Langwig, K. E. *et al.* Resistance in persisting bat populations after white-nose syndrome invasion. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**, (2017).
51. Dobony, C. A. *et al.* Little brown myotis persist despite exposure to white-nose syndrome. *Journal of Fish and Wildlife Management* **2**, 190–195 (2011).
52. Johnson, C. M. & King, R. A. Beneficial Forest Management Practices for WNS-affected Bats: Voluntary Guidance for Land Managers and Woodland Owners in the Eastern United States. *White-nose Syndrome National Plan* 39 (2018).
53. Szymanski, J. A., Runge, M. C., Parkin, M. J. & Armstrong, M. White-nose syndrome management: report on structured decision making initiative. *Department of Interior, US Fish and Wildlife Service, Fort Snelling, Minn* (2009).
54. U.S. Fish and Wildlife Service & Others. A national plan for assisting states, federal agencies, and tribes in managing white-nose syndrome in bats. (2011).
55. Maslo, B., Valent, M., Gumbs, J. F. & Frick, W. F. Conservation implications of ameliorating survival of little brown bats with white-nose syndrome. *Ecol. Appl.* **25**, 1832–1840 (2015).
56. Maslo, B. *et al.* High annual survival in infected wildlife populations may veil a persistent extinction risk from disease. *Ecosphere* **8**, e02001 (2017).
57. Clare, E. L., Barber, B. R., Sweeney, B. W., Hebert, P. D. N. & Fenton, M. B. Eating local: influences of habitat on the diet of little brown bats (*Myotis lucifugus*). *Mol. Ecol.* **20**, 1772–1780 (2011).
58. Clare, E. L. *et al.* The diet of *Myotis lucifugus* across Canada: assessing foraging quality and diet variability. *Mol. Ecol.* **23**, 3618–3632 (2014).
59. Clare, E. L., Symondson, W. O. C. & Fenton, M. B. An inordinate fondness for beetles? Variation in seasonal dietary preferences of night-roosting big brown bats (*Eptesicus fuscus*). *Mol. Ecol.* **23**, 3633–3647 (2014).
60. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).
61. Palmer, J. M., Jusino, M. A., Banik, M. T. & Lindner, D. L. Non-biological synthetic spike-in controls and the AMPtk software pipeline improve mycobiome data. *PeerJ* **6**, e4925 (2018).
62. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* (2019). doi:10.1038/s41587-019-0209-9
63. Bokulich, N. A. *et al.* mockrobiota: a Public Resource for Microbiome Bioinformatics Benchmarking. *mSystems* **1**, (2016).

64. Gohl, D. M. *et al.* Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat. Biotechnol.* **34**, 942–949 (2016).
65. Bokulich, N. A. *et al.* Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* **10**, 57–59 (2013).
66. Nearing, J. T., Douglas, G. M., Comeau, A. M. & Langille, M. G. I. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ* **6**, e5364 (2018).
67. Bokulich, N. A. *et al.* Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **6**, 90 (2018).
68. Beng, K. C. *et al.* The utility of DNA metabarcoding for studying the response of arthropod diversity and composition to land-use change in the tropics. *Sci. Rep.* **6**, 24965 (2016).
69. Galan, M. *et al.* Metabarcoding for the parallel identification of several hundred predators and their prey: Application to bat species diet analysis. *Mol. Ecol. Resour.* **18**, 474–489 (2018).
70. Divoll, T. J., Brown, V. A., Kinne, J., McCracken, G. F. & O'Keefe, J. M. Disparities in second-generation DNA metabarcoding results exposed with accessible and repeatable workflows. *Mol. Ecol. Resour.* **18**, 590–601 (2018).
71. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
72. Amir, A. *et al.* Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* **2**, (2017).
73. Rideout, J. R. *et al.* Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ* **2**, e545 (2014).
74. Glassman, S. I. & Martiny, J. B. H. Broad-scale Ecological Patterns Are Robust to Use of Exact Sequence Variants versus Operational Taxonomic Units. *mSphere* **3**, (2018).
75. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11**, 2639–2643 (2017).
76. Vesterinen, E. J., Puisto, A. I. E., Blomberg, A. S. & Lilley, T. M. Table for five, please: Dietary partitioning in boreal bats. *Ecol. Evol.* **8**, 10914–10937 (2018).
77. Kaunisto, K. M., Roslin, T., Sääksjärvi, I. E. & Vesterinen, E. J. Pellets of proof: First glimpse of the dietary composition of adult odonates as revealed by metabarcoding of feces. *Ecol. Evol.* **7**, 8588–8598 (2017).
78. Czenze, Z. J. *et al.* Spatiotemporal and demographic variation in the diet of New Zealand lesser short-tailed bats (*Mystacina tuberculata*). *Ecol. Evol.* **8**, 7599–7610 (2018).
79. Bohmann, K. *et al.* Using DNA metabarcoding for simultaneous inference of common vampire bat diet and population structure. *Mol. Ecol. Resour.* (2018). doi:10.1111/1755-0998.12891



80. Clare, E. L., Chain, F. J. J., Littlefair, J. E. & Cristescu, M. E. The effects of parameter choice on defining molecular operational taxonomic units and resulting ecological analyses of metabarcoding data. *Genome* **59**, 981–990 (2016).
81. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank. *Nucleic Acids Res.* **33**, D34–8 (2005).
82. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
83. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35**, 7188–7196 (2007).
84. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
85. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
86. Edgar, R. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv preprint 074161* (2016). doi:10.1101/074161
87. Bokulich, N. *et al.* q2-sample-classifier: machine-learning tools for microbiome classification and regression. *JOSS* **3**, 934 (2018).
88. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* **79**, 5112–5120 (2013).
89. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
90. Mata, V. A. *et al.* How much is enough? Effects of technical and biological replication on metabarcoding dietary analysis. *Mol. Ecol.* **28**, 165–175 (2019).
91. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl 1**, 4516–4522 (2011).
92. Thompson, L. R. *et al.* A communal catalogue reveals Earth’s multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
93. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2018).
94. Bisanz, J. E. *qiime2R: Importing QIIME2 artifacts and associated data into R sessions*. (2018).
95. Wickham, H. *tidyverse: Easily Install and Load the ‘Tidyverse’*. (2017).
96. Wickham, H. Reshaping Data with the reshape Package. *Journal of Statistical Software* **21**, 1–20 (2007).
97. McMurdie, P. J. & Holmes, S. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).

98. Oksanen, J. *et al.* *vegan: Community Ecology Package*. (2018).
99. Chao, A. *et al.* Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecol. Monogr.* **84**, 45–67 (2014).
100. Hsieh, T. C., Ma, K. H. & Chao, A. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods Ecol. Evol.* **7**, 1451–1456 (2016).
101. Bates, D. & Maechler, M. *Matrix: Sparse and Dense Matrix Classes and Methods*. (2018).
102. Ogle, D. H., Wheeler, P. & Dinno, A. *FSA: Fisheries Stock Analysis*. (2018).
103. Dinno, A. *dunn.test: Dunn's Test of Multiple Comparisons Using Rank Sums*. (2017).
104. Wilke, C. O. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. (2017).
105. Kassambara, A. *ggpubr: 'ggplot2' Based Publication Ready Plots*. (2018).
106. Slowikowski, K. *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'*. (2018).
107. Wilke, C. O. *ggridges: Ridgeline Plots in 'ggplot2'*. (2018).
108. Gagolewski, M. R package stringi: Character string processing facilities. (2019).
109. Wickham, H. *scales: Scale Functions for Visualization*. (2018).
110. Garnier, S. *viridis: Default Color Maps from 'matplotlib'*. (2018).
111. Porter, T. M. & Hajibabaei, M. Automated high throughput animal CO1 metabarcode classification. *Sci. Rep.* **8**, 4226 (2018).
112. Chamberlain, S. bold: interface to bold systems API. *R package* (2017).
113. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
114. Rodriguez-R, L. M. & Konstantinidis, K. T. Estimating coverage in metagenomic data sets and why it matters. *ISME J.* **8**, 2349–2351 (2014).
115. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27 (2017).
116. Chao, A. Nonparametric Estimation of the Number of Classes in a Population. *Scand. Stat. Theory Appl.* **11**, 265–270 (1984).
117. Willis, A. & Bunge, J. Estimating diversity via frequency ratios. *Biometrics* **71**, 1042–1049 (2015).
118. Willis, A. D. & Martin, B. D. DivNet: Estimating diversity in networked communities. *bioRxiv* 305045 (2018).

119. Jost, L. The Relation between Evenness and Diversity. *Diversity* **2**, 207–232 (2010).
120. Nilsson, R. H. *et al.* The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res.* **47**, D259–D264 (2019).
121. Porter, T. M. & Hajibabaei, M. Over 2.5 million COI sequences in GenBank and growing. *PLOS ONE* **13**, e0200177 (2018).
122. Ratnasingham, S. & Hebert, P. D. N. A DNA-based registry for all animal species: the barcode index number (BIN) system. *PLoS One* **8**, e66213 (2013).
123. Vesterinen, E. J. *et al.* What you need is what you eat? Prey selection by the bat *Myotis daubentonii*. *Mol. Ecol.* **25**, 1581–1594 (2016).
124. Almeida, A., Mitchell, A. L., Tarkowska, A. & Finn, R. D. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *Gigascience* **7**, (2018).
125. Gardner, P. P. *et al.* Identifying accurate metagenome and amplicon software via a meta-analysis of sequence to taxonomy benchmarking studies. *PeerJ* **7**, e6160 (2019).
126. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
127. Pearman, W. *et al.* New tools for diet analyses: nanopore sequencing of metagenomic DNA from stomach contents to quantify diet in an invasive population of rats. *bioRxiv* 363622 (2018). doi:10.1101/363622
128. Krehenwinkel, H. *et al.* Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. doi:10.1101/358572
129. Hall, D. H. Endangered and Threatened Wildlife and Plants; 90-Day and 12-Month Findings on a Petition To Revise Critical Habitat for the Indiana Bat. *Fed. Regist.* **72**, 5 (2007).
130. Thomson, C. E. *Myotis sodalis*. *Mammalian Species* **1** (1982). doi:10.2307/3504013
131. O'Shea, T. J. & Bogan, M. A. Monitoring trends in bat populations of the United States and territories: problems and prospects. *Publications of the US Geological* (2003).
132. U.S. Fish and Wildlife Service. Indiana Bat (*Myotis sodalis*) Draft Recovery Plan: First Revision. *U.S. Fish and Wildlife Service* (2007).
133. U.S. Fish and Wildlife Service. Recovery plan for the Indiana bat. *U.S. Fish Wildl. Serv.* (1983).
134. Lorch, J. M. *et al.* Experimental infection of bats with *Geomyces destructans* causes white-nose syndrome. *Nature* **480**, 376–378 (2011).
135. Warnecke, L. *et al.* Inoculation of bats with European *Geomyces destructans* supports the novel pathogen hypothesis for the origin of white-nose syndrome. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 6999–7003 (2012).

136. Turner, G. G., Reeder, D. & Coleman, J. T. H. A Five-year Assessment of Mortality and Geographic Spread of White-Nose Syndrome in North American Bats, with a Look at the Future. Update of White-Nose Syndrome in Bats. *Bat research news* **52**, 13 (2011).
137. Jachowski, D. S. *et al.* Disease and community structure: white-nose syndrome alters spatial and temporal niche partitioning in sympatric bat species. *Diversity and Distributions* **20**, 1002–1015 (2014).
138. Womack, K. M., Amelon, S. K. & Thompson, F. R. Resource selection by Indiana bats during the maternity season. *The Journal of Wildlife Management* **77**, 707–715 (2013).
139. Menzel, J. M. *et al.* RESEARCH NOTES: SUMMER HABITAT USE AND HOME-RANGE ANALYSIS OF THE ENDANGERED INDIANA BAT. *Journal of Wildlife Management* **69**, 430–436 (2005).
140. Callahan, E. V., Drobney, R. D. & Clawson, R. L. Selection of Summer Roosting Sites by Indiana Bats (*Myotis sodalis*) in Missouri. *J. Mammal.* **78**, 818–825 (1997).
141. Clark, B. K., Bowles, J. B. & Clark, B. S. Summer Status of the Endangered Indiana Bat in Iowa. *Am. Midl. Nat.* **118**, 32–39 (1987).
142. Humphrey, S. R., Richter, A. R. & Cope, J. B. Summer Habitat and Ecology of the Endangered Indiana Bat, *Myotis sodalis*. *J. Mammal.* **58**, 334–346 (1977).
143. Garner, J. D., Gardner, J. E. & Others. *Determination of summer distribution and habitat utilization of the Indiana bat (Myotis sodalis) in Illinois*. (Division of Natural Heritage, Illinois Department of Conservation, 1992).
144. Jachowski, D. S., Rota, C. T., Dobony, C. A., Ford, W. M. & Edwards, J. W. Seeing the Forest through the Trees: Considering Roost-Site Selection at Multiple Spatial Scales. *PLoS One* **11**, e0150011 (2016).
145. Clare, E. L., Fraser, E. E., Braid, H. E., Fenton, M. B. & Hebert, P. D. N. Species on the menu of a generalist predator, the eastern red bat (*Lasiurus borealis*): using a molecular approach to detect arthropod prey. *Mol. Ecol.* **18**, 2532–2542 (2009).
146. Hope, P. R. *et al.* Second generation sequencing and morphological faecal analysis reveal unexpected foraging behaviour by *Myotis nattereri* (Chiroptera, Vespertilionidae) in winter. *Front. Zool.* **11**, 39 (2014).
147. Dodd, L. E., Chapman, E. G., Harwood, J. D., Lacki, M. J. & Rieske, L. K. Identification of prey of *Myotis septentrionalis* using DNA-based techniques. *J. Mammal.* **93**, 1119–1128 (2012).
148. Gerhard, W. A. & Gunsch, C. K. Metabarcoding and machine learning analysis of environmental DNA in ballast water arriving to hub ports. *Environ. Int.* **124**, 312–319 (2019).
149. Roguet, A., Eren, A. M., Newton, R. J. & McLellan, S. L. Fecal source identification using random forest. *Microbiome* **6**, 185 (2018).

150. Dahl, C. *et al.* Gut microbiome of mothers delivering prematurely shows reduced diversity and lower relative abundance of Bifidobacterium and Streptococcus. *PLOS ONE* **12**, e0184336 (2017).
151. Bokulich, N. A. *et al.* Associations among Wine Grape Microbiome, Metabolome, and Fermentation Behavior Suggest Microbial Contribution to Regional Wine Characteristics. *mBio* **7**, (2016).
152. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
153. Adams, J. *et al.* Success of BrandenBark™, an artificial roost structure designed for use by Indiana bats (*Myotis sodalis*). *JASMR* **4**, (2015).
154. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71**, 8228–8235 (2005).
155. Chang, Q., Luan, Y. & Sun, F. Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics* **12**, 118 (2011).
156. Lozupone, C. A., Hamady, M., Kelley, S. T. & Knight, R. Quantitative and Qualitative  $\beta$  Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. *Appl. Environ. Microbiol.* **73**, 1576–1585 (2007).
157. Chen, J. *et al.* Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* **28**, 2106–2113 (2012).
158. McDonald, D. *et al.* Striped UniFrac: enabling microbiome analysis at unprecedented scale. *Nat. Methods* **15**, 847–848 (2018).
159. McKinney, W. & Others. Data structures for statistical computing in python. in *Proceedings of the 9th Python in Science Conference* **445**, 51–56 (Austin, TX, 2010).
160. McDonald, D. *et al.* The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Gigascience* **1**, 7 (2012).
161. Pedersen, T. L. & Robinson, D. *gganimate: A Grammar of Animated Graphics*. (2019).
162. Lumley, T., from Randall Munroe, U. D. & Participants, A. 2. 2e5 S. *xkcdcolors: Color Names from the XKCD Color Survey*. (2016).
163. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
164. Wickham, H. *stringr: Simple, Consistent Wrappers for Common String Operations*. (2019).
165. Chamberlain, S. A. & Szöcs, E. taxize: taxonomic search and retrieval in R. *F1000Res.* **2**, 191 (2013).
166. Chamberlain, S. *et al.* taxize: Taxonomic information from around the web. *R package version 0. 3. 0* (2014).
167. Ren, K. & Russell, K. *formattable: Create 'Formattable' Data Structures*. (2016).

168. Wickham, H. *rvest: Easily Harvest (Scrape) Web Pages*. (2019).
169. Agosta, S. J. & Morton, D. Diet of the big brown bat, *Eptesicus fuscus*, from Pennsylvania and western Maryland. *Northeastern Naturalist* **10**, 89–105 (2003).
170. Robert M. R. Barclay. Population Structure of Temperate Zone Insectivorous Bats in Relation to Foraging Behaviour and Energy Demand. *J. Anim. Ecol.* **60**, 165–178 (1991).
171. Fraser, E. E. & Fenton, M. B. Age and food hardness affect food handling by insectivorous bats. *Can. J. Zool.* **85**, 985–993 (2007).
172. Frenckell, B. von & Barclay, R. M. R. Bat activity over calm and turbulent water. *Can. J. Zool.* **65**, 219–222 (1987).
173. Kaupas, L. A. & Barclay, R. M. R. Temperature-dependent consumption of spiders by little brown bats (*Myotis lucifugus*), but not northern long-eared bats (*Myotis septentrionalis*), in northern Canada. *Can. J. Zool.* **96**, 261–268 (2018).
174. Agosta, S. J., Morton, D. & Kuhn, K. M. Feeding ecology of the bat *Eptesicus fuscus*: ‘preferred’ prey abundance as one factor influencing prey selection and diet breadth. *Journal of Zoology* **260**, 169–177 (2003).
175. Hamilton, I. M. & Barclay, R. M. R. Diets of Juvenile, Yearling, and Adult Big Brown Bats (*Eptesicus fuscus*) in Southeastern Alberta. *J. Mammal.* **79**, 764–771 (1998).
176. Moosman, P. R., Thomas, H. H. & Veilleux, J. P. Food habits of eastern small-footed bats (*Myotis leibii*) in New Hampshire. *Am. Midl. Nat.* **158**, 354–361 (2007).
177. Long, B. L., Kurta, A. & Clemans, D. L. Analysis of DNA from Feces to Identify Prey of Big Brown Bats (*Eptesicus fuscus*) Caught in Apple Orchards. *Am. Midl. Nat.* **170**, 287–297 (2013).
178. Boyles, J. G. & Storm, J. J. The Perils of Picky Eating: Dietary Breadth Is Related to Extinction Risk in Insectivorous Bats. *PLoS ONE* **2**, e672 (2007).
179. Blehert, D. S. *et al.* Bat white-nose syndrome: an emerging fungal pathogen? *Science* **323**, 227 (2009).
180. Hallmann, C. A. *et al.* More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLoS One* **12**, e0185809 (2017).
181. Lister, B. C. & Garcia, A. Climate-driven declines in arthropod abundance restructure a rainforest food web. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E10397–E10406 (2018).
182. Sánchez-Bayo, F. & Wyckhuys, K. A. G. Worldwide decline of the entomofauna: A review of its drivers. *Biol. Conserv.* **232**, 8–27 (2019).
183. Hagen, E. M. & Sabo, J. L. Influence of river drying and insect availability on bat activity along the San Pedro River, Arizona (USA). *J. Arid Environ.* **84**, 1–8 (2012).
184. Treitler, J. T., Heim, O., Tschapka, M. & Jung, K. The effect of local land use and loss of forests on bats and nocturnal insects. *Ecol. Evol.* **6**, 4289–4297 (2016).

185. Wickramasinghe, L. P., Harris, S., Jones, G. & Vaughan Jennings, N. Abundance and Species Richness of Nocturnal Insects on Organic and Conventional Farms: Effects of Agricultural Intensification on Bat Foraging. *Conserv. Biol.* **18**, 1283–1292 (2004).
186. Mineau, P. & Callaghan, C. Neonicotinoid Insecticides and Bats: An Assessment of the Direct and Indirect Risks. *Canadian Wildlife Federation, Ontario, Canada* 87 (2018).
187. Kurta, A., Bell, G. P., Nagy, K. A. & Kunz, T. H. Energetics of Pregnancy and Lactation in Freeranging Little Brown Bats (*Myotis lucifugus*). *Physiol. Zool.* **62**, 804–818 (1989).
188. Kurta, A., Kunz, T. H. & Nagy, K. A. Energetics and Water Flux of Free-Ranging Big Brown Bats (*Eptesicus fuscus*) during Pregnancy and Lactation. *J. Mammal.* **71**, 59–65 (1990).
189. Anthony, E. L. P. & Kunz, T. H. Feeding Strategies of the Little Brown Bat, *Myotis Lucifugus*, in Southern New Hampshire. *Ecology* **58**, 775–786 (1977).
190. Faith, D. P. Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* **61**, 1–10 (1992).
191. Team, R. & Others. RStudio: integrated development for R. *RStudio, Inc., Boston, MA URL* <http://www.rstudio.com> **42**, 14 (2015).
192. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
193. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
194. Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, (2018).
195. Hafen, R. *geofacet: 'ggplot2' Faceting Utilities for Geographical Data*. (2019).
196. Kahle, D. & Wickham, H. ggmap: Spatial Visualization with ggplot2. *The R Journal* **5**, 144–161 (2013).
197. RStudio & Inc. *htmltools: Tools for HTML*. (2017).
198. Golemund, G., Wickham, H. & Others. Dates and times made easy with lubridate. *J. Stat. Softw.* **40**, 1–25 (2011).
199. Chang, W. *webshot: Take Screenshots of Web Pages*. (2018).
200. Chandler, D. S. & Sikes, D. S. University of New Hampshire insect and arachnid collections. (2001). Available at: <http://furious-d.unh.edu/ento/home.php>. (Accessed: 1st July 2019)
201. Vesterinen, E. J., Lilley, T., Laine, V. N. & Wahlberg, N. Next generation sequencing of fecal DNA reveals the dietary diversity of the widespread insectivorous predator Daubenton's Bat (*Myotis daubentonii*) in Southwestern Finland. *PLoS One* **8**, e82168 (2013).
202. Aizpurua, O. *et al.* Agriculture shapes the trophic niche of a bat preying on multiple pest arthropods across Europe: Evidence from DNA metabarcoding. *Mol. Ecol.* **27**, 815–825 (2018).

203. Krüger, F., Clare, E. L., Symondson, W. O. C., Keiřs, O. & Pētersons, G. Diet of the insectivorous bat *Pipistrellus nathusii* during autumn migration and summer residence. *Mol. Ecol.* **23**, 3672–3683 (2014).
204. Kunz, T. H., Braun de Torrez, E., Bauer, D., Lobo, T. & Fleming, T. H. Ecosystem services provided by bats. *Ann. N. Y. Acad. Sci.* **1223**, 1–38 (2011).
205. Boyles, J. G., Cryan, P. M., McCracken, G. F. & Kunz, T. H. Conservation. Economic importance of bats in agriculture. *Science* **332**, 41–42 (2011).
206. Williams-Guillén, K., Perfecto, I. & Vandermeer, J. Bats limit insects in a neotropical agroforestry system. *Science* **320**, 70 (2008).
207. Brown, V. A., Braun de Torrez, E. & McCracken, G. F. Crop pests eaten by bats in organic pecan orchards. *Crop Prot.* **67**, 66–71 (2015).
208. Justice, D., Deely, A. K. & Rubin, F. Land Cover and Land Use Classification for the State of New Hampshire, 1996-2001. (2016). doi:10.3334/ORNLDAAAC/1305
209. Morin, R. S. & Lombard, K. *Forests of New Hampshire, 2016*. (U.S. Department of Agriculture, Forest Service, Northern Research Station, 2017). doi:10.2737/FS-RU-124



## APPENDIX

### SUPPLEMENTARY TABLES

<b>Site</b>	<b>Town</b>	<b>Samples</b>
ALS	Alstead, NH	120
BRN	Brown Lane, Hollis, NH	178
CHI	Chichester, NH	92
CNB	Canterbury, NH	162
EPS	Epsom, NH	85
FAR	Fairfield, ME	52
FOX	Fox State Forest, Hillsborough, NH	177
GIL	Gilsum, NH	44
HOL	Holderness, NH	165
HOP	Hopkinton, NH	182
MAP	Maple Hill, Hollis, NH	209
MAS	Massabeseic Lake, Auburn, NH	85
MTV	Mont Vernon, NH	91
PEN	Penacook, NH	47
ROL	Rollinsford, NH	13

**Table S1.** Guano samples collected per location. Guano samples were passively collected from 15 sites in New Hampshire (14) and Maine (1) in 2016. The following table includes the site abbreviation used in **Figure S1**, the town name where samples were collected, and the number of samples sequenced from each site.

	Basic					Standard					Extra				
Exact	dada2   deblur   vsearch					dada2   deblur   vsearch					dada2   deblur   vsearch				
	libA	22	23	19	19	libA	22	22	19	19	libA	22	22	19	19
	libB	22	22	21	21	libB	22	22	21	21	libB	22	22	21	21
	libC	22	22	18	18	libC	22	22	18	18	libC	22	22	18	18
	libD	23	23	19	19	libD	22	22	19	19	libD	22	22	19	19
		22	22	16	shared ASVs		22	22	16	shared ASVs		22	22	16	shared ASVs
Partial	dada2   deblur   vsearch					dada2   deblur   vsearch					dada2   deblur   vsearch				
	libA	3	4	28	2	libA	2	3	4	2	libA	2	4	4	2
	libB	2	2	35	2	libB	2	2	6	2	libB	2	2	3	2
	libC	2	2	59	2	libC	2	2	20	2	libC	2	2	8	2
	libD	2	138	173	2	libD	2	7	10	2	libD	2	4	6	2
		2	2	2	shared ASVs		2	2	2	shared ASVs		2	2	2	shared ASVs
Miss	dada2   deblur   vsearch					dada2   deblur   vsearch					dada2   deblur   vsearch				
	libA	5	60	67	1	libA	2	11	26	1	libA	0	0	0	
	libB	8	96	184	2	libB	6	16	62	2	libB	0	0	0	
	libC	7	56	90	5	libC	6	12	44	5	libC	0	0	0	
	libD	20	364	412	8	libD	12	23	92	7	libD	0	0	0	
		0	0	3	shared ASVs		0	0	3	shared ASVs					

**Table S2.** Summary of per-library or per-Pipeline overlaps in shared ASVs. Shared ASVs within a library (orange, left to right) reflect those sequences common across all filtering pipelines. Shared ASVs within a pipeline (blue, top to bottom) reflect sequences common across all mock community samples/libraries. Tables are grouped by vertical faceting reflecting filtering parameters: “Basic” represents default parameters for each filtering method; “Standard” requires a sample to have > 5000 reads, and an OTU to be present in > 1 sample; “Extra” includes “Standard” filters in addition to subtracting a fixed number of reads from all observations. Horizontal faceting describes how detected sequence variants in a mock sample aligned to expected mock community reference sequences: “Exact” are 100% identity matches; “Partial” are 97-99.9% identical; “Miss” are less than 97% identical to a reference. DADA2 and Deblur identify between 22 to 23 (of 24 possible) “Exact” sequences, while VSEARCH identifies fewer, whether or not additional filtering is applied. The remaining 1 or 2 expected ASV sequences are detected as “Partial” matches in every sequencing pipeline, regardless of filtering strategy. “Miss” ASVs are much more abundant in a VSEARCH AND Deblur than DADA2 for “basic” and “standard” filtering parameters. Among denoising “Miss” ASVs, no sequence occurs in multiple libraries, suggesting these ASVs are a sequencing artifact and not a product of wet-bench contamination.

	DADA2	Deblur	Vsearch
basic	12,793	22,733	16,064
standard	4,744	5,402	8,505
extra	4,353	3,913	2,740

**Table S3.** Number of variants observed per filtering parameter and program. Distinct sequence variants (ASVs for DADA2 and Deblur, OTUs for VSEARCH) are reduced with additional filtering parameters.

Comparison	Z	P.unadj	P.adj
deblur-basic - vsearch-extra	3.926	0	0.002
vsearch-basic - vsearch-extra	3.961	0	0.003
dada2-extra - deblur-basic	-2.979	0.003	0.026
dada2-extra - vsearch-basic	-3.014	0.003	0.031
deblur-basic - deblur-extra	2.576	0.01	0.06
deblur-extra - vsearch-basic	-2.611	0.009	0.065
dada2-basic - vsearch-extra	2.156	0.031	0.124
vsearch-basic - vsearch-standard	1.98	0.048	0.132
deblur-basic - vsearch-standard	1.945	0.052	0.133
deblur-basic - deblur-standard	2.173	0.03	0.134
dada2-standard - vsearch-extra	1.91	0.056	0.135
deblur-standard - vsearch-basic	-2.208	0.027	0.14
vsearch-extra - vsearch-standard	-1.98	0.048	0.143
dada2-standard - deblur-basic	-2.015	0.044	0.144
dada2-standard - vsearch-basic	-2.05	0.04	0.145
deblur-standard - vsearch-extra	1.753	0.08	0.159
dada2-basic - vsearch-basic	-1.805	0.071	0.16
dada2-basic - deblur-basic	-1.77	0.077	0.162
deblur-extra - vsearch-extra	1.349	0.177	0.336
dada2-basic - dada2-extra	1.209	0.227	0.408
dada2-extra - vsearch-standard	-1.034	0.301	0.516
dada2-extra - vsearch-extra	0.946	0.344	0.538
dada2-extra - dada2-standard	-0.964	0.335	0.548
dada2-basic - deblur-extra	0.806	0.42	0.605
dada2-extra - deblur-standard	-0.806	0.42	0.63
deblur-extra - vsearch-standard	-0.631	0.528	0.731
dada2-standard - deblur-extra	0.561	0.575	0.767
dada2-extra - deblur-extra	-0.403	0.687	0.824
dada2-basic - deblur-standard	0.403	0.687	0.853
deblur-extra - deblur-standard	-0.403	0.687	0.883
deblur-standard - vsearch-standard	-0.228	0.82	0.922
dada2-standard - deblur-standard	0.158	0.875	0.926
dada2-basic - dada2-standard	0.245	0.806	0.936
dada2-basic - vsearch-standard	0.175	0.861	0.939
dada2-standard - vsearch-standard	-0.07	0.944	0.971
deblur-basic - vsearch-basic	-0.035	0.972	0.972

**Table S4.** Dunn's test for Hill Number 0 estimates using mock data. Dunn's post hoc test for pairwise differences in diversity estimates of rarefied mock data among filtering method and filtering parameter levels for Hill Number  $q=0$ . Pairwise comparisons were considered significant if the Benjamini-Hochberg adjusted p value (P.adj) was  $< 0.05$ .

Comparison	Z	P.unadj	P.adj
deblur-basic - vsearch-extra	3.926	0	0.003
deblur-basic - vsearch-standard	3.691	0	0.004
deblur-basic - vsearch-basic	3.154	0.002	0.014
dada2-basic - vsearch-extra	3.222	0.001	0.015
dada2-standard - vsearch-extra	2.953	0.003	0.019
dada2-basic - vsearch-standard	2.987	0.003	0.02
dada2-extra - vsearch-extra	2.718	0.007	0.03
dada2-standard - vsearch-standard	2.718	0.007	0.034
dada2-basic - vsearch-basic	2.45	0.014	0.051
dada2-extra - vsearch-standard	2.483	0.013	0.052
dada2-standard - vsearch-basic	2.181	0.029	0.095
deblur-basic - deblur-extra	2.114	0.035	0.104
deblur-basic - deblur-standard	2.047	0.041	0.113
dada2-extra - vsearch-basic	1.946	0.052	0.133
deblur-standard - vsearch-extra	1.879	0.06	0.145
deblur-extra - vsearch-extra	1.812	0.07	0.157
deblur-standard - vsearch-standard	1.644	0.1	0.212
deblur-extra - vsearch-standard	1.577	0.115	0.229
dada2-basic - deblur-extra	1.409	0.159	0.301
dada2-basic - deblur-standard	1.342	0.179	0.323
dada2-extra - deblur-basic	-1.208	0.227	0.389
dada2-standard - deblur-extra	1.141	0.254	0.415
deblur-standard - vsearch-basic	1.107	0.268	0.42
dada2-standard - deblur-standard	1.074	0.283	0.424
deblur-extra - vsearch-basic	1.04	0.298	0.429
dada2-standard - deblur-basic	-0.973	0.33	0.458
dada2-extra - deblur-extra	0.906	0.365	0.487
dada2-extra - deblur-standard	0.839	0.402	0.516
vsearch-basic - vsearch-extra	0.772	0.44	0.546
dada2-basic - deblur-basic	-0.705	0.481	0.577
vsearch-basic - vsearch-standard	0.537	0.591	0.687
dada2-basic - dada2-extra	0.503	0.615	0.692
dada2-extra - dada2-standard	-0.235	0.814	0.838
dada2-basic - dada2-standard	0.268	0.788	0.86
vsearch-extra - vsearch-standard	-0.235	0.814	0.862
deblur-extra - deblur-standard	-0.067	0.946	0.946

**Table S5.** Pairwise Dunn's test for Hill Number 1 estimates using mock data. Dunn's post hoc test for pairwise differences in diversity estimates of rarefied mock data among filtering method and filtering parameter levels for Hill Number  $q=1$ . Pairwise comparisons were considered significant if the Benjamini-Hochberg adjusted p value (P.adj) was  $< 0.05$ .

Comparison	Z	P.unadj	P.adj
dada2-extra - vsearch-standard	2.852	0.004	0.026
dada2-basic - vsearch-standard	2.886	0.004	0.028
dada2-standard - vsearch-standard	2.953	0.003	0.028
dada2-extra - vsearch-basic	2.618	0.009	0.032
dada2-basic - vsearch-basic	2.651	0.008	0.032
dada2-standard - vsearch-basic	2.718	0.007	0.034
dada2-extra - vsearch-extra	2.987	0.003	0.034
deblur-basic - vsearch-extra	2.651	0.008	0.036
deblur-basic - vsearch-standard	2.517	0.012	0.039
dada2-basic - vsearch-extra	3.02	0.003	0.045
deblur-basic - vsearch-basic	2.282	0.022	0.067
dada2-standard - vsearch-extra	3.087	0.002	0.073
deblur-extra - vsearch-extra	1.879	0.06	0.155
deblur-standard - vsearch-extra	1.879	0.06	0.167
deblur-extra - vsearch-standard	1.745	0.081	0.182
deblur-standard - vsearch-standard	1.745	0.081	0.194
deblur-extra - vsearch-basic	1.51	0.131	0.262
deblur-standard - vsearch-basic	1.51	0.131	0.277
dada2-extra - deblur-extra	1.107	0.268	0.402
dada2-standard - deblur-extra	1.208	0.227	0.409
dada2-basic - deblur-extra	1.141	0.254	0.415
dada2-extra - deblur-standard	1.107	0.268	0.42
dada2-standard - deblur-standard	1.208	0.227	0.43
dada2-basic - deblur-standard	1.141	0.254	0.435
deblur-basic - deblur-extra	0.772	0.44	0.61
deblur-basic - deblur-standard	0.772	0.44	0.634
dada2-standard - deblur-basic	0.436	0.663	0.884
dada2-basic - deblur-basic	0.369	0.712	0.884
dada2-extra - deblur-basic	0.336	0.737	0.885
vsearch-basic - vsearch-extra	0.369	0.712	0.915
vsearch-basic - vsearch-standard	0.235	0.814	0.946
dada2-basic - dada2-extra	0.034	0.973	1
dada2-basic - dada2-standard	-0.067	0.946	1
dada2-extra - dada2-standard	-0.101	0.92	1
deblur-extra - deblur-standard	0	1	1
vsearch-extra - vsearch-standard	-0.134	0.893	1

**Table S6.** Dunn's post hoc test for pairwise differences in diversity estimates of rarefied mock data among filtering method and filtering parameter levels for Hill Number  $q=2$ . Pairwise comparisons were considered significant if the Benjamini-Hochberg adjusted p value (P.adj) was  $< 0.05$ .

Comparison	Z	P.unadj	P.adj
vsearch-basic - vsearch-extra	23.509	0	0
deblur-basic - vsearch-extra	22.688	0	0
dada2-basic - vsearch-extra	22.049	0	0
vsearch-extra - vsearch-standard	-21.384	0	0
deblur-extra - vsearch-basic	-17.697	0	0
deblur-standard - vsearch-extra	17.634	0	0
dada2-standard - vsearch-extra	17.459	0	0
deblur-basic - deblur-extra	17.053	0	0
dada2-basic - deblur-extra	16.281	0	0
deblur-extra - vsearch-standard	-15.633	0	0
dada2-extra - vsearch-basic	-12.722	0	0
dada2-extra - deblur-basic	-12.178	0	0
deblur-extra - deblur-standard	-12.156	0	0
dada2-standard - deblur-extra	11.842	0	0
dada2-basic - dada2-extra	11.274	0	0
dada2-extra - vsearch-extra	10.667	0	0
dada2-extra - vsearch-standard	-10.609	0	0
dada2-extra - deblur-standard	-7.185	0	0
dada2-extra - dada2-standard	-6.744	0	0
dada2-standard - vsearch-basic	-5.961	0	0
dada2-standard - deblur-basic	-5.585	0	0
dada2-extra - deblur-extra	5.248	0	0
deblur-standard - vsearch-basic	-5.201	0	0
deblur-extra - vsearch-extra	5.145	0	0
deblur-basic - deblur-standard	4.855	0	0
dada2-basic - dada2-standard	4.514	0	0
dada2-standard - vsearch-standard	-3.845	0	0
dada2-basic - deblur-standard	3.79	0	0
deblur-standard - vsearch-standard	-3.137	0.002	0.002
vsearch-basic - vsearch-standard	2.125	0.034	0.04
deblur-basic - vsearch-standard	1.844	0.065	0.076
dada2-basic - vsearch-basic	-1.45	0.147	0.165
dada2-basic - deblur-basic	-1.186	0.235	0.257
dada2-basic - vsearch-standard	0.674	0.5	0.53
dada2-standard - deblur-standard	-0.609	0.543	0.558
deblur-basic - vsearch-basic	-0.227	0.821	0.821

**Table S7.** Dunn's post hoc test for pairwise differences in diversity estimates of rarefied bat guano data among filtering method and filtering parameter levels for Hill Number  $q=0$ . Pairwise comparisons were considered significant if the Benjamini-Hochberg adjusted p value (P.adj) was  $< 0.05$ .

Comparison	Z	P.unadj	P.adj
dada2-basic - vsearch-extra	10.833	0	0
dada2-standard - vsearch-extra	9.322	0	0
dada2-basic - deblur-extra	8.619	0	0
dada2-basic - vsearch-standard	8.18	0	0
dada2-extra - vsearch-extra	7.75	0	0
dada2-basic - vsearch-basic	7.766	0	0
dada2-basic - deblur-standard	7.356	0	0
dada2-standard - deblur-extra	7.158	0	0
dada2-standard - vsearch-standard	6.674	0	0
dada2-standard - vsearch-basic	6.261	0	0
dada2-basic - deblur-basic	6.143	0	0
dada2-standard - deblur-standard	5.898	0	0
dada2-extra - deblur-extra	5.635	0	0
dada2-extra - vsearch-standard	5.104	0	0
dada2-standard - deblur-basic	4.683	0	0
dada2-extra - vsearch-basic	4.691	0	0
deblur-basic - vsearch-extra	4.441	0	0
dada2-extra - deblur-standard	4.375	0	0
deblur-standard - vsearch-extra	3.186	0.001	0.003
dada2-extra - deblur-basic	3.157	0.002	0.003
vsearch-basic - vsearch-extra	3.098	0.002	0.003
dada2-basic - dada2-extra	3.023	0.003	0.004
vsearch-extra - vsearch-standard	-2.681	0.007	0.011
deblur-basic - deblur-extra	2.453	0.014	0.021
deblur-extra - vsearch-extra	1.91	0.056	0.081
deblur-basic - vsearch-standard	1.832	0.067	0.093
dada2-extra - dada2-standard	-1.548	0.122	0.162
dada2-basic - dada2-standard	1.473	0.141	0.181
deblur-basic - vsearch-basic	1.426	0.154	0.191
deblur-extra - deblur-standard	-1.24	0.215	0.258
deblur-basic - deblur-standard	1.21	0.226	0.263
deblur-extra - vsearch-basic	-1.096	0.273	0.307
deblur-extra - vsearch-standard	-0.691	0.489	0.534
deblur-standard - vsearch-standard	0.585	0.559	0.592
vsearch-basic - vsearch-standard	0.418	0.676	0.696
deblur-standard - vsearch-basic	0.18	0.857	0.857

**Table S8.** Dunn's post hoc test for pairwise differences in diversity estimates of rarefied bat guano data among filtering method and filtering parameter levels for Hill Number  $q=1$ . Pairwise comparisons were considered significant if the Benjamini-Hochberg adjusted p value (P.adj) was < 0.05.

Comparison	Z	P.unadj	P.adj
dada2-basic - vsearch-extra	8.286	0	0
dada2-standard - vsearch-extra	7.175	0	0
dada2-basic - vsearch-standard	6.923	0	0
dada2-basic - vsearch-basic	6.666	0	0
dada2-basic - deblur-extra	6.56	0	0
dada2-extra - vsearch-extra	6.177	0	0
dada2-basic - deblur-standard	5.974	0	0
dada2-standard - vsearch-standard	5.813	0	0
dada2-standard - vsearch-basic	5.557	0	0
dada2-standard - deblur-extra	5.487	0	0
dada2-basic - deblur-basic	5.15	0	0
dada2-standard - deblur-standard	4.901	0	0
dada2-extra - vsearch-standard	4.815	0	0
dada2-extra - vsearch-basic	4.558	0	0
dada2-extra - deblur-extra	4.52	0	0
dada2-standard - deblur-basic	4.075	0	0
dada2-extra - deblur-standard	3.934	0	0
dada2-extra - deblur-basic	3.105	0.002	0.004
deblur-basic - vsearch-extra	2.948	0.003	0.006
deblur-standard - vsearch-extra	2.091	0.037	0.066
dada2-basic - dada2-extra	2.063	0.039	0.067
vsearch-basic - vsearch-extra	1.646	0.1	0.163
deblur-basic - vsearch-standard	1.6	0.109	0.171
deblur-extra - vsearch-extra	1.493	0.135	0.203
vsearch-extra - vsearch-standard	-1.387	0.166	0.229
deblur-basic - deblur-extra	1.408	0.159	0.229
deblur-basic - vsearch-basic	1.348	0.178	0.237
dada2-basic - dada2-standard	1.082	0.279	0.359
dada2-extra - dada2-standard	-0.98	0.327	0.406
deblur-basic - deblur-standard	0.827	0.408	0.49
deblur-standard - vsearch-standard	0.747	0.455	0.529
deblur-extra - deblur-standard	-0.58	0.562	0.633
deblur-standard - vsearch-basic	0.495	0.621	0.677
vsearch-basic - vsearch-standard	0.26	0.795	0.842
deblur-extra - vsearch-standard	0.149	0.881	0.906
deblur-extra - vsearch-basic	-0.102	0.918	0.918

**Table S9.** Dunn's post hoc test for pairwise differences in diversity estimates of rarefied bat guano data among filtering method and filtering parameter levels for Hill Number  $q=2$ . Pairwise comparisons were considered significant if the Benjamini-Hochberg adjusted p value (P.adj) was < 0.05.



TestGroup	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr..F.
Method	2	0.049	0.024	1.973	0.08	0.011
Filt	2	0.067	0.033	2.704	0.11	0.001
Method:Filt	4	0.162	0.041	3.276	0.26	0.001
Residuals	27	0.334	0.012	NA	0.55	NA
Total	35	0.612	NA	NA	1.00	NA

**Table S10.** Permutational Multivariate Analysis of Variance Using Distance Matrices (ADONIS) of mock community samples by filtering method (Method) and filtering parameter (Filt). Distances measured using Dice-Sorensen index. Samples rarefied to 5000 sequences.

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr..F.
Method	2	0.135	0.067	0.588	0.03	0.656
Filt	2	0.699	0.349	3.050	0.17	0.04
Method:Filt	4	0.229	0.057	0.501	0.06	0.848
Residuals	27	3.092	0.115	NA	0.74	NA
Total	35	4.154	NA	NA	1.00	NA

**Table S11.** Permutational Multivariate Analysis of Variance Using Distance Matrices (ADONIS) of mock community samples by filtering method (Method) and filtering parameter (Filt). Distances measured using Bray-Curtis index. Samples rarefied to 5000 sequences.

TestGroup	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr..F.
Method	2	0.015	0.007	1.133	0.06	0.387
Filt	2	0.039	0.020	3.054	0.16	0.036
Method:Filt	4	0.024	0.006	0.927	0.09	0.516
Residuals	27	0.173	0.006	NA	0.69	NA
Total	35	0.251	NA	NA	1.00	NA

**Table S12.** Permutational Multivariate Analysis of Variance Using Distance Matrices (ADONIS) of mock community samples by filtering method (Method) and filtering parameter (Filt). Distances measured using Morisita-Horn index. Samples rarefied to 5000 sequences.

TestGroup	Df	SumsOfSqs	MeanSqs	F.Model1	R2	Pr..F.
Method	2	0.391	0.196	0.543	0.00	1
Filt	2	1.140	0.570	1.582	0.00	0.006
MonthStart	3	43.392	14.464	40.159	0.14	0.001
Method:Filt	4	0.602	0.151	0.418	0.00	1
Residuals	710	255.721	0.360	NA	0.85	NA
Total	721	301.246	NA	NA	1.00	NA

**Table S13.** Permutational Multivariate Analysis of Variance Using Distance Matrices (ADONIS) of select guano samples collected from single location (Fox State Forest, Hillsborough NH) between April through October 2016 by denoising method (Method), filtering parameter (Filt), and date of sample collection (MonthStart). Distances measured using Dice-Sorensen index. Samples rarefied to 5000 sequences.

TestGroup	Df	SumsOfSqs	MeanSqs	F.Model1	R2	Pr..F.
Method	2	0.097	0.049	0.114	0.00	1
Filt	2	0.548	0.274	0.643	0.00	0.99
MonthStart	3	28.147	9.382	22.013	0.08	0.001
Method:Filt	4	0.116	0.029	0.068	0.00	1
Residuals	710	302.618	0.426	NA	0.91	NA
Total	721	331.526	NA	NA	1.00	NA

**Table S14.** Permutational Multivariate Analysis of Variance Using Distance Matrices (ADONIS) of select guano samples collected from single location (Fox State Forest, Hillsborough NH) between April through October 2016 by denoising method (Method), filtering parameter (Filt), and date of sample collection (MonthStart). Distances measured using Bray-Curtis index. Samples rarefied to 5000 sequences.

TestGroup	Df	SumsOfSqs	MeanSqs	F.Model1	R2	Pr..F.
Method	2	0.058	0.029	0.069	0.00	1
Filt	2	0.340	0.170	0.404	0.00	1
MonthStart	3	29.465	9.822	23.325	0.09	0.001
Method:Filt	4	0.090	0.023	0.054	0.00	1
Residuals	710	298.961	0.421	NA	0.91	NA
Total	721	328.915	NA	NA	1.00	NA

**Table S15.** Permutational Multivariate Analysis of Variance Using Distance Matrices (ADONIS) of select guano samples collected from single location (Fox State Forest, Hillsborough NH) between April through October 2016 by denoising method (Method), filtering parameter (Filt), and date of sample collection (MonthStart). Distances measured using Morisita-Horn index. Samples rarefied to 5000 sequences.

Database	Species	Genus	Family	Order	Class
tidybug	138,012	27,434	2,023	112	17
Palmer	76,493	20,148	1,492	96	16
Porter	82,874	21,937	1,773	111	17

**Table S16.** Number of unique arthropod taxa within each of the three COI databases. Reference records with shared sequences but distinct taxonomies were collapsed to least common ancestral taxa prior to dereplication and subsequent counts. While both Palmer and tidybug databases were derived from BOLD, timing of database design and filtering strategies lead to different number of unique records.

Database	Species	Genus	Family	Order	Class
Palmer	551,862	862,610	1,196,305	1,565,669	1,565,808
Porter	504,274	513,656	515,385	515,736	515,774
tidybug	946,309	1,311,216	1,836,852	1,840,258	1,841,305

**Table S17.** Number of unique arthropod sequences within each of the three COI databases. The Porter database design stipulates that only records with known Species names are retained, while tidybug database design required at least Family-rank information to be present. Thus the Porter database retains fewer records at higher taxonomic ranks.

Database	Species	Genus	Family	Order	Class
Palmer	1,013,969	703,221	369,526	162	23
Porter	11,506	2,124	395	44	6
tidybug	895,637	530,730	5,094	1,688	641

**Table S18.** Taxonomic missingness among COI databases. Number of records that are missing taxonomic information at a particular rank among retained arthropod sequences in each of the three COI databases. Because the Porter design required Species-rank information for inclusion in database construction, few records are lacking information at Species level, while many records in the Palmer and tidybug datasets do not include Species or Genus-rank information.

Classifier	Class	Order	Family	Genus	Species
blast	8,348	8,346	8,053	7,059	4,911
vsearch	8,201	8,199	7,986	7,046	4,930
nbayes	12,921	10,864	8,830	8,092	5,687
sintax	12,245	10,360	9,119	7,687	4,663
Bold	8,781	7,392	5,748	5,497	3,074

**Table S19.** Common taxonomic names assigned among classifiers. Number of bat guano ASVs assigned taxonomic information at Class through Species rank for each classifier. A total of 13,407 ASVs were included in the guano dataset for potential classification

<b><u>Arthropod Order</u></b>	<b><u>Samples detected</u></b>
Diptera	275
Lepidoptera	262
Araneae	253
Hemiptera	196
Coleoptera	168
Hymenoptera	119
Psocodea	104
Trichoptera	90
Neuroptera	56
Ephemeroptera	43
Trombidiformes	36
Blattodea	27
Orthoptera	22
Isopoda	16
Odonata	6
Mesostigmata	4
Poduromorpha	2
Entomobryomorpha	1
Mecoptera	1
Sarcoptiformes	1
Strepsiptera	1

**Table S20.** Breadth and detection frequency of arthropod Orders. More than half of all 285 bat guano samples contained at least one detection of a sequence variant (ASV) classified to Dipteran, Lepidopteran, Aranean, Hemipteran, or Coleopteran taxa. These and other frequently observed taxa such as Hymenoptera, Psocodea, and Trichoptera are among the expected arthropods previously described in morphological studies. It's unclear whether the infrequently observed taxa represent non-dietary components due to our passive sampling regime, gleaning behavior, or are instead rarely foraged. For example, a Springtail species (Entomobryomorpha) was identified in just a single sample and generated thousands of sequences, but these flightless arthropods are unlikely to be consumed by a bat and instead represent a non-target detection as a consequence of our passive collection regime. However, other taxa like Orthoptera and Blattodea may represent actively foraged, yet perhaps less seasonally available targets: 24 of 28 Blattodea samples were observed in June, while only 1 in 23 Orthopteran samples were detected in that same month. It's most likely that various mite ASVs (Sarcoptiformes, Mesostigmata, and Trombidiformes) detected are a product of either the bats directly gleaning parasitic mites or consuming an arthropod that is itself infected.

ASValias	Order	Family	Genus	Species	Samples	SeqCounts	Top 30 %	Top 50 %	Top 70 %
ASV-1	Araneae	Araneidae	Eustala		204	117864	TRUE	TRUE	TRUE
ASV-6	Diptera	Limoniidae	Epiphragma	Epiphragma solatrix	141	42300	TRUE	TRUE	FALSE
ASV-3	Diptera	Limoniidae	Rhipidia		164	56294	TRUE	TRUE	FALSE
ASV-13	Araneae	Araneidae	Eustala		88	22392	TRUE	FALSE	FALSE
ASV-64	Diptera	Chironomidae	Chironomus		88	4484	TRUE	FALSE	FALSE
ASV-10	Diptera	Culicidae	Aedes	Aedes vexans	136	23606	TRUE	FALSE	FALSE
ASV-20	Diptera	Culicidae	Culex	Culex erraticus	98	19355	TRUE	FALSE	FALSE
ASV-5	Diptera	Limoniidae	Erioptera	Erioptera caliptera	138	58751	TRUE	FALSE	FALSE
ASV-4	Diptera	Limoniidae	Erioptera	Erioptera parva	122	55561	TRUE	FALSE	FALSE
ASV-17	Diptera	Limoniidae			87	13096	TRUE	FALSE	FALSE
ASV-11	Diptera	Limoniidae			89	27673	TRUE	FALSE	FALSE
ASV-85	Diptera	Chironomidae	Chironomus		69	3219	FALSE	FALSE	FALSE
ASV-110	Diptera	Chironomidae	Glyptotendipes		56	2255	FALSE	FALSE	FALSE
ASV-34	Diptera	Chironomidae	Glyptotendipes		65	8803	FALSE	FALSE	FALSE
ASV-36	Diptera	Culicidae	Culex	Culex erraticus	60	8475	FALSE	FALSE	FALSE
ASV-18	Diptera	Culicidae	Culex	Culex erraticus	73	12992	FALSE	FALSE	FALSE
ASV-80	Diptera	Culicidae	Uranotaenia	Uranotaenia sapphirina	62	4438	FALSE	FALSE	FALSE
ASV-23	Diptera	Limoniidae	Helius	Helius flavipes	58	13877	FALSE	FALSE	FALSE
ASV-15	Diptera	Limoniidae	Metalimnobia	Metalimnobia triocellata	66	22461	FALSE	FALSE	FALSE
ASV-21	Diptera	Limoniidae	Rhipidia		58	11069	FALSE	FALSE	FALSE
ASV-16	Diptera	Tipulidae	Nephrotoma	Nephrotoma ferruginea	69	14503	FALSE	FALSE	FALSE
ASV-9	Lepidoptera	Tineidae	Acrolophus	Acrolophus mortipennella	71	23219	FALSE	FALSE	FALSE
ASV-58	Lepidoptera	Tortricidae	Choristoneura		62	4915	FALSE	FALSE	FALSE
ASV-14	Psocodea	Psocidae	Metylophorus	Metylophorus novaescotiae	67	17802	FALSE	FALSE	FALSE
ASV-19	Trichoptera	Hydropsychidae	Potamyia	Potamyia flava	60	20588	FALSE	FALSE	FALSE

**Table S21.** Core sequence variant detection and sequence counts. Only a small fraction of all ASVs detected (2,575 ASVs in total) are repeatedly observed. We detected just 25 ASVs in at least 20% of samples (shown below), and 63 ASVs in at least 10% of samples. The majority of these frequently detected ASVs are classified as Dipteran; only Dipteran or Aranean taxa are identified in 30% or 50% of samples. The total sequencing depth (SeqCounts) represents the sum of the rarefied counts per ASV, the highest of which was classified as an unknown *Eustala* spider (ASV-1). The unambiguous classification information is provided for each sequence variant observed, though not all Species or Genus names were assigned to these highly detected ASVs. Multiple ASVs were frequently assigned to repeatedly detected ASVs—often with Species names—indicating population genetic variation likely exists for this marker gene in many taxa.

Site	Month	Hill_qType	meanQ	sdQ
EN	June	q=0	39.63	11.95
EN	July	q=0	42.17	22.78
EN	September	q=0	30.35	13.03
HB	June	q=0	42.31	28.06
HB	July	q=0	39.35	24.20
HB	September	q=0	29.55	22.33
EN	June	q=1	9.08	5.84
EN	July	q=1	12.29	9.92
EN	September	q=1	7.55	4.53
HB	June	q=1	11.47	10.45
HB	July	q=1	15.22	11.60
HB	September	q=1	9.06	7.37
EN	June	q=2	4.97	3.03
EN	July	q=2	6.87	5.76
EN	September	q=2	4.40	2.56
HB	June	q=2	6.54	5.65
HB	July	q=2	9.28	7.19
HB	September	q=2	5.66	4.30

**Table S22.** Summary of alpha diversity estimates for collection Site and Month groups. Three Hill Numbers were used to estimate species diversity (i.e. unique ASVs) per sample. The mean (meanQ) and standard deviation (sdQ) reported for each Site (Egner, “EN”; Hickory Bottoms, “HB”) and Month reflect classic alpha diversity measures that are normalized to a common scale: q=0 represents observed richness, q=1 Shannon’s entropy, and q=2 Simpson’s 1-D index. Estimates that increasingly weight relative abundance information result lower species diversity, indicating that rare variants are dominating the observed richness per sample.

Comparison	Z	P.unadj	P.adj
July-EN - September-HB	3.524	0.000	0.003
June-EN - September-HB	3.634	0.000	0.004
June-HB - September-HB	3.004	0.003	0.013
June-EN - September-EN	2.727	0.006	0.019
July-HB - September-HB	2.807	0.005	0.019
July-EN - September-EN	2.586	0.010	0.024
June-HB - September-EN	2.020	0.043	0.093
July-HB - September-EN	1.843	0.065	0.123
July-HB - June-EN	-0.980	0.327	0.491
June-EN - June-HB	0.883	0.377	0.514
September-EN - September-HB	0.993	0.321	0.534
July-EN - July-HB	0.773	0.439	0.549
July-EN - June-HB	0.668	0.504	0.582
July-EN - June-EN	-0.232	0.817	0.875
July-HB - June-HB	-0.128	0.898	0.898

**Table S23.** Dunn's Test for observed richness (Hill  $q=0$ ). Pairwise comparisons for each Site+Month group (Comparison) includes a Benjamini Hochberg-adjusted p-value (P.adj). Most comparisons for Hill Number diversity equivalent to observed richness ( $q=0$ ) for June or July to September between and within Sites are significant (P.adj < 0.05) while comparisons between June to July, or for September or not. These data suggest that Month, but not Site, contributes to the variability in observed richness among our samples.

Comparison	Z	P.unadj	P.adj
July-HB - September-EN	2.953	0.003	0.047
July-HB - September-HB	2.454	0.014	0.106
July-HB - June-HB	1.758	0.079	0.236
July-HB - June-EN	1.963	0.05	0.248
July-EN - September-EN	1.779	0.075	0.282
June-HB - September-EN	1.305	0.192	0.411
July-EN - July-HB	-1.162	0.245	0.46
July-EN - September-HB	1.313	0.189	0.473
June-HB - September-HB	0.833	0.405	0.552
July-EN - June-EN	0.834	0.404	0.606
June-EN - September-EN	0.888	0.375	0.625
September-EN - September-HB	-0.422	0.673	0.721
July-EN - June-HB	0.544	0.586	0.733
June-EN - June-HB	-0.34	0.734	0.734
June-EN - September-HB	0.458	0.647	0.747

**Table S24.** Dunn's Test for Shannon entropy (Hill  $q=1$ ). Pairwise comparisons for each Site+Month group (Comparison) includes a Benjamini Hochberg-adjusted p-value (P.adj). Most comparisons for Hill Number diversity equivalent to Shannon's entropy ( $q=1$ ) are not significant. These data indicate that when abundance information is incorporated in a diversity analysis, we observe little difference in these estimates of the 'commonly abundant' taxa.

<b>Comparison</b>	<b>Z</b>	<b>P.unadj</b>	<b>P.adj</b>
July-HB - September-EN	3.102	0.002	0.029
July-HB - June-EN	2.386	0.017	0.128
July-HB - September-HB	2.225	0.026	0.131
July-HB - June-HB	1.901	0.057	0.215
July-EN - July-HB	-1.678	0.093	0.280
July-EN - September-EN	1.419	0.156	0.390
June-HB - September-EN	1.318	0.187	0.402
July-EN - September-HB	0.590	0.555	0.694
June-EN - September-EN	0.613	0.540	0.736
June-HB - September-HB	0.461	0.645	0.744
July-EN - June-EN	0.760	0.447	0.745
June-EN - June-HB	-0.638	0.523	0.785
September-EN - September-HB	-0.792	0.428	0.803
July-EN - June-HB	0.155	0.876	0.876
June-EN - September-HB	-0.170	0.865	0.927

**Table S25.** Dunn's Test for Simpson's 1-D index (Hill  $q=2$ ). Pairwise comparisons for each Site+Month group (Comparison) includes a Benjamini Hochberg-adjusted p-value (P.adj). Most comparisons for Hill Number diversity equivalent to Simpson's 1-D index ( $q=2$ ) are not significant. As with Table S5, these data indicate that when abundance information is incorporated in a diversity analysis, we observe little difference in these estimates of the 'commonly abundant' taxa.



	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr..F.
Month	2	9.067236	4.5336181	12.800823	0.08233079	0.001
Site	1	2.481712	2.4817123	7.007198	0.02253403	0.001
Month:Site	2	2.603807	1.3019036	3.675969	0.02364265	0.001
Residuals	271	95.979024	0.3541661	NA	0.87149254	NA
Total	276	110.131780	NA	NA	1.00000000	NA

**Table S26.** ADONIS for Dice-Sorensen distance metric.

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr..F.
Month	2	6.626122	3.3130611	8.231117	0.05446043	0.001
Site	1	2.192721	2.1927207	5.447693	0.01802208	0.001
Month:Site	2	3.771026	1.8855130	4.684453	0.03099425	0.001
Residuals	271	109.078701	0.4025044	NA	0.89652324	NA
Total	276	121.668570	NA	NA	1.00000000	NA

**Table S27.** ADONIS for Bray-Curtis distance metric.

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr..F.
Month	2	6.831555	3.415778	8.633658	0.05642219	0.001
Site	1	2.373880	2.373880	6.000177	0.01960601	0.001
Month:Site	2	4.656704	2.328352	5.885101	0.03845998	0.001
Residuals	271	107.217095	0.395635	NA	0.88551183	NA
Total	276	121.079234	NA	NA	1.00000000	NA

**Table S28.** ADONIS for Morisita-Horn distance metric.

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr..F.
Month	2	6.567287	3.2836436	12.978074	0.08389806	0.001
Site	1	1.490458	1.4904579	5.890795	0.01904082	0.001
Month:Site	2	1.652255	0.8261277	3.265137	0.02110781	0.001
Residuals	271	68.566987	0.2530147	NA	0.87595332	NA
Total	276	78.276988	NA	NA	1.00000000	NA

**Table S29.** ADONIS for unweighted Unifrac distance metric.

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr..F.
Month	2	0.9022023	0.45110115	10.204535	0.06570457	0.001
Site	1	0.2611661	0.26116611	5.907940	0.01901991	0.001
Month:Site	2	0.5880148	0.29400739	6.650856	0.04282328	0.001
Residuals	271	11.9798121	0.04420595	NA	0.87245224	NA
Total	276	13.7311953	NA	NA	1.00000000	NA

**Table S30.** ADONIS for weighted Unifrac distance metric.

Site	Year	SamplesCollected	SamplesAnalyzed	SamplingWindowsSurveyed	SamplingWindowsAnalyzed
ALS	2016	99	47	9	6
BRN	2015	33	11	8	4
BRN	2016	153	49	12	8
CHI	2016	170	43	12	6
CNA	2016	96	36	8	6
CNB	2016	158	38	12	8
COR	2015	45	20	7	4
COR	2016	65	25	5	4
EPS	2016	157	41	15	6
FOX	2015	38	18	5	5
FOX	2016	254	88	21	15
GIL	2015	14	5	2	2
GIL	2016	44	16	3	3
GRN	2015	14	0	3	0
HOL	2016	212	81	16	9
HOP	2015	58	28	11	8
HOP	2016	185	56	16	7
MAP	2015	29	16	7	6
MAP	2016	191	73	14	11
MAS	2015	21	11	5	3
MAS	2016	51	5	9	3
MTV	2016	158	69	12	9
PEN	2016	82	39	7	5
ROL	2016	9	0	2	0
SWZ	2015	65	43	9	6
WLD	2015	54	27	6	6
WLT	2015	34	23	6	6

**Table 31.** Summary of data collected and analyzed for each site and year. Individual guano pellets were collected (largely by volunteers) at 20 different sites: 11 of these sites were sampled in 2015, while 16 sites were collected in 2016. Among these sites, 7 were sampled across both years (BRN, COR, FOX, GIL, HOP, MAP, MAS). Passive collection methods generated thousands of samples (SamplesCollected) but only about half of these typically produced sufficient sequence throughput (SamplesAnalyzed). After filtering for minimum read abundance per sample and removing sampling windows (two-week spans) with only a single sample per Site, there were typically fewer surveyed sampling periods (SamplingWindowsSurveyed) per site that what was analyzed (SamplingWindowsAnalyzed).

Order	Family	Genus	ASVs	Sites	CollectionWindows	Samples
Coleoptera	Scarabaeidae	Phyllophaga	838	18	23	469
Coleoptera	Scarabaeidae	Maladera	3	18	21	325
Coleoptera	Elateridae	Hemicrepidius	42	18	23	267
Diptera	Tipulidae	Tipula	119	18	22	251
Coleoptera	Pyrochroidae	Dendroides	99	18	23	241
Coleoptera	Elateridae	Melanotus	101	18	22	194
Lepidoptera	Tortricidae	NA	13	18	18	110
Coleoptera	Coccinellidae	Harmonia	24	18	21	83
Coleoptera	Scarabaeidae	Diploptaxis	34	17	19	188
Lepidoptera	Gracillariidae	Caloptilia	17	17	18	100
Trichoptera	Hydropsychidae	Hydropsyche	41	16	17	141
Diptera	Limoniidae	Pseudolimnophila	42	16	16	137
Megaloptera	Corydalidae	Chauliodes	76	16	20	104
Diptera	Limoniidae	Erioptera	42	16	20	91
Lepidoptera	Tortricidae	Olethreutes	10	16	13	67
Diptera	Bombyliidae	Lepidophora	17	16	14	57
Diptera	Chironomidae	NA	101	15	21	146
Diptera	Chironomidae	Chironomus	32	15	19	128
Diptera	Chironomidae	Procladius	41	15	19	112
Coleoptera	Scarabaeidae	NA	11	15	19	104
Trichoptera	Limnephilidae	Platycentropus	16	15	15	71
Lepidoptera	Blastobasidae	Asaphocrita	17	15	14	60
Coleoptera	Chrysomelidae	Paria	12	15	13	41

**Table S32.** Summary of most frequently detected arthropod genera. Multiple genera of beetles and flies are repeatedly detected at most sites for most time periods, though some particular moth, caddisfly, and a fishfly genera are also widely detected throughout the New Hampshire locations sampled. Samples were grouped in 14-day increments (CollectionWindows); a taxon detected in the same Collection Window across both years was counted individually. Not only are bats capable of foraging a diverse assortment of multiple orders, but they appear to be continually consuming a broad collection of arthropods throughout the foraging season across a range of New Hampshire habitats.

**Dice-Sorensen**

<b>Terms</b>	<b>Df</b>	<b>SumsOfSqs</b>	<b>MeanSqs</b>	<b>F.Model</b>	<b>R2</b>	<b>Pr (&gt;F)</b>
Site	8	16.758	2.095	5.039	0.0691	0.001
Window	1	3.1244	3.124	7.516	0.0129	0.001
Site:Window	8	10.137	1.267	3.048	0.0418	0.001
Residuals	511	212.419	0.416		0.876	
Total	528	242.438			1	

**Bray-Curtis**

<b>Terms</b>	<b>Df</b>	<b>SumsOfSqs</b>	<b>MeanSqs</b>	<b>F.Model</b>	<b>R2</b>	<b>Pr (&gt;F)</b>
Site	8	14.500	1.812	4.227	0.059	0.001
Window	1	2.195	2.195	5.120	0.009	0.001
Site:Window	8	10.981	1.373	3.201	0.044	0.001
Residuals	511	219.097	0.429		0.888	
Total	528	246.773			1	

**Unweighted Unifrac**

<b>Terms</b>	<b>Df</b>	<b>SumsOfSqs</b>	<b>MeanSqs</b>	<b>F.Model</b>	<b>R2</b>	<b>Pr (&gt;F)</b>
Site	8	13.869	1.734	5.929	0.080	0.001
Window	1	2.139	2.139	7.314	0.012	0.001
Site:Window	8	8.039	1.005	3.436	0.046	0.001
Residuals	511	149.427	0.292		0.861	
Total	528	173.473			1	

**Weighted Unifrac**

<b>Terms</b>	<b>Df</b>	<b>SumsOfSqs</b>	<b>MeanSqs</b>	<b>F.Model</b>	<b>R2</b>	<b>Pr (&gt;F)</b>
Site	8	4.87	0.609	6.671	0.088	0.001
Window	1	0.446	0.446	4.885	0.008	0.002
Site:Window	8	2.996	0.374	4.101	0.055	0.001
Residuals	511	46.658	0.091		0.849	
Total	528	54.973			1	

**Table S33.** PERMANOVA summaries for site and location main effects. Testing for main effects for site and sampling location (Window) and their interaction was conducted using the Adonis function in Vegan on distance estimates from four metrics: Dice-Sorensen (unweighted abundance, unweighted phylogenetic), Bray-Curtis (weighted abundance, unweighted phylogenetic), unweighted Unifrac (unweighted abundance, weighted phylogenetic), and weighted Unifrac (weighted abundance, weighted phylogenetic). Main effects for site and window, as well as their interaction, were significant for all distance estimates tested. A greater proportion of variation is explained by these main effects when incorporating abundance and/or phylogenetic information.

diff	lwr	upr	p.adj	Pairs
14.67	6.75	22.598	0	MAP-FOX
-15.29	-23.69	-6.881	0	MTV-MAP
14.55	6.472	22.628	0	MAP-HOL
14.75	5.859	23.642	0	MAP-HOP
11	0.975	21.03	0.019	PEN-MTV
10.39	0.762	20.019	0.023	PEN-FOX
10.27	0.511	20.023	0.031	PEN-HOL
-10.52	-20.628	-0.405	0.034	MTV-CNB
-9.9	-19.621	-0.189	0.042	FOX-CNB
10.47	0.028	20.906	0.049	PEN-HOP
-9.34	-18.692	0.01	0.05	MTV-BRN
-9.78	-19.623	0.061	0.053	HOL-CNB
-8.73	-17.652	0.192	0.061	FOX-BRN
-8.61	-17.665	0.453	0.078	HOL-BRN
-9.98	-20.501	0.539	0.078	HOP-CNB
-8.81	-18.598	0.985	0.117	HOP-BRN
-8.27	-18.56	2.022	0.232	MTV-CNA
-7.66	-17.561	2.245	0.281	FOX-CNA
-7.53	-17.56	2.492	0.319	HOL-CNA
-7.73	-18.427	2.959	0.372	HOP-CNA
7.02	-3.178	17.21	0.444	MAP-CNA
5.94	-3.3	15.188	0.541	MAP-BRN
4.77	-5.243	14.782	0.862	MAP-CNB
-4.28	-14.211	5.644	0.917	PEN-MAP
2.73	-8.836	14.302	0.998	PEN-CNA
2.25	-9.395	13.889	1	CNB-CNA
1.66	-9.08	12.402	1	PEN-BRN
1.18	-9.645	11.995	1	CNB-BRN
-1.07	-12.06	9.916	1	CNA-BRN
-0.74	-8.935	7.465	1	MTV-HOL
-0.61	-8.66	7.437	1	MTV-FOX
-0.54	-9.538	8.468	1	MTV-HOP
0.49	-10.924	11.895	1	PEN-CNB
-0.2	-8.899	8.499	1	HOP-HOL
0.12	-7.583	7.831	1	HOL-FOX
-0.08	-8.633	8.48	1	HOP-FOX

**Table S34** – Tukey's pairwise comparison of observed richness by site.

<b>diff</b>	<b>lwr</b>	<b>upr</b>	<b>p.adj Pairs</b>
0.89	0.407	1.378	0 MAP-FOX
-0.78	-1.293	-0.263	0 MTV-MAP
0.73	0.186	1.275	0.001 MAP-HOP
-0.73	-1.274	-0.181	0.001 FOX-BRN
-0.74	-1.331	-0.141	0.004 FOX-CNB
0.71	0.125	1.304	0.006 PEN-FOX
-0.61	-1.186	-0.04	0.026 MTV-BRN
-0.62	-1.227	-0.014	0.041 FOX-CNA
-0.62	-1.241	-0.002	0.048 MTV-CNB
0.6	-0.014	1.214	0.061 PEN-MTV
0.46	-0.011	0.933	0.062 HOL-FOX
-0.57	-1.165	0.034	0.083 HOP-BRN
-0.57	-1.218	0.07	0.125 HOP-CNB
0.43	-0.063	0.926	0.144 MAP-HOL
0.55	-0.087	1.192	0.153 PEN-HOP
-0.51	-1.136	0.124	0.234 MTV-CNA
-0.46	-1.113	0.197	0.42 HOP-CNA
-0.35	-0.849	0.156	0.441 MTV-HOL
-0.3	-0.832	0.234	0.716 HOP-HOL
-0.27	-0.821	0.288	0.857 HOL-BRN
-0.28	-0.878	0.328	0.889 HOL-CNB
0.27	-0.352	0.897	0.912 MAP-CNA
0.25	-0.344	0.851	0.924 PEN-HOL
0.16	-0.362	0.686	0.989 HOP-FOX
-0.18	-0.786	0.43	0.992 PEN-MAP
0.17	-0.401	0.731	0.992 MAP-BRN
-0.16	-0.773	0.455	0.997 HOL-CNA
0.16	-0.457	0.77	0.997 MAP-CNB
0.11	-0.379	0.607	0.998 MTV-FOX
0.12	-0.597	0.829	1 CNB-CNA
-0.11	-0.78	0.566	1 CNA-BRN
0.09	-0.614	0.803	1 PEN-CNA
-0.05	-0.599	0.504	1 MTV-HOP
-0.02	-0.72	0.677	1 PEN-CNB
-0.01	-0.671	0.645	1 PEN-BRN
0.01	-0.654	0.671	1 CNB-BRN

**Table S35** – Tukey's pairwise comparison of Shannon's entropy (diversity) by site.

diff	lwr	upr	p.adj	Pairs
-1.01	-1.449	-0.572	0	MAP-CNB
1.06	0.589	1.537	0	CNB-BRN
-0.74	-1.093	-0.385	0	MAP-HOL
1.04	0.534	1.554	0	CNB-CNA
0.79	0.395	1.188	0	HOL-BRN
0.77	0.333	1.211	0	HOL-CNA
0.72	0.285	1.155	0	PEN-MAP
0.77	0.302	1.243	0	PEN-BRN
0.6	0.228	0.964	0	MTV-MAP
0.65	0.239	1.058	0	MTV-BRN
0.75	0.247	1.26	0	PEN-CNA
-0.68	-1.145	-0.223	0	HOP-CNB
-0.63	-1.055	-0.204	0	FOX-CNB
0.63	0.178	1.08	0.001	MTV-CNA
0.43	0.043	0.824	0.017	FOX-BRN
-0.38	-0.728	-0.034	0.019	MAP-FOX
-0.41	-0.793	-0.031	0.023	HOP-HOL
0.36	0.02	0.695	0.028	HOL-FOX
0.41	-0.019	0.848	0.074	FOX-CNA
-0.42	-0.858	0.028	0.086	MTV-CNB
0.38	-0.05	0.808	0.132	HOP-BRN
0.39	-0.064	0.851	0.157	PEN-HOP
-0.33	-0.716	0.063	0.183	MAP-HOP
0.34	-0.083	0.761	0.232	PEN-FOX
0.36	-0.108	0.828	0.288	HOP-CNA
0.27	-0.125	0.663	0.456	MTV-HOP
-0.27	-0.703	0.159	0.569	HOL-CNB
0.21	-0.138	0.567	0.616	MTV-FOX
-0.29	-0.79	0.209	0.673	PEN-CNB
-0.14	-0.502	0.216	0.946	MTV-HOL
0.12	-0.315	0.563	0.994	PEN-MTV
-0.05	-0.429	0.32	1	HOP-FOX
0.05	-0.352	0.457	1	MAP-BRN
0.03	-0.413	0.48	1	MAP-CNA
-0.02	-0.446	0.408	1	PEN-HOL
0.02	-0.462	0.5	1	CNA-BRN

**Table S36** – Tukey's pairwise comparison of Faith's phylogenetic diversity by site.

**Dice Sorensen: COR, HOL, and PEN sites in 2016**

<b>Terms</b>	<b>Df</b>	<b>SumsOfSqs</b>	<b>MeanSqs</b>	<b>F.Model</b>	<b>R2</b>	<b>Pr (&gt;F)</b>
Site	2	7.016	3.508	9.107	0.109	0.001
Window	1	2.015	2.015	5.231	0.031	0.001
Site:Window	2	1.912	0.956	2.482	0.030	0.001
Residuals	139	53.540	0.385		0.830	
Total	144	64.483			1.000	

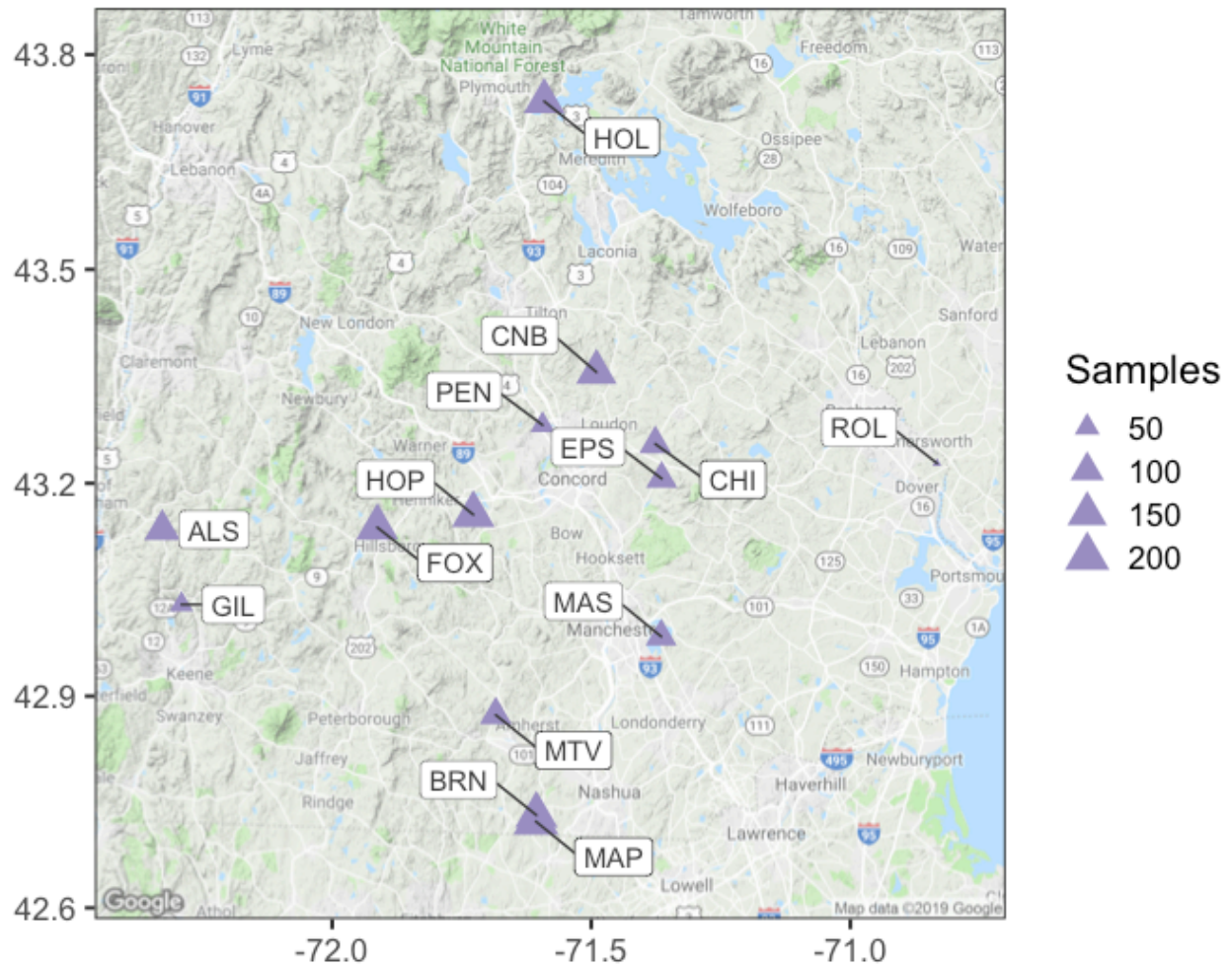
**Dice Sorensen: MAP and BRN sites in 2016**

<b>Terms</b>	<b>Df</b>	<b>SumsOfSqs</b>	<b>MeanSqs</b>	<b>F.Model</b>	<b>R2</b>	<b>Pr (&gt;F)</b>
Site	1	0.479	0.479	1.231	0.010	0.172
Window	1	2.871	2.871	7.383	0.057	0.001
Site:Window	1	0.725	0.725	1.863	0.015	0.020
Residuals	118	45.895	0.389		0.918	
Total	121	49.970			1.000	

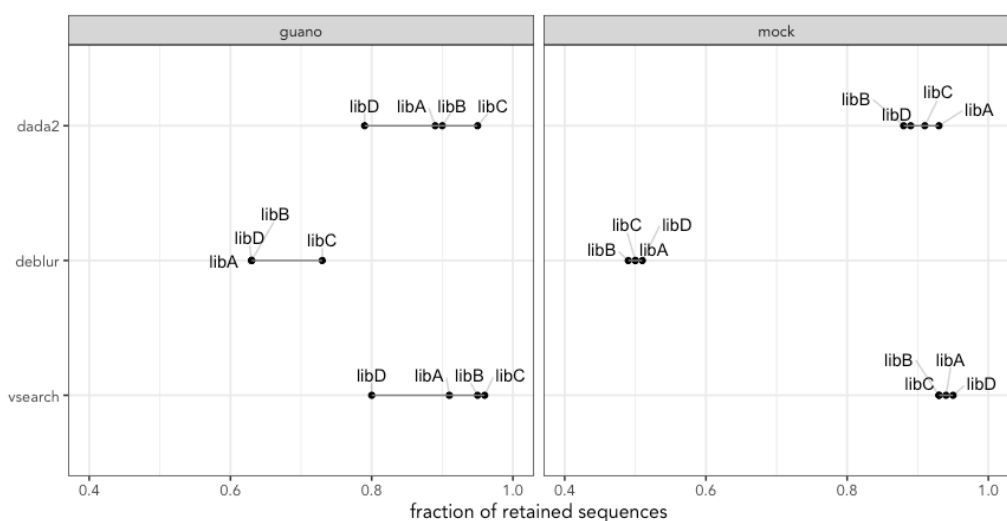
**Table S37** - PERMANOVA testing for selected sites. Main effects for site and sampling location (Window) and their interaction were calculated using the Adonis function in Vegan with a Dice-Sorensen distance estimate. Selected sites consist of 2016 samples collected among shared landscape features. (A) Cornish (COR), Holderness (HOL), and Penacook (PEN) are roosts with close proximity to large water bodies (Connecticut River, Squam Lake, and Merrimack River, respectively).



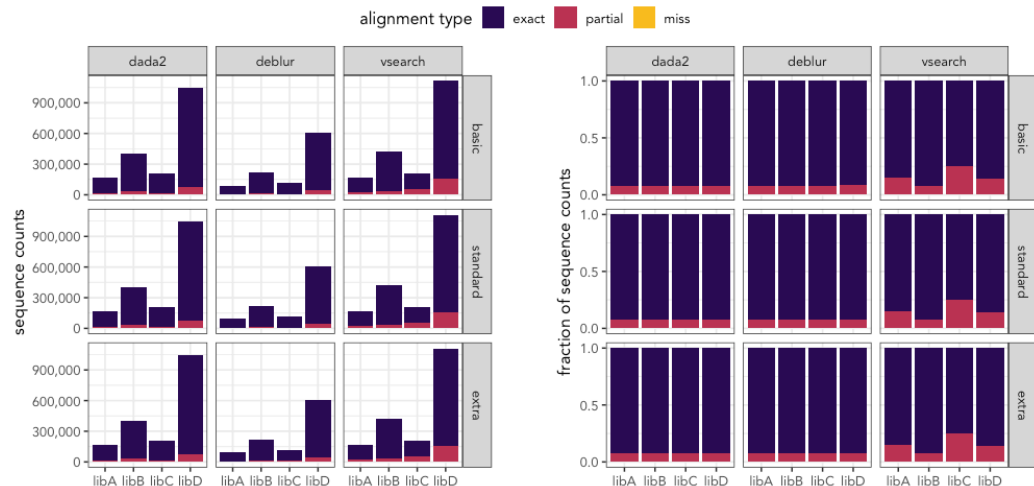
## SUPPLEMENTARY FIGURES



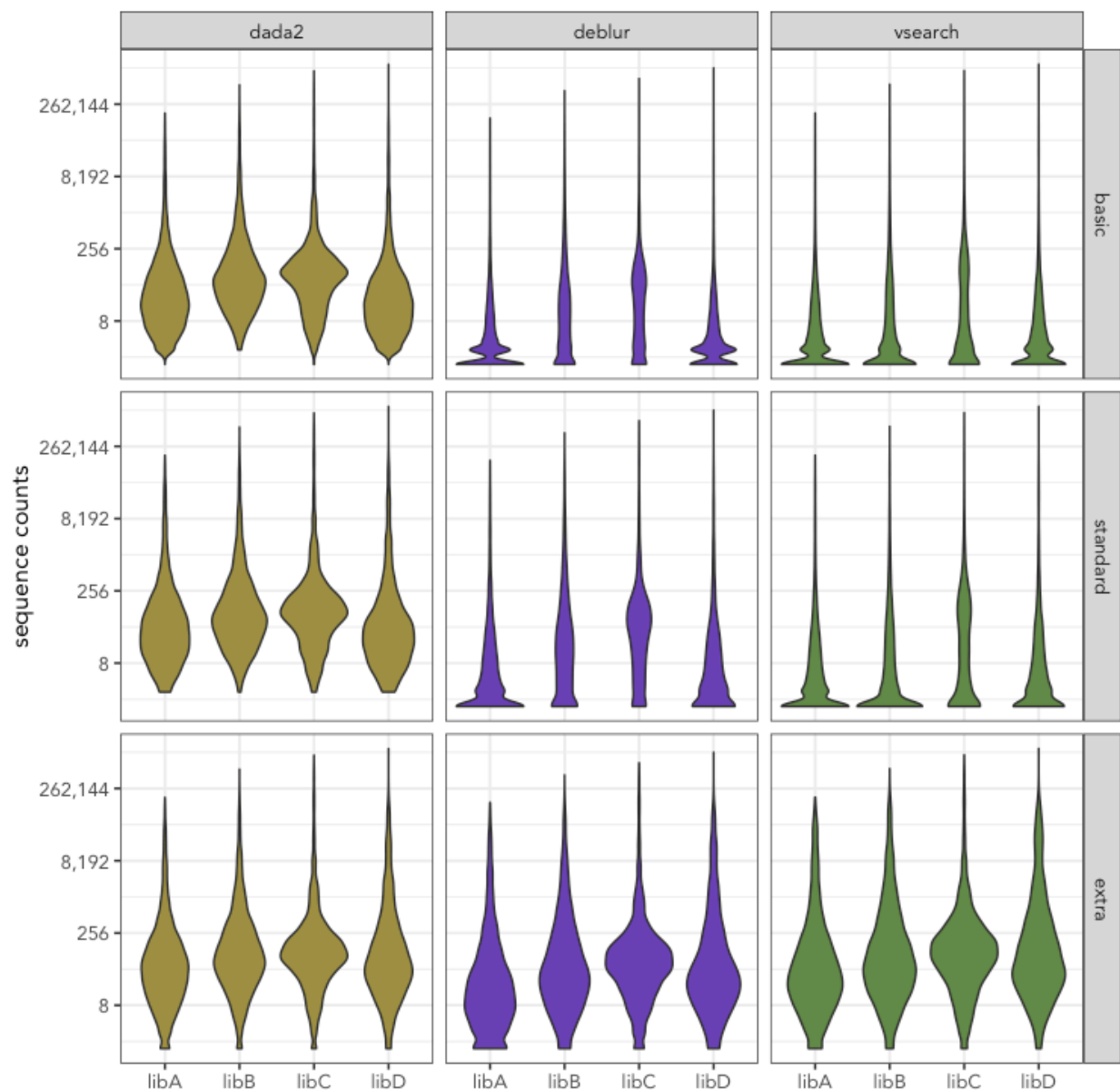
**Figure S1.** Locations for New England guano sampling. Guano samples included in this project constitute a subset of a broader project that included collections from 21 locations across New Hampshire and Maine in 2015 and 2016. The samples used in this study were from sites collected in 2016 in towns listed in **Table S1**, and shown in the map of New Hampshire below. Samples collected from Fairfield ME are not shown in this map.



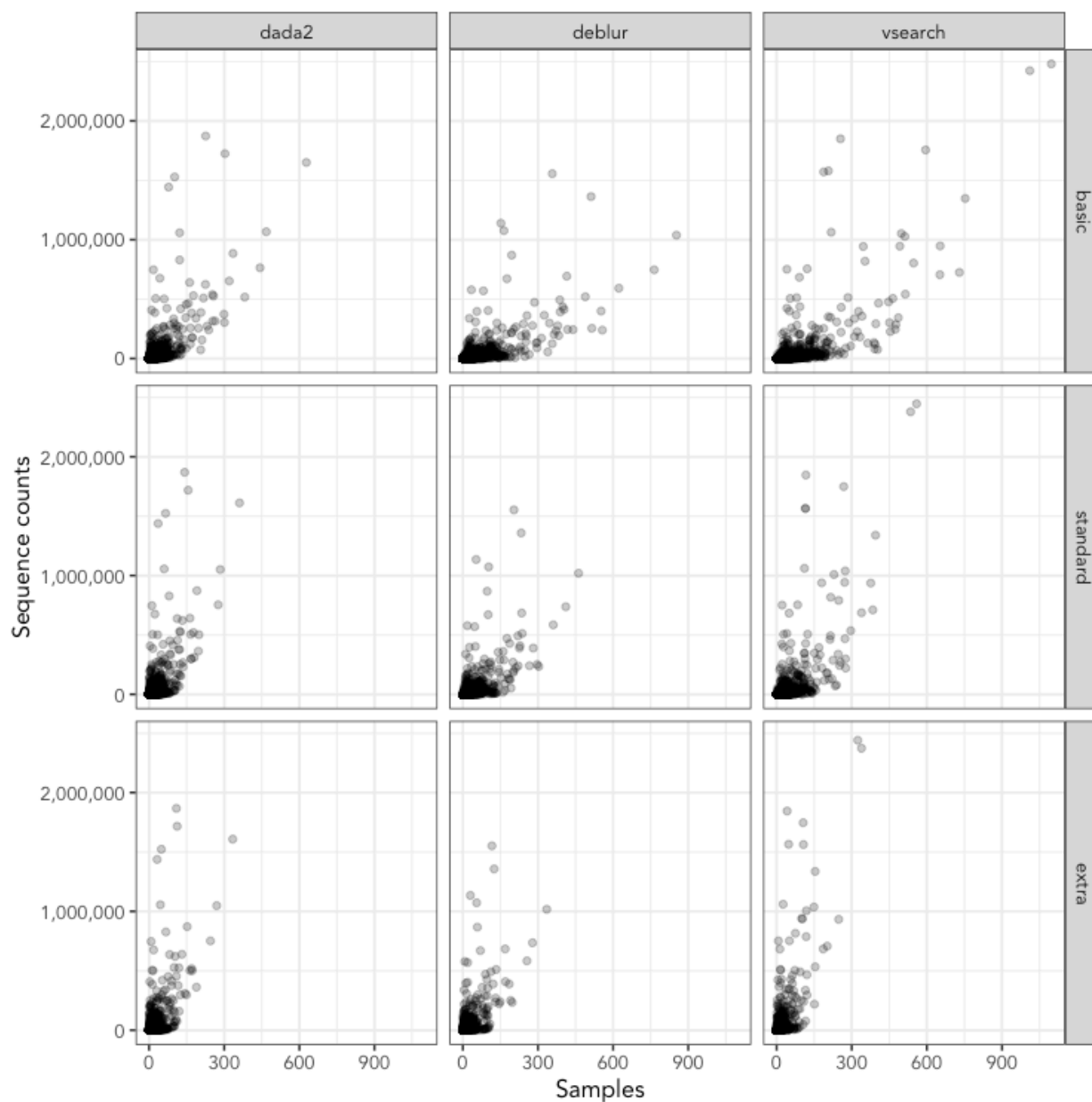
**Figure S2.** Per-library throughput among denoising programs. The proportion of filtered sequences varies by filtering method for both guano and mock community samples. Each point in the left facet (“guano”) represents fraction of all guano samples retained following processing with one of the three pipelines. The right facet (“mock”) points describe only a single data point of the individual mock sample sequenced for each library. Deblur retains fewer sequences per library than DADA2 or VSEARCH.



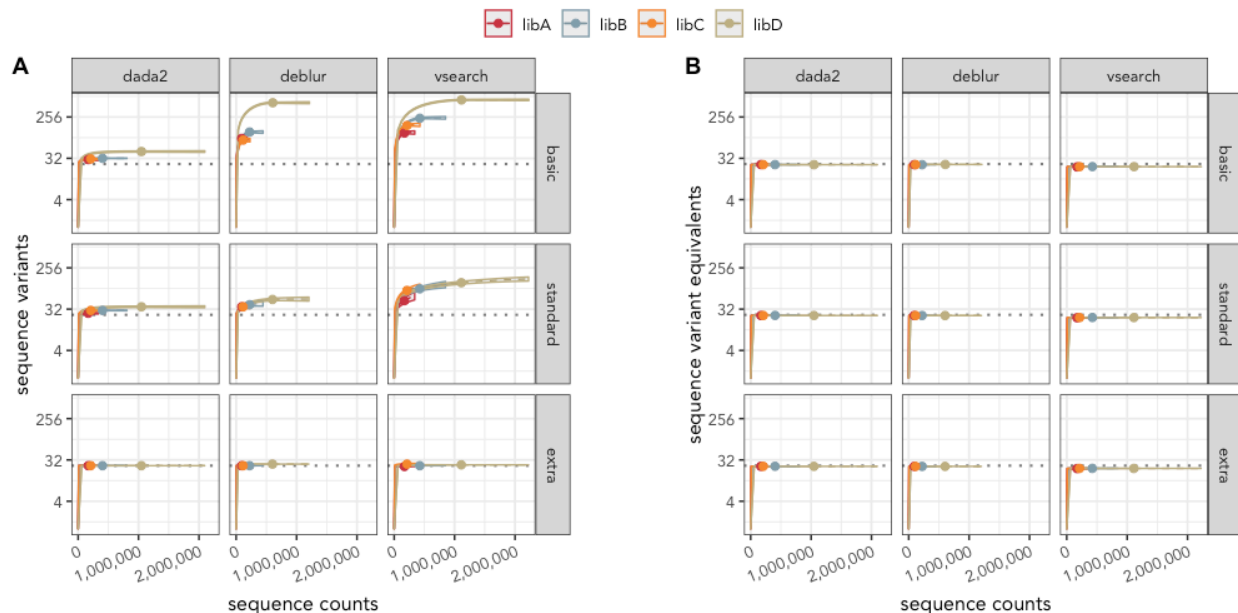
**Figure S3.** Total and fraction of expected reads in mock samples. Total (left plot) and Fraction (right plot) of expected reads in each mock community sample is influenced by Denoising pipeline (vertical facets). Exact matches reflect 100% alignment identity between ASV detected in mock community and a known mock sequence, partial match reflects between 97-99.9% identity, and miss represents an ASV with less than 97% identity to known mock sequences. All queries must span at least 97% of the reference target. While filtering parameters (horizontal facets) reduce the number of “miss” or “partial” detections (see Figure 2), the proportion of “partial” or “miss” reads is very low relative to the expected read abundances regardless of filtering strategy. However, the fraction of “partial” reads in VSEARCH is proportionally larger than in either denoising pipeline, indicating that the read correction strategies of the denoisers reduces the abundance and number of observed spurious sequence variants.



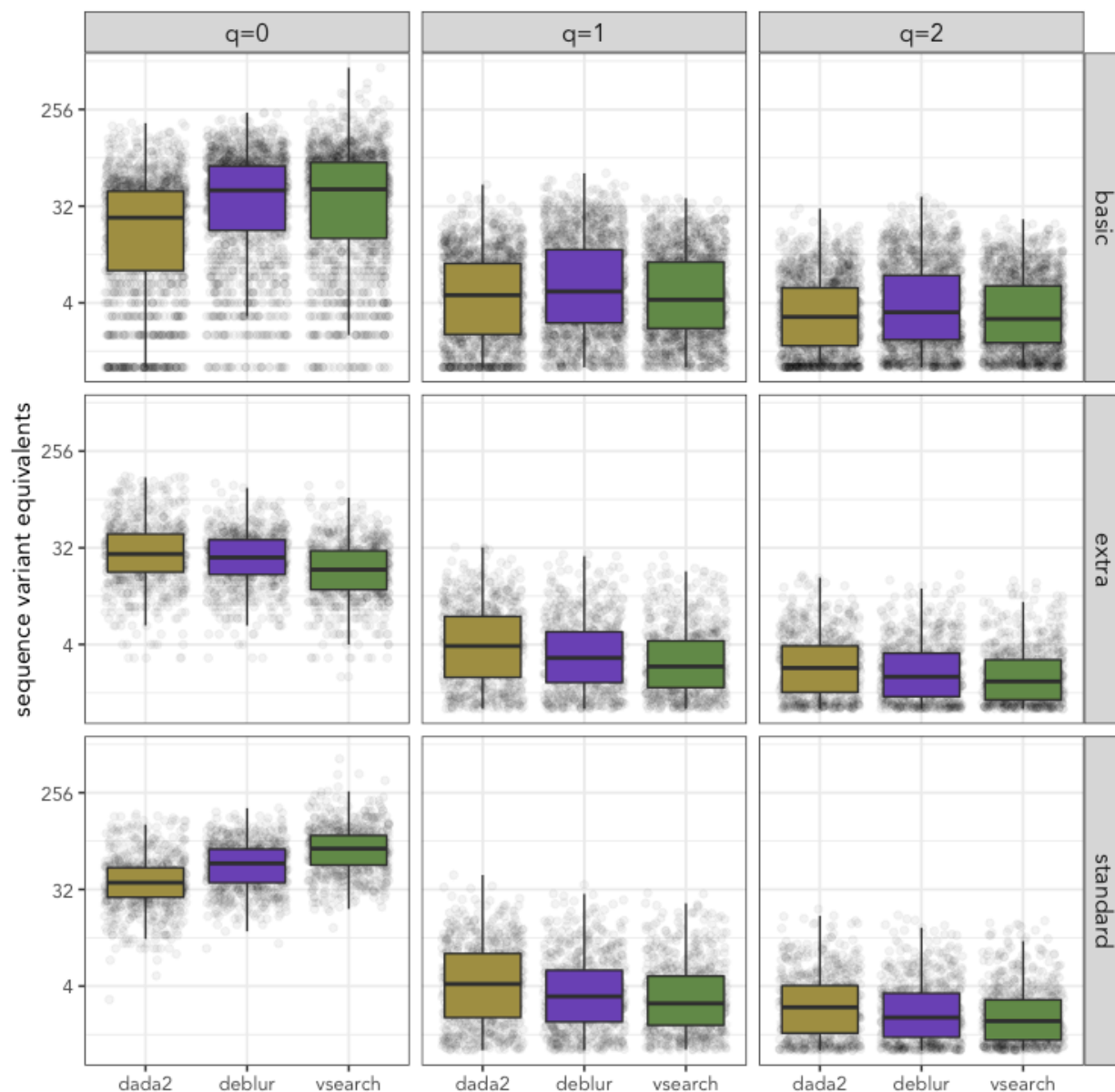
**Figure S4.** Read abundances per ASV for each library of bat guano data. DADA2 retains more sequence variants with larger abundances than Deblur or VSEARCH for default outputs and “standard” filter in each of the four libraries sequenced. Subtracting a fixed integer of reads from all ASVs with the “extra” filtering parameter discards many low-abundance ASVs from Deblur and VSEARCH datasets, resulting in distributions of read abundances that are more similar to those in DADA2.



**Figure S5.** Bat guano data generates a sparse matrix of ASVs among samples. Points represent sequence variants detected among all four libraries of guano data, suggesting the vast majority of ASVs are detected in few samples and generate relatively few sequences. In particular, note how the application of the “standard” and “extra” filters tend to reduce the number of samples an ASV is detected (x axis). These filters are dropping low abundance samples and ASVs, reducing the number of times a particular sequence variant is detected, resulting in a horizontal shift left of data points within a given denoising method (vertical facet).

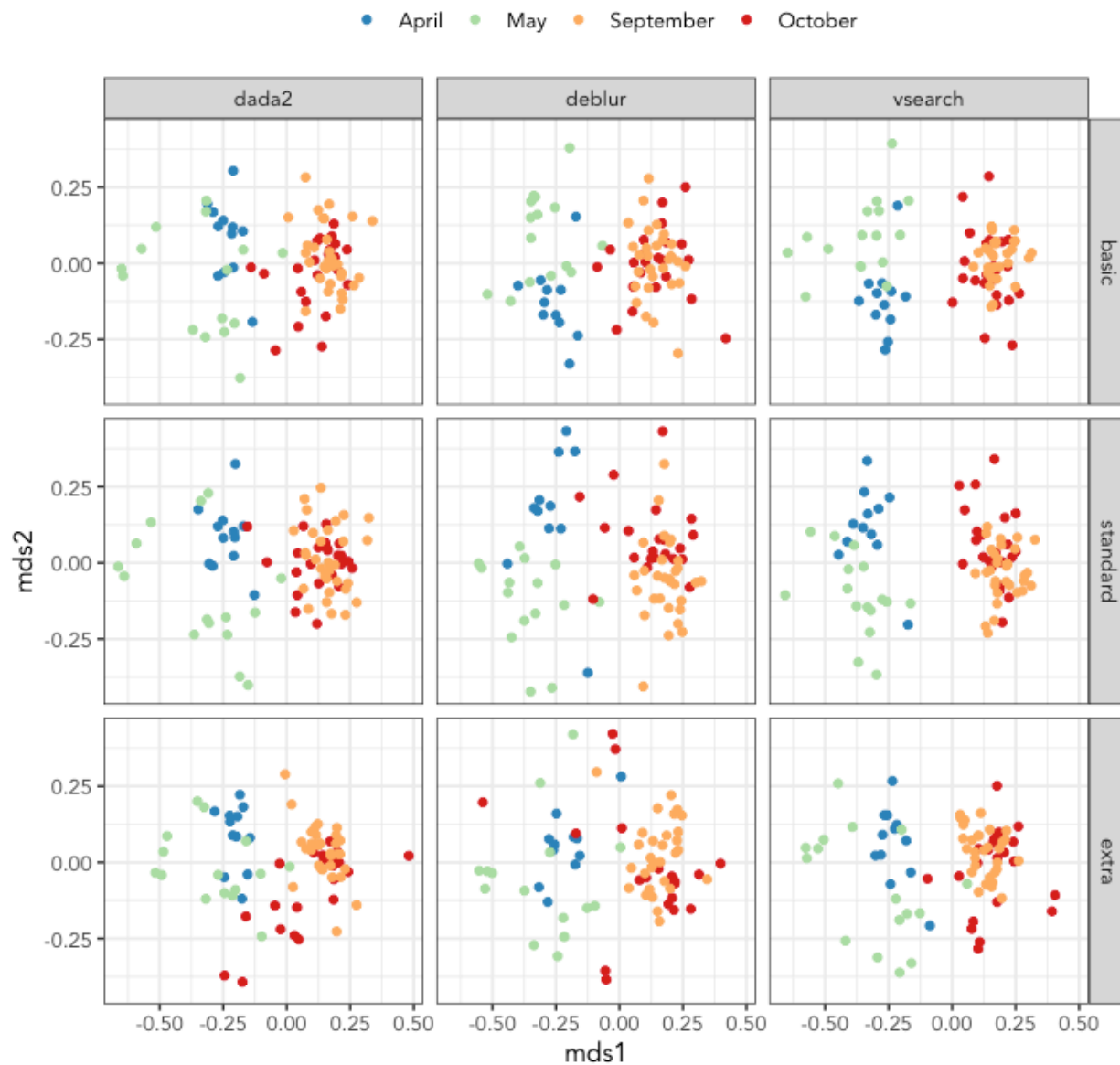


**Figure S6.** Accumulation curves of mock samples that are not rarefied. Interpolated (solid line) and extrapolated (dashed line) diversity estimates are shown for each mock sample relative to expected diversity (dotted line) for each denoising method (vertical facets) and filtering parameter (horizontal facets). Estimated diversity shown in (A) is calculated using Hill Number  $q=0$ , while estimates in (B) are calculated using Hill Number  $q=2$ . Note that the number of observed ASVs is more than twice as large compared to rarefied data (Figure 4). While DADA2 has fewer unexpected sequence variants compared to Deblur or VSEARCH, because all the unexpected ASVs are of low abundance these differences are negligible among diversity estimates that incorporate abundance information.



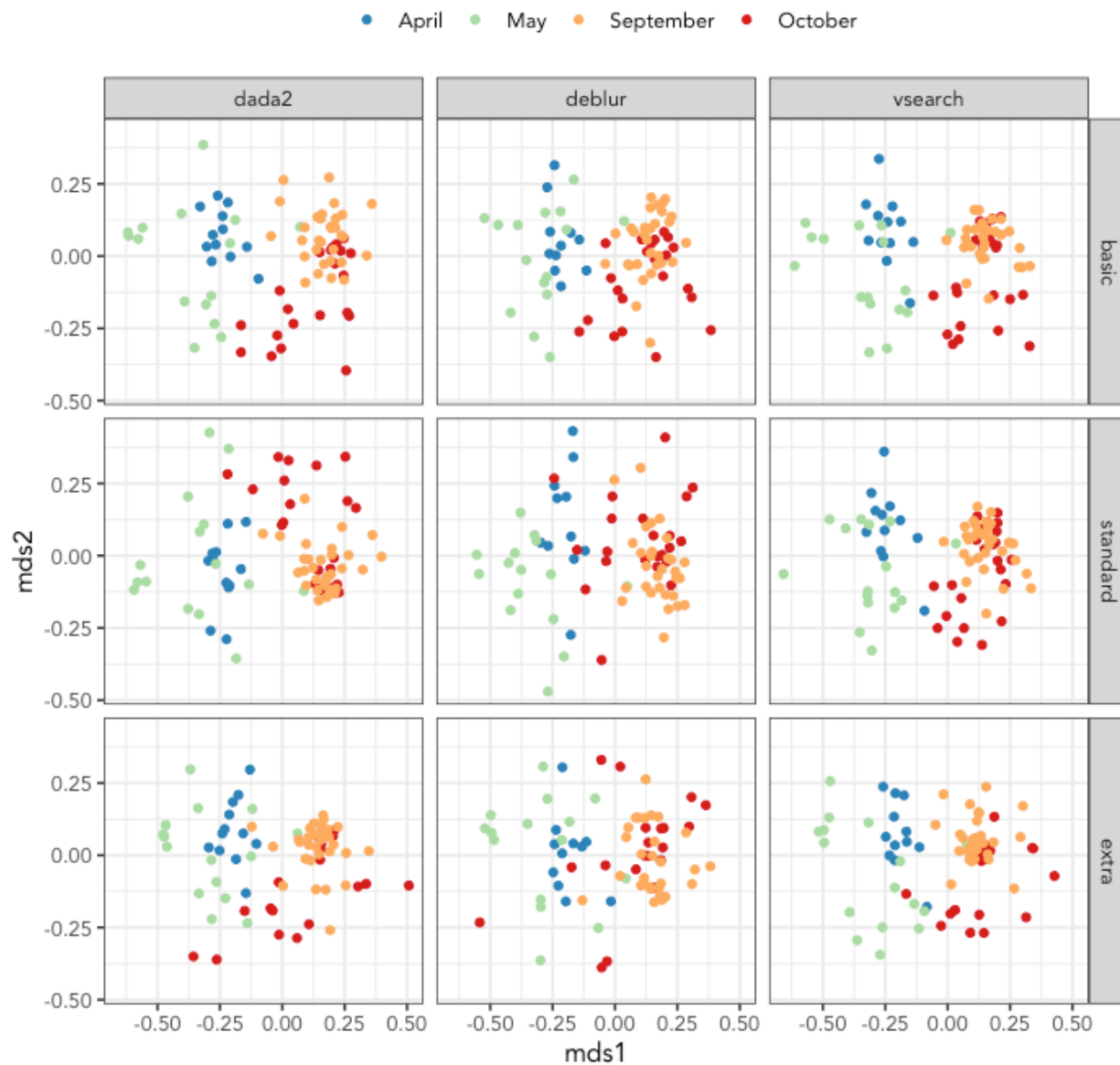
**Figure S7.** Diversity estimates for unrarefied bat guano. Denoising methods are compared within each subplot according to the respective Hill value (vertical facet) and filtering parameter (horizontal facet). Sequence diversity is reduced with increasing Hill number. Applying additional filtering parameters increases diversity estimates at a Hill a value of 0 (equivalent to observed OTUs) because rare ASVs are dropped and samples with low numbers of sequences are discarded from analysis. Diversity estimates for  $q=1$  and  $q=2$  are reduced compared to  $q=0$  estimates, indicating that relatively few distinct sequences encompass the majority of the sequence information in a given sample.

Dice-Sorensen index used to calculate distnaces



**Figure S8.** Ordination of distances calculated by Dice-Sorensen index using NMDS for select guano data for each combination of denoising method and filtering parameter. Samples (points) are colored according to the month they were collected.

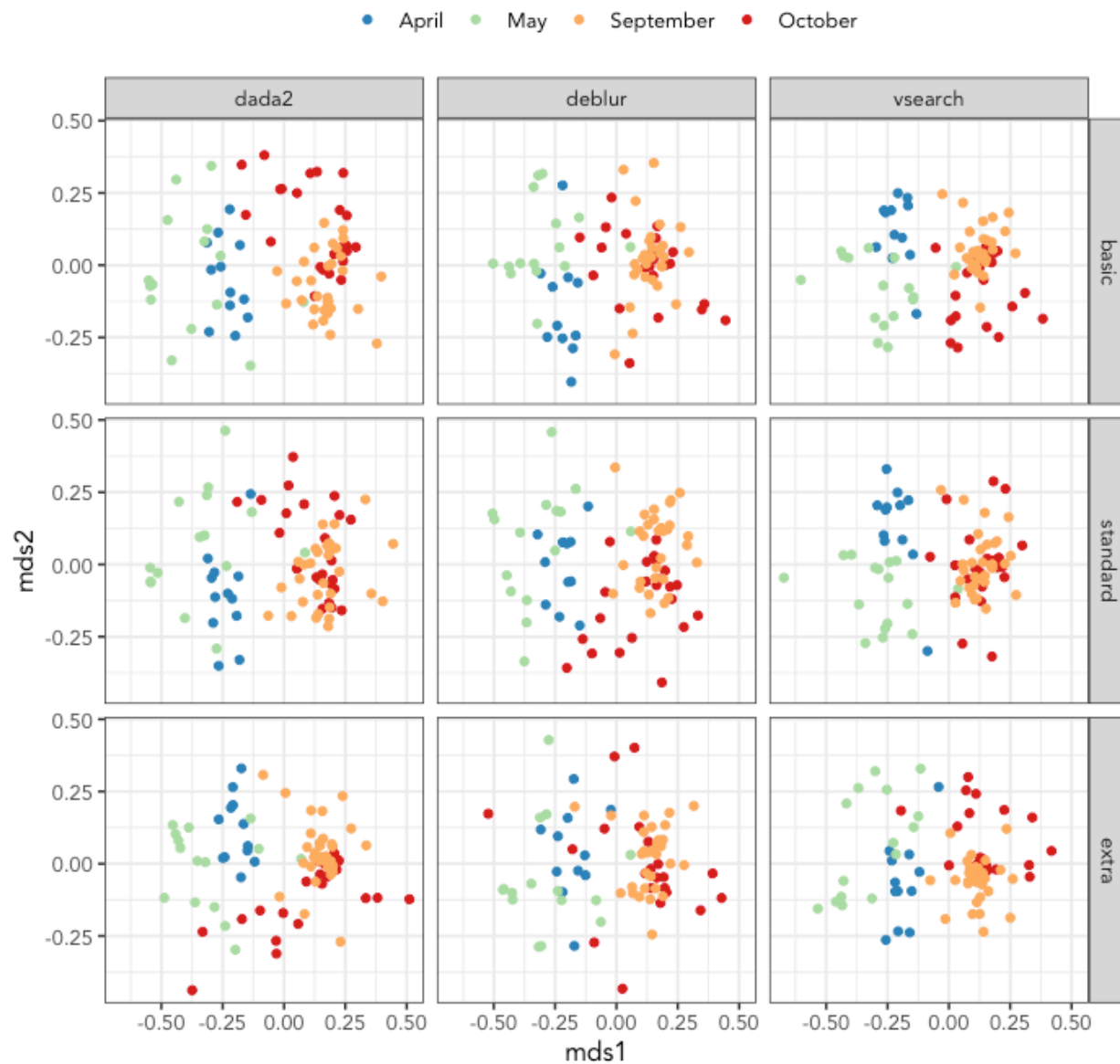
Bray-Curtis index used to calculate distnaces



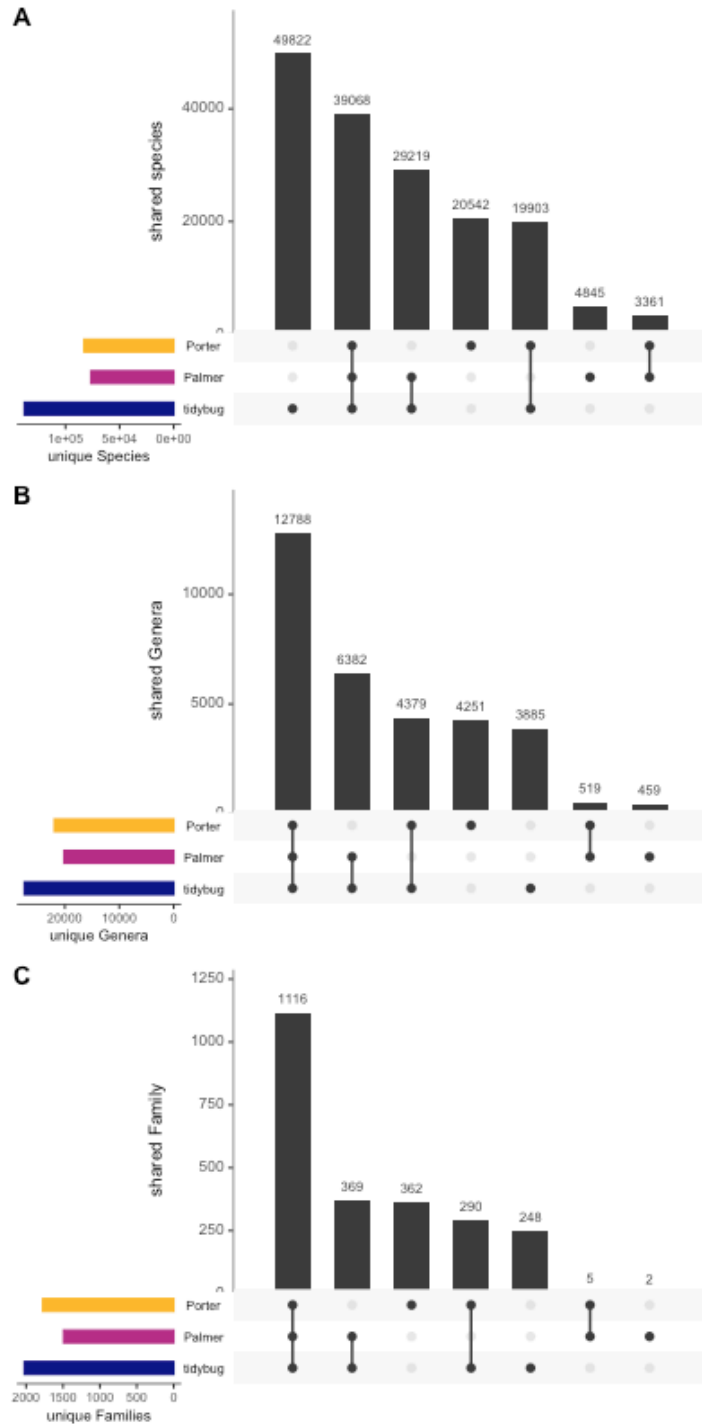
**Figure S9.** Ordination of distances calculated by Bray-Curtis index using NMDS for select guano data for each combination of denoising method and filtering parameter. Samples (points) are colored according to the month they were collected.



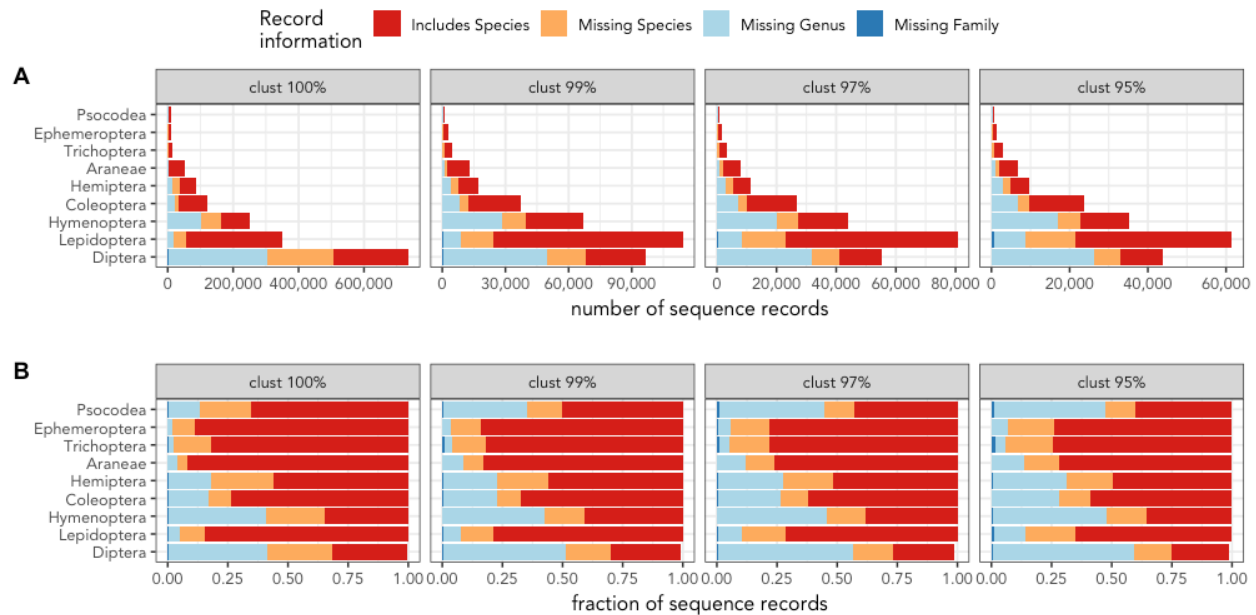
Morisita-Horn index used to calculate distnaces



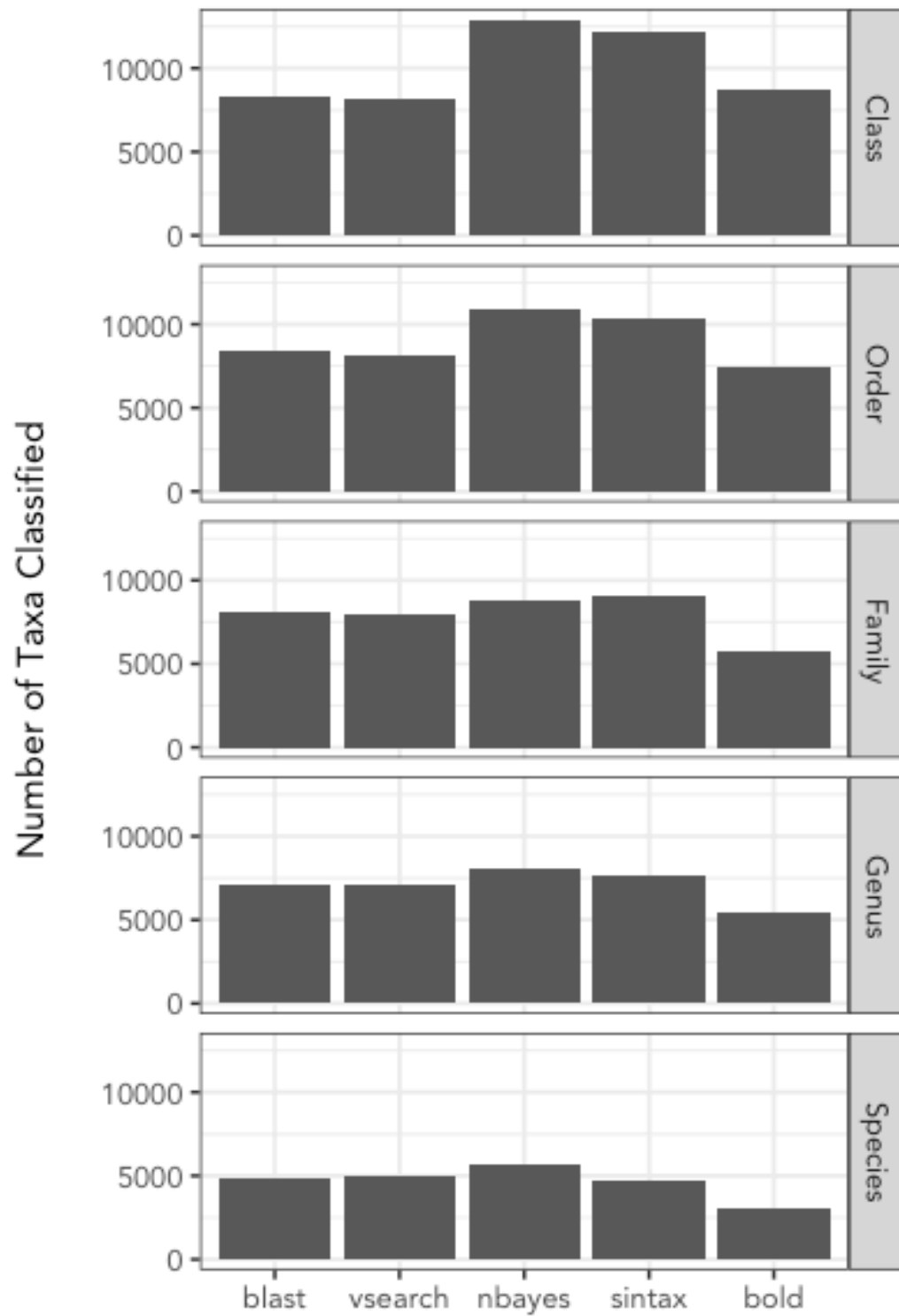
**Figure S10.** Ordination of distances calculated by Morisita-Horn index using NMDS for select guano data for each combination of denoising method and filtering parameter. Samples (points) are colored according to the month they were collected.



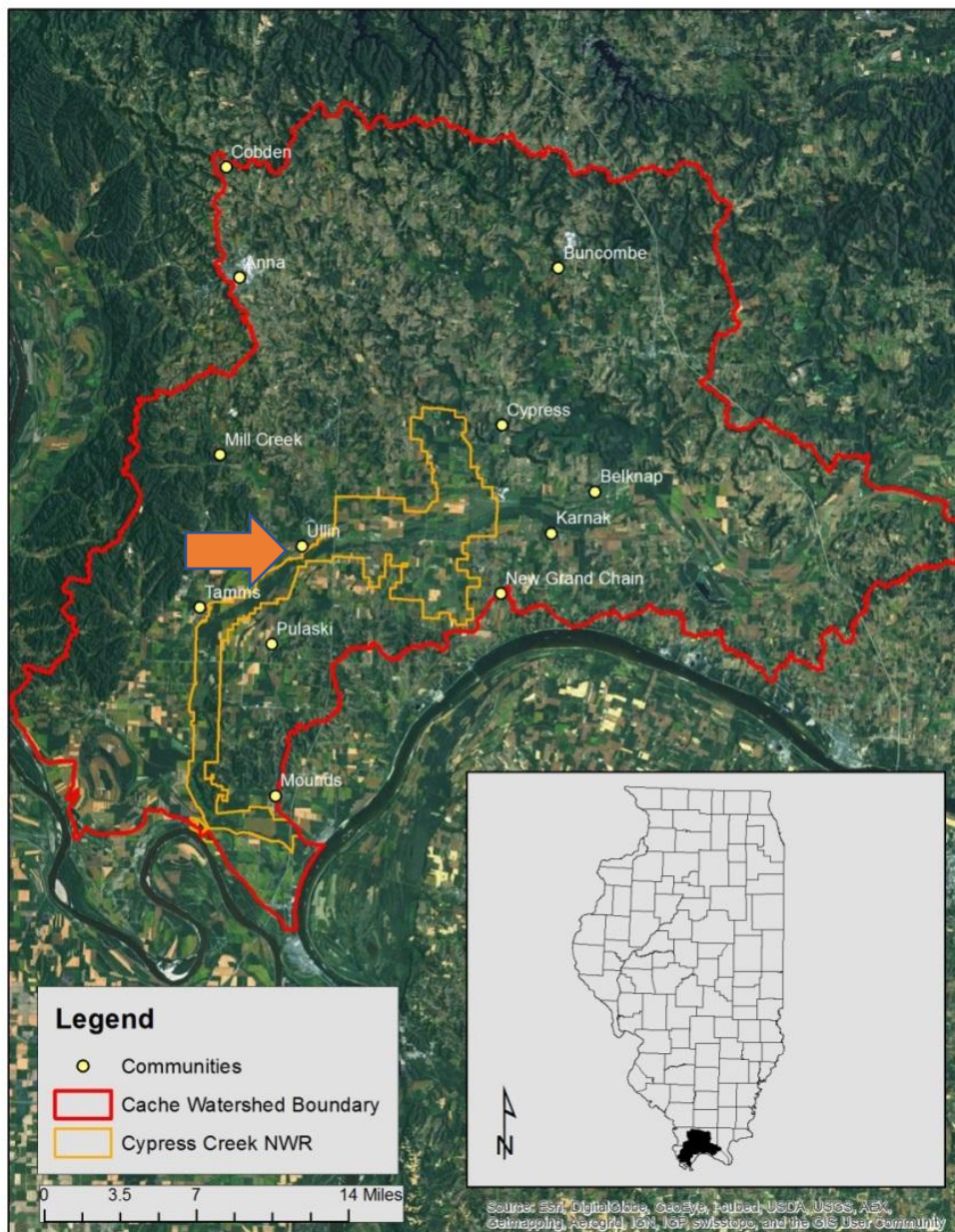
**Figure S11.** Unique and shared taxa among three databases. Vertical bars indicate the number of shared taxa among databases (points connected in matrix) or those that are unique to a single database (unconnected points in matrix). Horizontal bars indicate the total number of taxa in a database. Taxa are represented at three levels: **(A)** references with distinct Species, **(B)** references with distinct Genus, **(C)**, references with distinct Family. While all three databases share the greatest number of shared Genera and Family taxa, our tidybug database contains more distinct Species than what is shared among all datasets.



**Figure S12.** Effect of clustering on taxonomic diversity by arthropod order. Clustering the tidybug database results in non-uniform reduction of sequence diversity and taxonomic diversity among top abundant arthropod Orders. Some references contain all taxonomic information (red), while other records are missing Species name (orange), Species and Genus names (light blue), or Species, Genus, and Family information. **(A)** The total number of unique arthropod COI sequences are most abundant for Dipteran Order among the the dereplicated dataset (clust 100%), but additional clustering results in fewer overall Dipteran sequences relative to Lepidopteran records. **(B)** The fraction of Clustering results in non-uniform reduction of information content among top abundant arthropod Order sequence records. Though clustering uniformly reduces information content, some Orders (ex. Ephemeroptera) retain a greater fraction of information after clustering than others (ex. Diptera, Psocodea), while others are relatively unchanged (ex. Hemiptera, Hymenoptera).

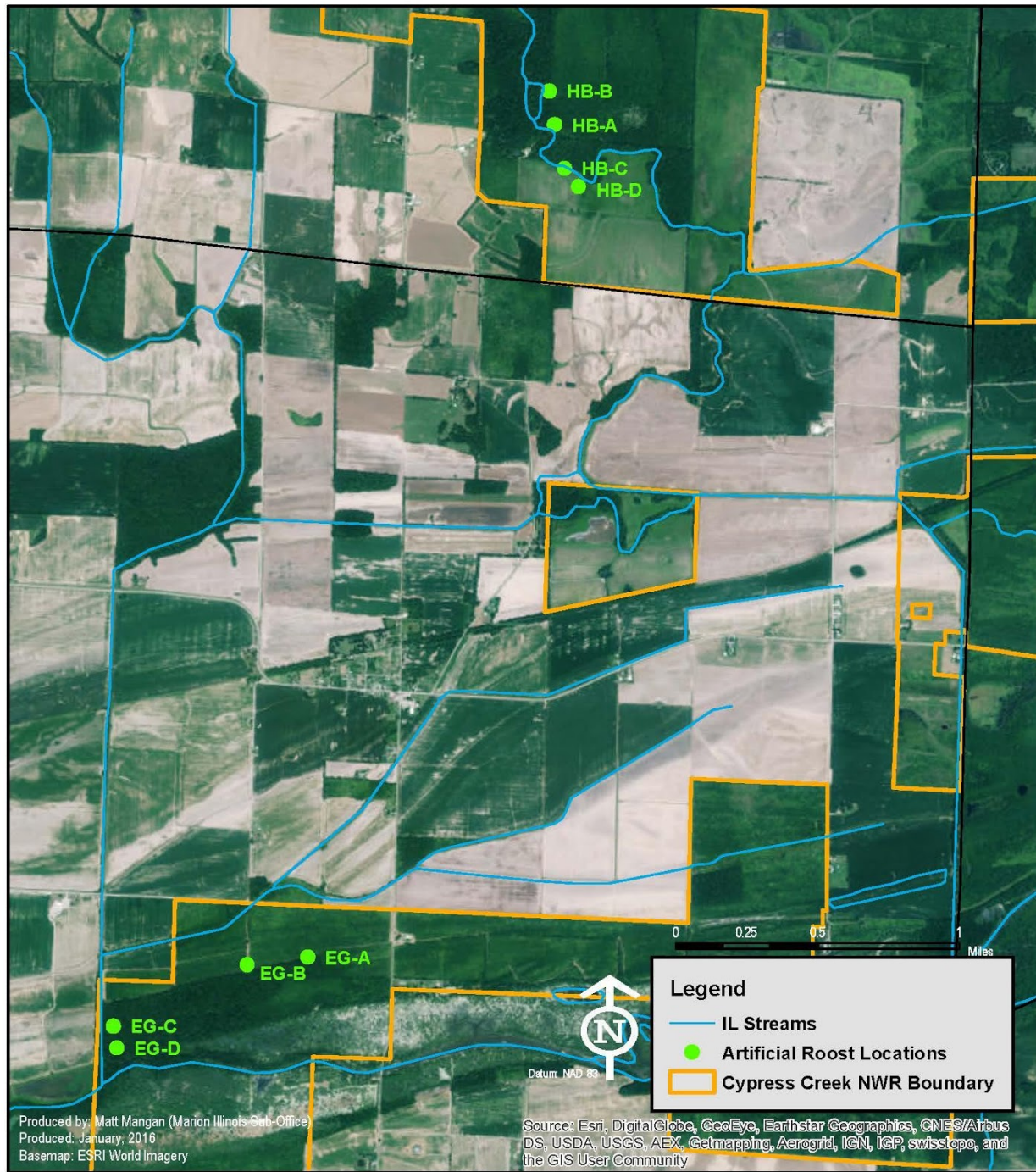


**Figure S13.** Number of ASVs assigned taxonomic information at Class through Species rank among classifiers for guano data. The kmer-based classifiers Naive Bayes and SINTAX assign more ASVs taxonomic identities than alignment based VSEARCH and BLAST classifiers.

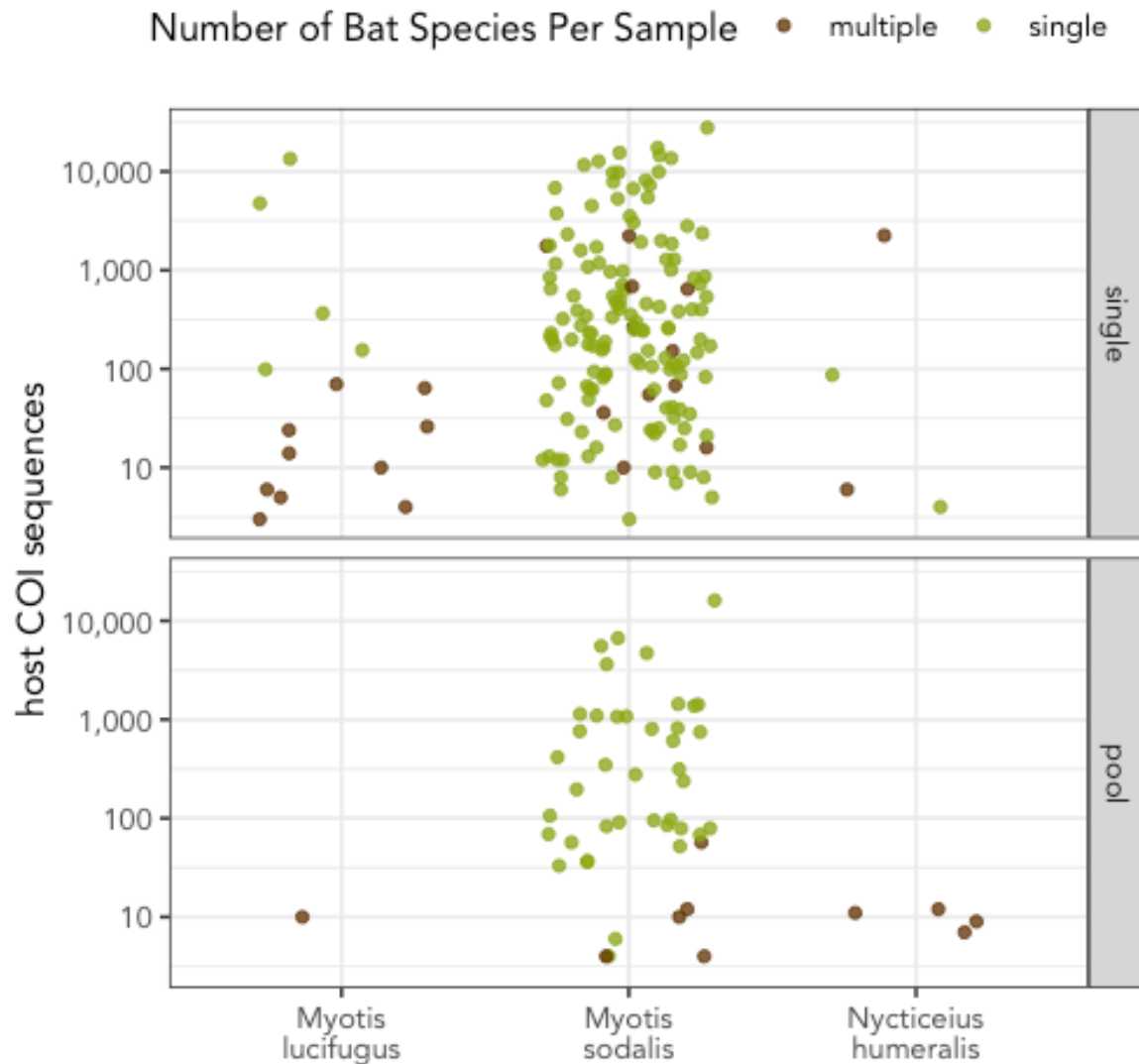


**Figure S14.** Map of Cypress Creek National Wildlife Refuge. The Cypress Creek National Wildlife Refuge (orange boundary line) is situated in southern Illinois and consists of 35,320 acres of wetland, forested, and grassland habitats. Among the actively managed units within the CCNWR, the Cache River Unit comprises just 1,136 acres of riparian wetland and forests essential to Indiana Bat foraging and roosting. Guano collections were focused in an area just east of Ullin, IL, denoted as a yellow point in the figure with the orange arrow.



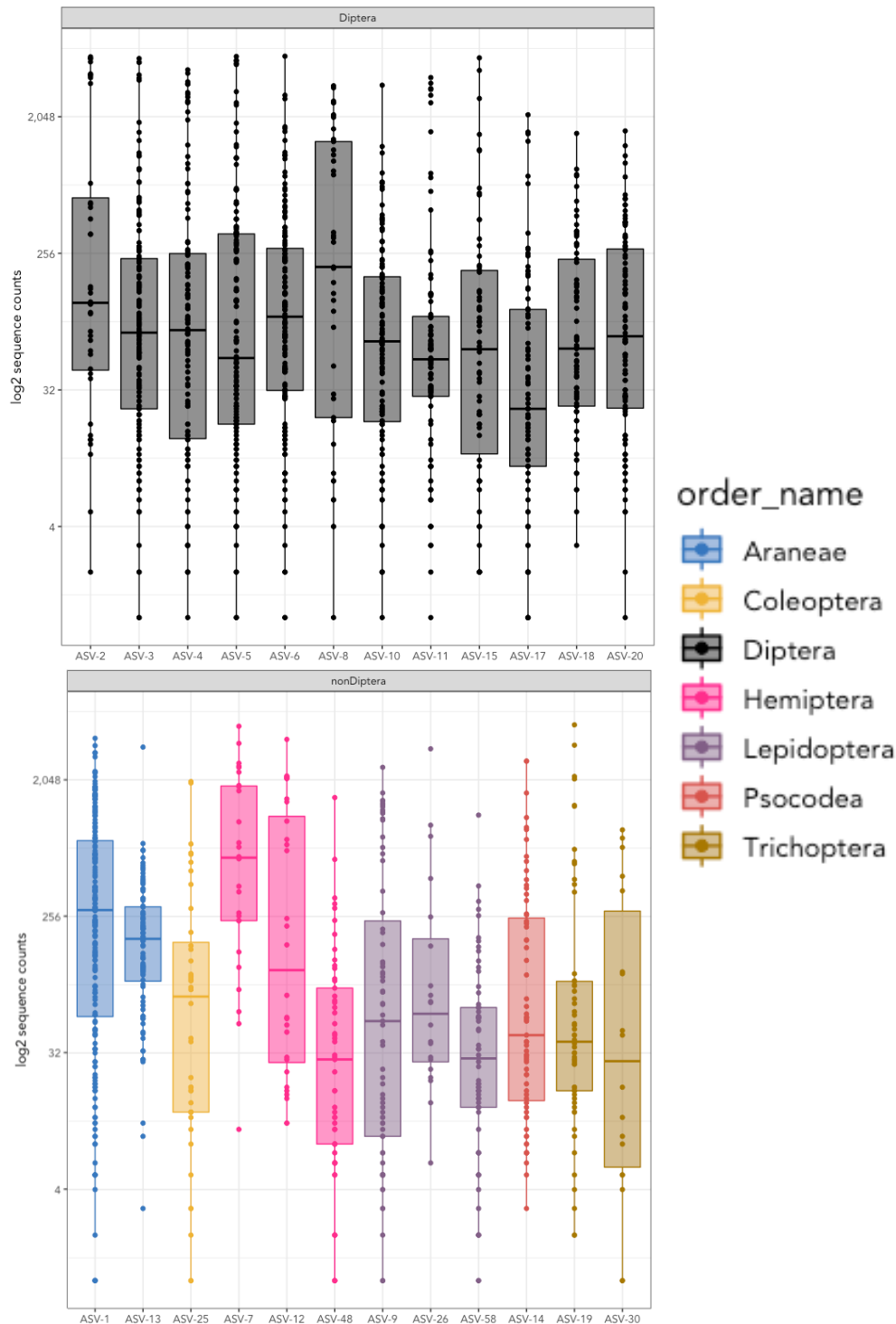


**Figure S15.** Artificial Roost locations at Cypress Creek NWR. Eight artificial roost structures (green dots) were installed in 2014 in areas where limited or no roosting habitat was previously detected. Emergence surveys, mist netting, and acoustic monitoring in 2016 confirmed presence of Indiana Bat use of each structure. Note that while roost sites are within CCNWR boundaries, the landscape is a mix of protected wetland and forested landscapes fragmented amongst privately owned agricultural areas. Our data suggest the bats are frequently foraging aquatic invertebrates, thus protection strategies will require not only conserving suitable roosting habitat, but likely also a renewed focus on the stream systems these artificial roosts were near.

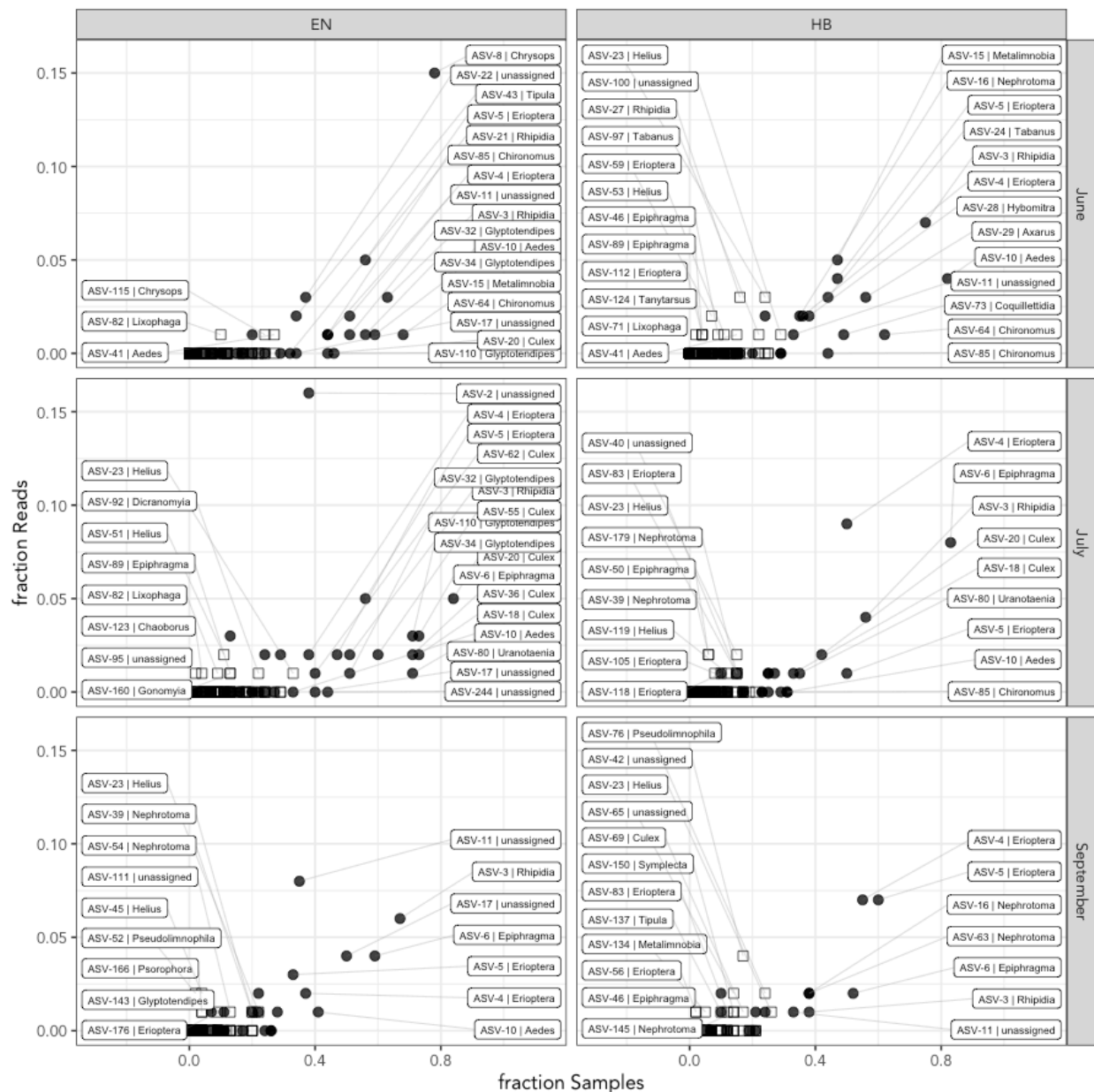


**Figure S16** Sequence depths per ASV classified as bat. Bat sequence information indicates guano is largely from Indiana Bats. Our COI sequence data consisted of both arthropod and non-arthropod sequences, and among these were some variants classified to one of three bat species: *Myotis lucifugus* (Little Brown Bat), *Myotis sodalis* (Indiana bat), and *Nycticeius humeralis* (Evening Bat). The majority of guano DNA amplified generated some bat sequence COI data (184 of 279 samples), though the vast majority of these samples were classified as Indiana Bat (177 samples). Instances of non-*sodalis* detections generally contained very few total reads (y axis) and were often found in samples in which *M. sodalis* DNA was also detected. In the plot below, points represent samples, with colors “multiple” (brown dot) and “single” (green dot) indicating whether or not a sample contained bat COI sequences classified to one or more species. We were no more likely to collect these low-abundant non-*sodalis* COI bat sequences whether or not the DNA was extracted from single guano pellets (“single” facet, top) or from a batch of guano pooled into a single extraction (“pool”, bottom facet), nor were the pools more likely to contain multiple bat species per sample. It is therefore likely that non-*sodalis* bats were transiently occupying these roosts, and that our passive guano sampling regime permitted the infrequent cross contamination of multiple bat species.

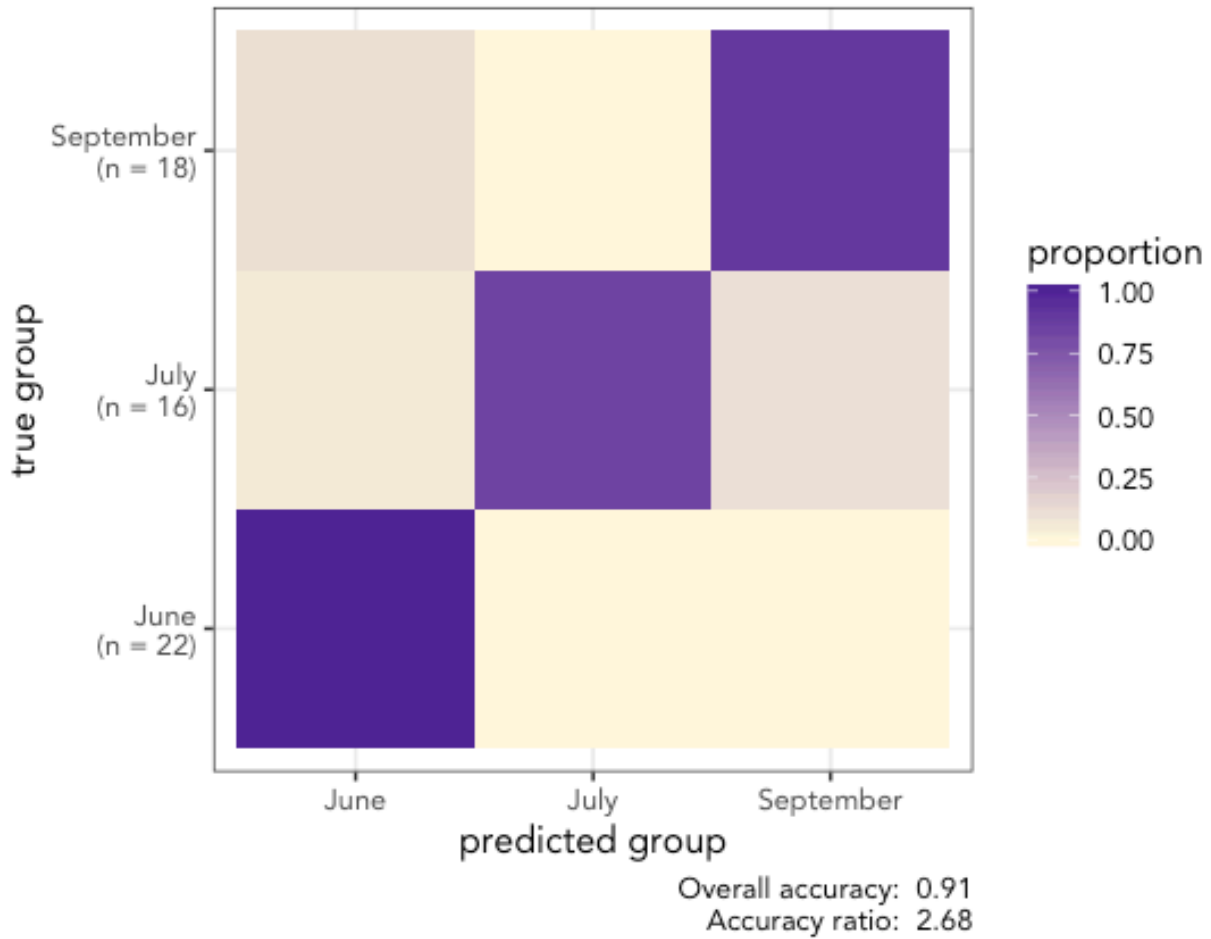
**Figure S17.** Distribution of sequence counts among frequently detected ASVs. Primers are not biased towards Dipteran sequences. Per-sample rarefied sequence counts of (points) vary among many of the most frequently detected Dipteran (“Diptera”, left facet) and non-Dipteran (“nonDiptera”, right facet) ASVs. We would predict a proportionally larger number of sequences per sample among Dipteran ASV than non-Dipteran ASV if primers preferentially amplified Dipteran taxa, but we observe ASVs that generate highly abundant amplicons within a given sample among many arthropod Orders.



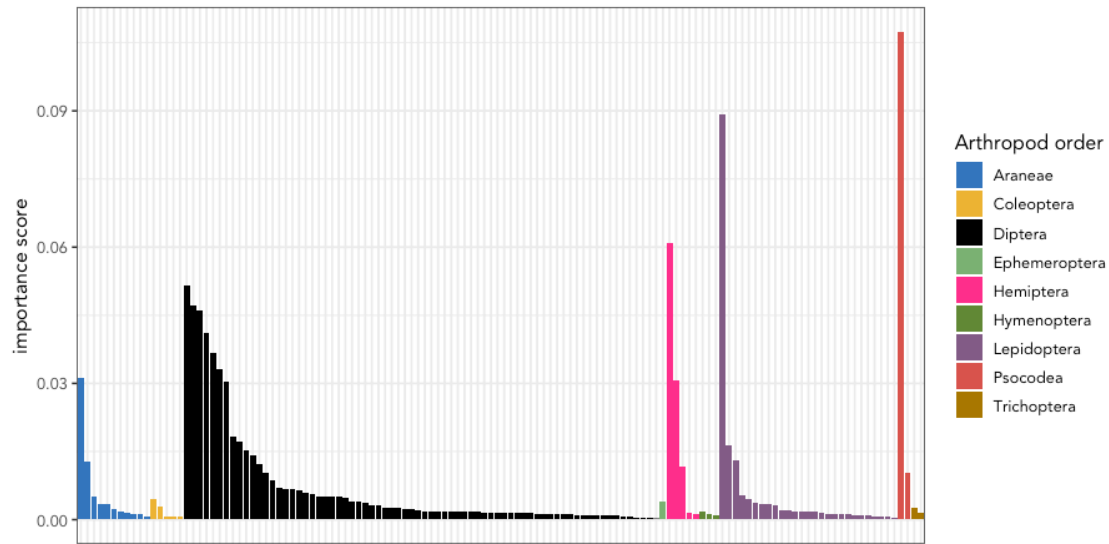




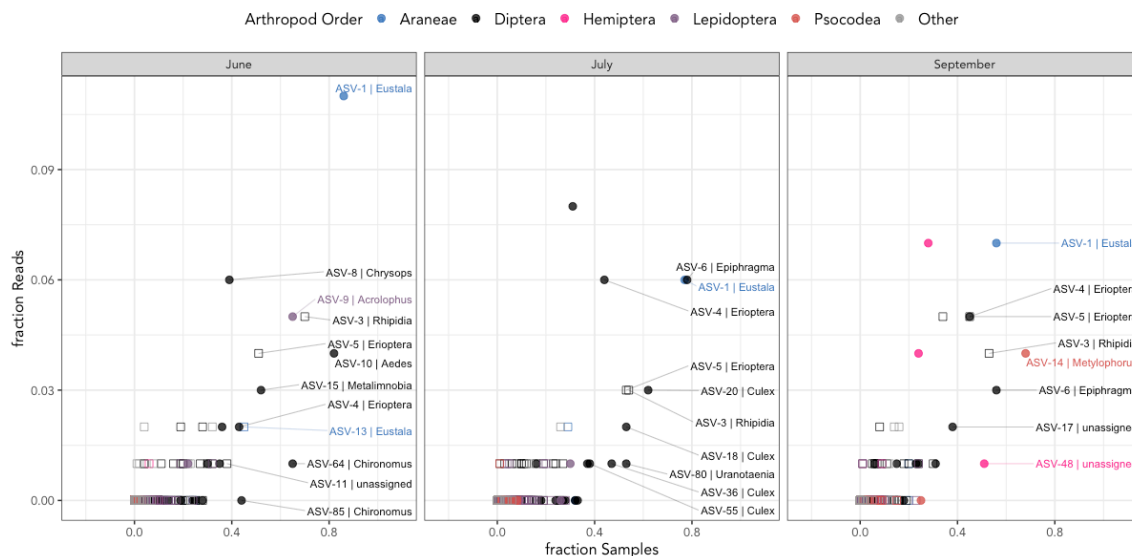
**Figure S18.** Dipteran ASV abundances and detections per site and month. The fraction of rarefied reads (y axis) and fraction of samples (x axis) for a particular Dipteran-classified ASV often change by Month and/or Site. In the plot below, a Random Forest model identified specific ASVs as important to classifying samples by Site and Month (solid points, labeled on the right of each plot). Some of these ASV's have high abundances and detections in both Sites, but the direction of those changes differ by Month of collection (e.g. ASV-3 is more frequently detected and generates the most reads in Egner collections in June, while the same ASV is more abundant and detected in July for Hickory Bottom Samples). Other ASV's generate similar trends in detections per Month, but have distinct magnitudes with respect to read abundances (e.g. ASV2 was identified in 12 samples at Egner and 17 samples at Hickory Bottoms sites in July, yet Egner samples collectively generated 37,047 sequences compared to just 1,126 in Hickory Bottoms). However, a few ASVs not identified by the classifier had similarly high fractions of detections or abundances as ASVs identified in the model, yet these were typically constant across Months (ex. ASV-23). Collectively, these data suggest that the bats are repeatedly targeting Dipteran taxa, yet many of the specific targets are changing with the summer season.



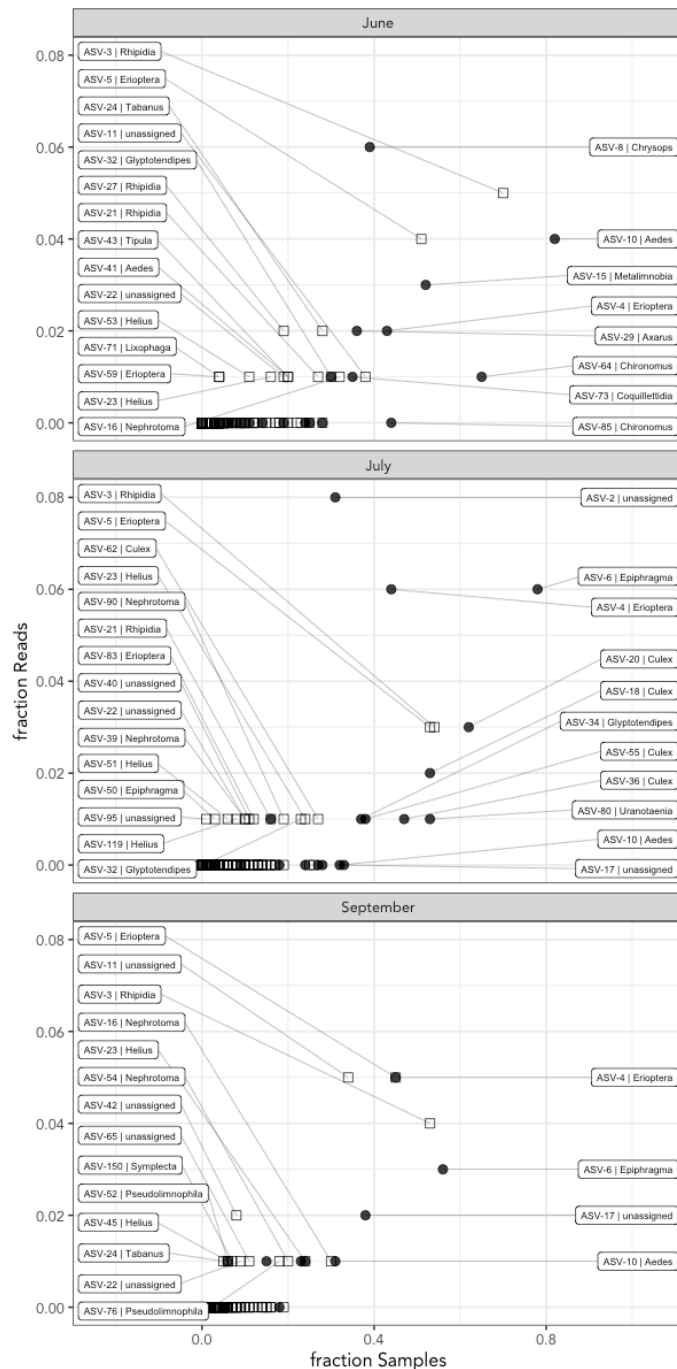
**Figure S19.** Random forest classifier accuracy by Month. A Random Forest model was trained to identify the features (ASVs) relevant to classifying a guano sample to a particular collection Month (June, July, or September). We trained the model using 80% of the dataset, then tested the model accuracy by classifying the remaining 20% of samples. This confusion matrix demonstrates that the model was highly successful at matching the predicted group with the true group for most samples. Fewer features were required for model training for this Month-only classifier than the Site and Month classifier (128 ASVs vs. 237 ASVs) but the same subset of ASVs were identified in both sites, and the majority of these were Dipteran. Thus while these bat guano samples contain largely Dipteran sequence data, many of the Dipteran-classified ASVs have substantial changes in abundances and detections by Month, indicating that bats are particularly variable in their summer fly foraging habits.



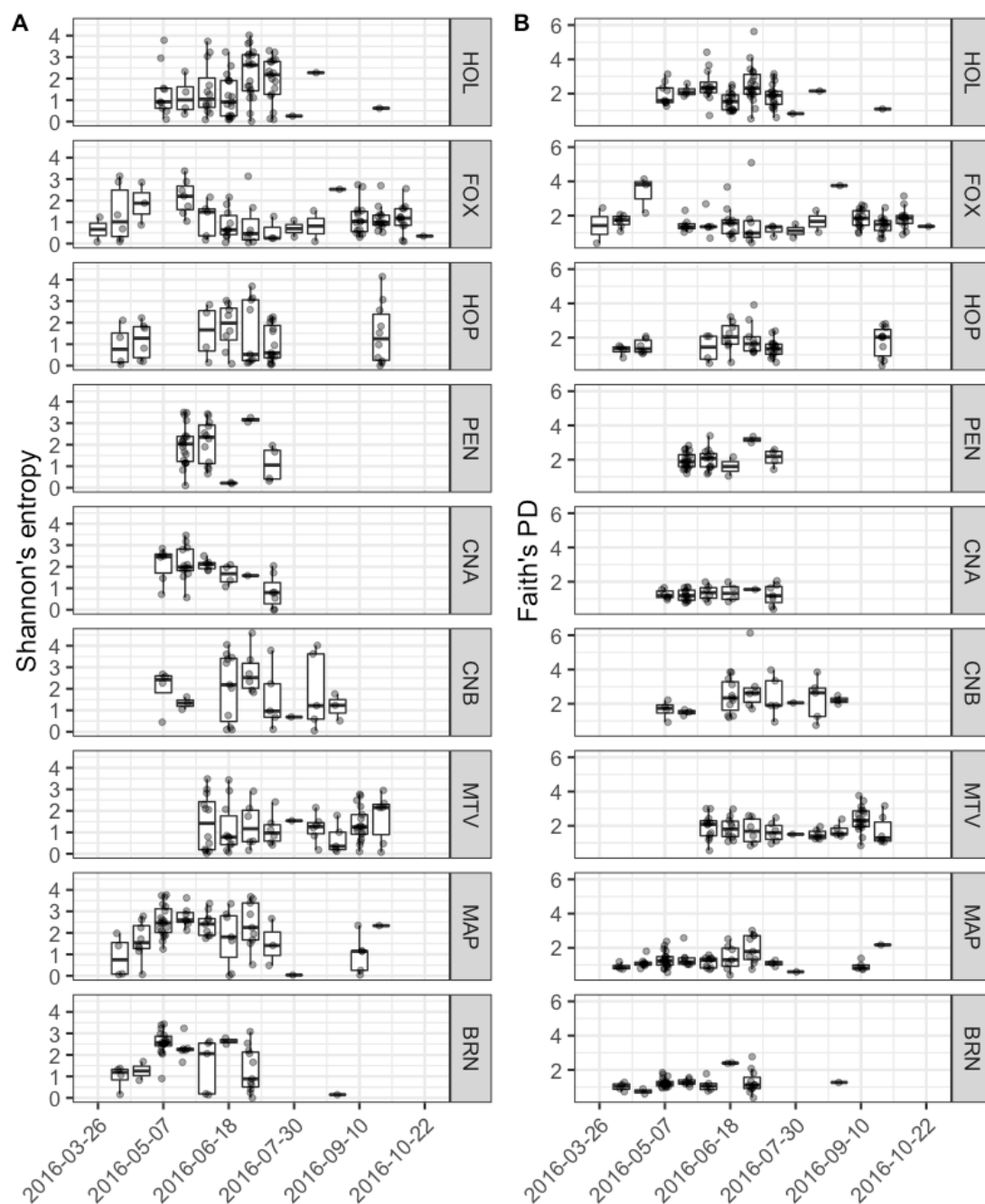
**Figure S20.** Important ASVs to Random Classifier model by Month. More fly ASVs vary temporally than any other arthropod order, but the individual ASVs most sensitive to temporal variation are non-dipteran orders. Among the 75th percentile of ASV importance to a random forest classifier trained to discriminate sample composition by month of collection, most of these ASVs are dipteran, though three other orders (Hemiptera, Lepidoptera, and Psocodea) contain the highest individual ASV importance values, an indication of the strongest temporal signal in the model.



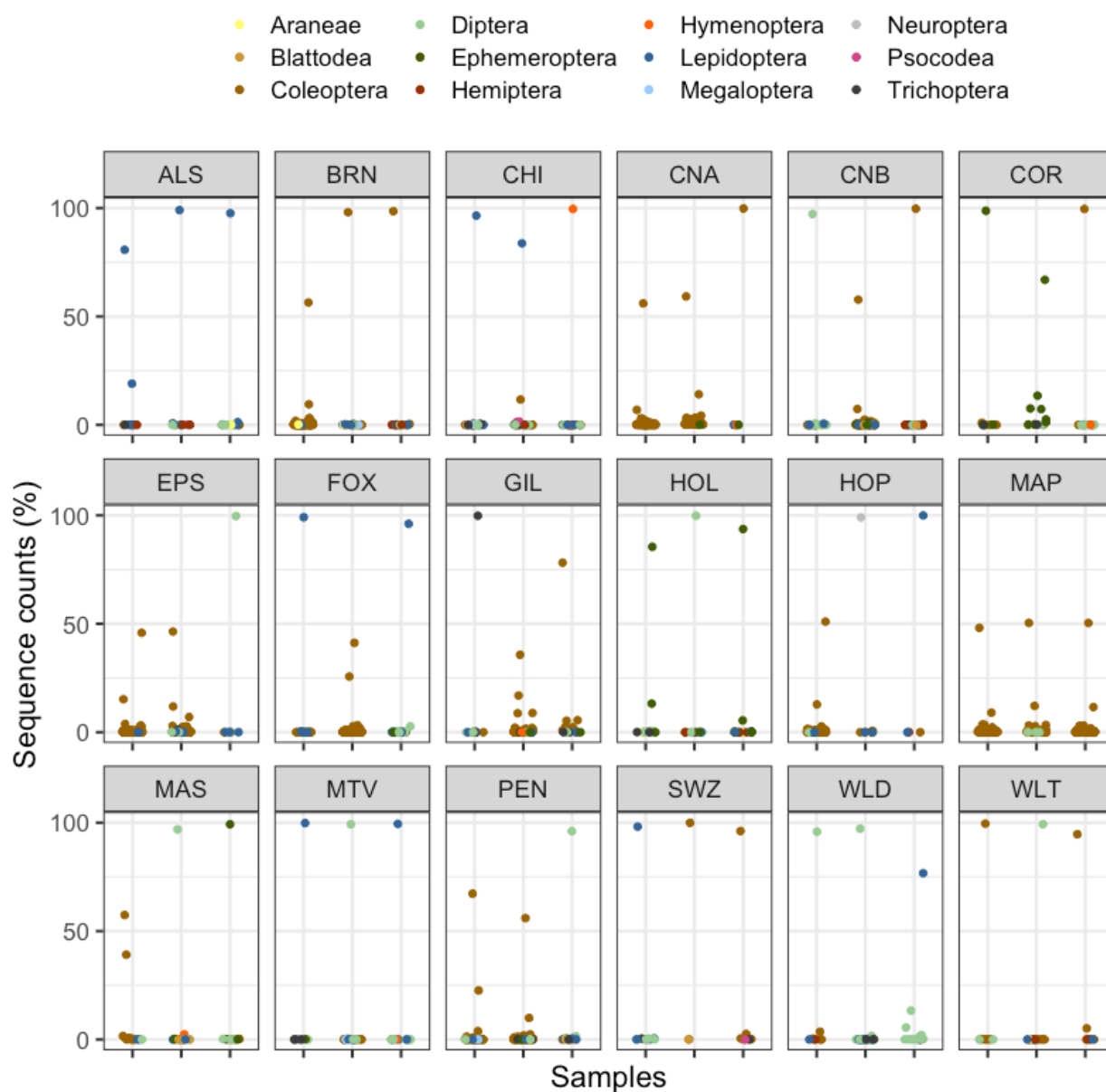
**Figure S21.** Per Month ASV abundance and detections. What's for dinner depends on the Month, but most likely, it's a fly. The fraction of rarefied reads per Month (y axis) and fraction of samples per Month (x axis) varies for some ASVs but not others. Some ASVs were identified as important to a Random Forest model trained to classify samples by collection Month (circles), though other ASVs were highly abundant throughout the entire collection season and therefore not relevant to the model (squares). Hemipteran and Psocodean taxa are highly abundant in September, suggestive of bats opportunistically foraging on particular non-Dipteran taxa that emerge on the wing as adults in the latter summer. However the majority of variable monthly prey is detected among the Dipteran taxa. For example, despite summer-long detections of mosquitoes, different Genera are more abundant in early months (e.g. *Aedes*; ASV-10) than in later months (e.g. *Culex*; ASV-18, ASV-20).



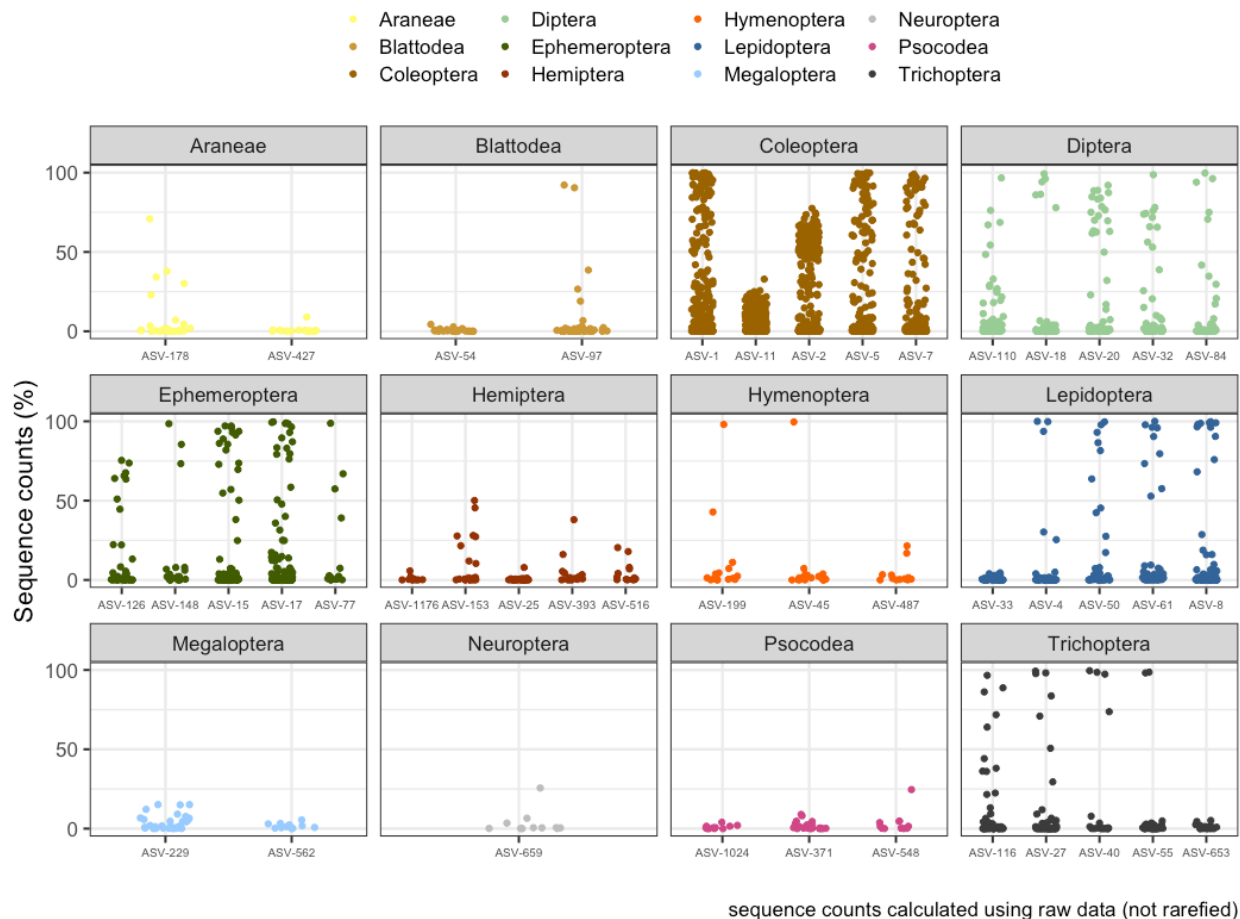
**Figure S22.** Dipteran ASVs relevant to Random Forest classifier. The fraction of rarefied reads per Month (y axis) and fraction of samples per Month (x axis) varies for some Dipteran ASVs but not others. This plot represents the Dipteran-specific subset of data described in Figure S7. ASVs identified as important to a Random Forest model classifying samples by Month (circles, labels at right of plot), and ASVs not important to the model (squares, labels at left of plot) are largely similar to those features that were relevant to a separate supervised learning model that was classifying samples by Site and Month. However, some of the most frequently detected/abundant taxa are no longer important to the Month-only classifier (but were previously relevant to the Site and Month classifier). For examples, ASV-3 and ASV-5 and are foraged repeatedly by the bats, but these taxa were more frequently foraged at one site than another; these site differences were relevant to the earlier classifier but are not here. These taxa highlight the distinction between "important" to classifier versus "important" to the bat; clearly these ASVs are relevant prey targets.



**Figure S23.** Alpha diversity estimates of select 2016 New Hampshire sites. Intra sample diversity using abundance (Shannon's entropy) and phylogenetic (Faith's PD) information. Samples were selected from 2016 locations with the greatest sampling intensity. Samples are binned in 14-day windows per site. **(A)** Shannon's entropy values indicate that most samples have relatively low information, indicating that few sequence variants are commonly observed in a sample (i.e. most sequence variants are low abundance), particularly at the earliest and latest parts of the sampling season. **(B)** Instances of elevated Faith's PD values are observed asynchronously among sampling windows across sites, indicating a collection of sequence variants with more diverse evolutionary backgrounds.



**Figure S24.** Proportions of sequences per ASV per sample. Highly abundant samples from each site are shown (the most abundant sample in three independent weeks was selected for each site). The most abundant ASV in a sample can arise from many different arthropod orders, indicating that a broad range of taxonomic sequences can be highly amplified by our molecular workflow, however, even the most deeply sequenced samples tend to have highly uneven abundances per ASV. Abundance-based metrics may overemphasize the importance of a single dietary component in similarity estimates, while binary measures may instead bias towards rare sequence variants.

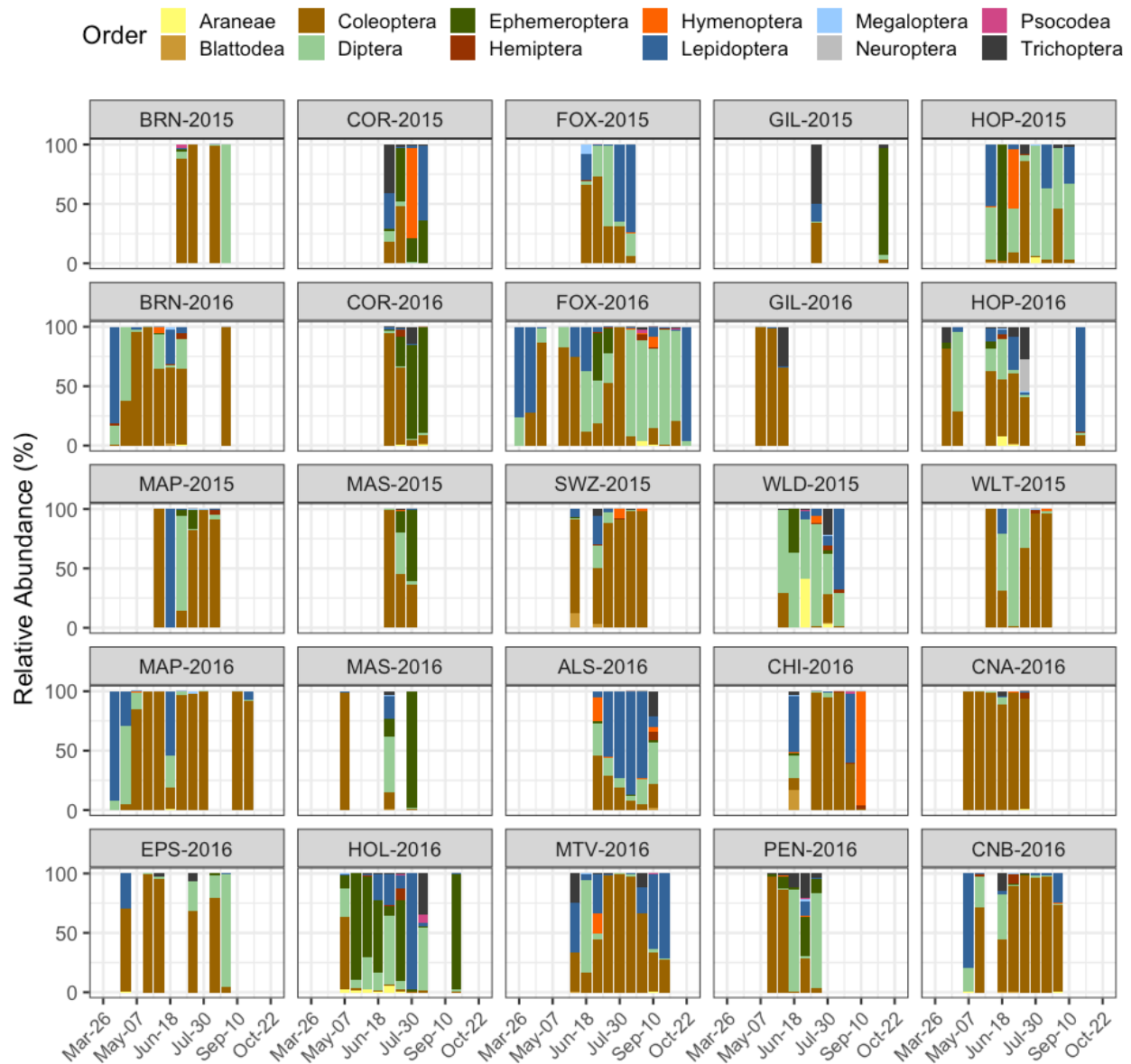


**Figure S25.** Distributions of sequence abundances per select ASVs. Among the top 5 most prevalent ASVs in each order (minimum 10 samples), the proportion of sequences that ASV comprises in a single sample has a broad distribution from negligible to nearly representing the entirety of the sample. These data suggest that primer bias is insufficient to explain why some orders are more prevalent in our data than others. Rather, a complex combination of biomass, digestion rate, and prey preference are all likely to explain why some orders are more frequently detected and generate larger fractions of sequence counts. Note that rarefied data produces similar distributions (not shown).

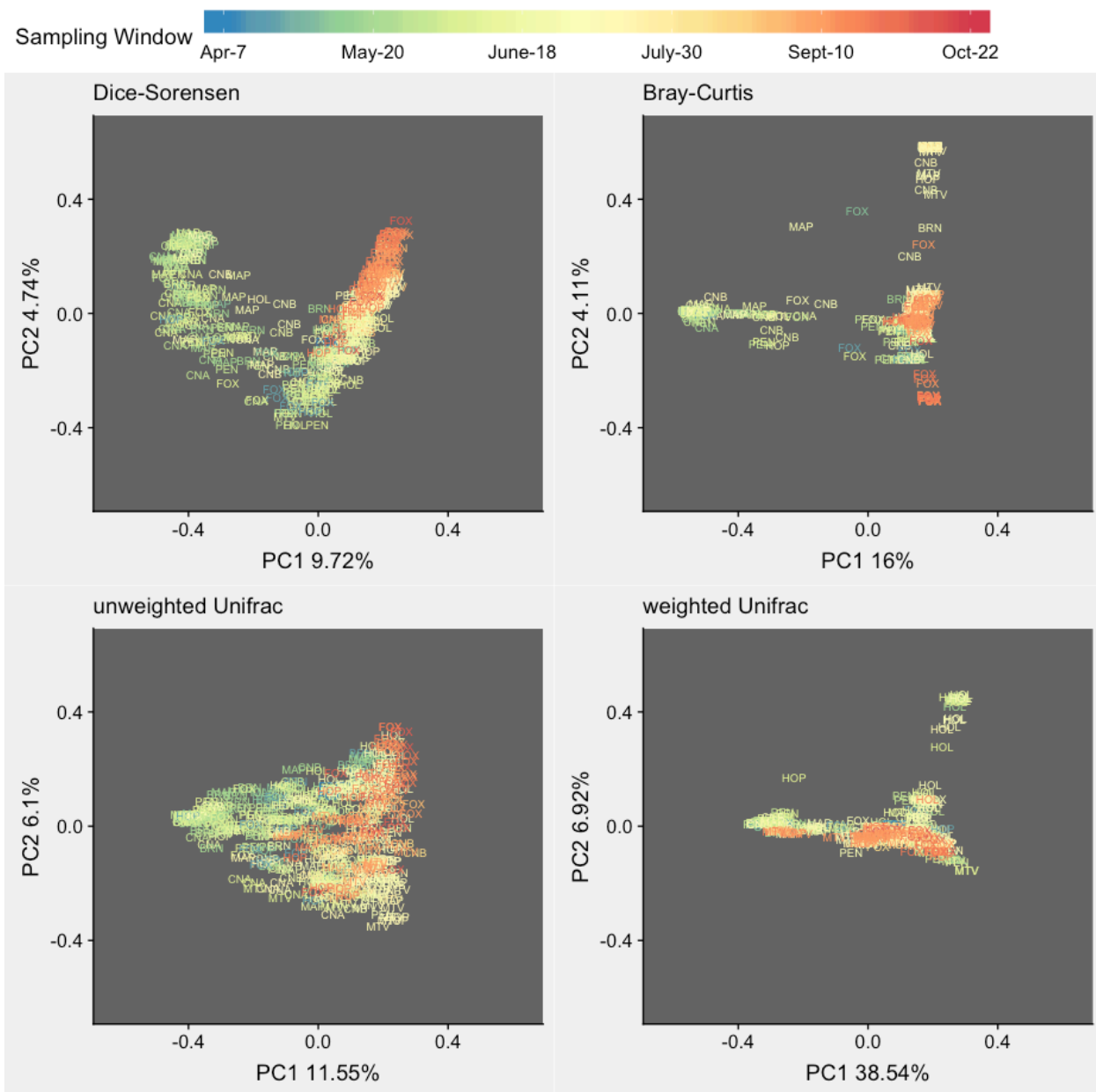


**Figure 26.** Detection frequency of ASVs in both or single year per site. Among sites surveyed in both years, the most prevalent ASVs are detected in both years (green points, “shared”), while ASVs detected only in one year or the other (distinct) are detected in no more than four samples in that one year. Many ASVs detected in both years are Coleopteran (e.g. ASV’s 1 and 5 are present in each of the six locations both years), although some non-beetles are also repeatedly observed at particular sites with varying detection frequencies. For example, trichopteran ASVs classified as *Phryganea sayi* were identified in 18 samples in one location (HOP), 7 samples at another (MTV), but only in three or fewer samples at six other locations





**Figure 27.** Relative abundances of arthropod COI per site per week per year. Sampling windows (vertical bars) are grouped in 14-day intervals; facet labels indicate sampling site and year. Ordinal proportions are more extreme with respect to proportions of individual taxa than what is observed in a presence-absence detection context (see **Figure 2**), but similar patterns of largely beetle, moth, and dipteran mixtures with distinctly local compositional changes are observed with both detection-based or abundance-based approaches.



**Figure S28.** Principal correspondence analysis of dietary composition for select 2016 New Hampshire locations using weighted and unweighted dissimilarity measures. Addition of abundance (Bray-Curtis) or phylogenetic (unweighted Unifrac) information increase the proportion of variation explained in the first two principal components compared to Dice-Sorensen (unweighted for both abundance and phylogenetic information). Combining both components (weighted Unifrac) captures the greatest proportion of variation, particularly for the first principal component axis (PC1). Samples in each ordination suggest that variation in composition is least prevalent in mid summer, while differences among samples are most apparent in spring or fall.