Resolving gut microbiome networks within Chiropterans

Timothy J. Rogers, Richard A. White, Laurel R. Yohe†

To whom correspondence should be addressed

# Abstract

# Introduction

* Bats are known carriers of human associated pathogens
* The reason bats are bioreactors is not understood
* The diet of bats may contribute to the gut microbiota makeup
* Phage associated with these microbiota can benifit and hinder microbial populations and have an impack on bat immune responses
* Some ideas suggest viral tolerance is linked to
  + Uniqueness of bats and their variation in Diets
    - Diversity
    - Ecological role
    - Pathogenic role
* Methods for characterizing microbiome and virome. As well as methods for linking the two
  + Culture dependant vs independent
    - What has been discovered with these methods
    - How these methods have been applied to bats
      * What has been found
      * What has yet to be discribed
* Goal of this study

**Figure X. Bat Holobiont** # Materials and Methods \* Sample location discription \* Sample collection \* Sample processing ### Phase genomics portion \* (AND)

## Sample collection

### ProxiMeta Hi-C and metagenomic pipeline

Guano samples were then used to generate Hi-C libraries using the protocol included in the ProxiMeta Hi-C kit (**Phase Genomics**). Hi-C operates by chemically cross-linking DNA fragments that lie in close spatial proximity inside intact cells, allowing for active phage infection identification. High quality Hi-C libraries were produced for **X** guano samples from **X** bat species. **X** libraries were successfully extracted and sequenced from bat species **X**, **Y**, and **Z**. Briefly, extracted genomic DNA along with chemically fixed fecal samples were sent to Phase Genomics (Seattle, WA, USA) for Hi-C proximity ligation and sequencing. Hi-C libraries were constructed using the Phase Genomics ProxiMeta Hi-C v4.0 Kit, while metagenomic shotgun libraries were generated in parallel with ProxiMeta’s standard library preparation reagents. Both library types were sequenced on an Illumina NovaSeq instrument to produce 150 bp paired-end reads. (**Need to add in here the methods that ProxiMeta uses to construct the MAGs**). The resulting **X** million Hi-C and **X** million shotgun metagenomic paired-end read fastq files were then uploaded to the Phase Genomics cloud portal for downstream analysis.

## Microbial Community Profiling

SingleM (vXXX; **CITATION**) “pipe” was used to extract metagenome operational taxonomic units (mOTUs) from 59 single copy conserved genes (22 Bacteria-specific, 24 Archaea-specific, and another 13 targeting both domains) identified in the raw reads and create an OTU table. Then, SingleM “pipe” was ran on the dereplicated metagenome assembled genomes (MAGs; see below) as well as the unbinned contigs. The OTU tables from the raw reads, MAGs, and unbinned contigs were then compared using SingleM “appraise” using a sequence identity of 89% to determine how much of the metagenomes were represented in the assemblies and MAGs at the genus level.

### MAG curration and taxonomic identification

The ProxiMeta produced a total of 441 medium to high quality MAGs ( 50% complete with 10% redundancy) across all bat species. MAGs were retrieved from the Phase Genomics cloud portal and stored on the University of North Carolina at Charlotte’s high-performance computing (HPC) cluster. To allow for comparision of MAGs across samples, dRep (**CITATION**) was used to dereplicate the MAGs at 98% ANI using the command “dRep cluster –SkipMash –S\_algorithm fastANI -comp 50”. This resulted in a final MAG set of 239 medium to high quality dereplicated MAGs. MAG quality was determined by CheckM2 (**v x.x.x**). MAG taxonomy was assigned using the Genome Taxonomy Database Toolkit (GTDB-Tk(**CITATION**)). GTDB nomenclature is used throughout the manuscript. However, other commonly used nomenclatures are used when they aid in clarity.

### vOTU curration

Viral contigs (vContigs) identified by ProxiMeta were retrieved from the Phase Genomics cloud portal. To identify potential vContigs not identified by ProxiMeta, genomade (**V**) was used to scan the assembly files. Anvi’o (**V**) anvi-script-reformate was using to clean up all identied vContig headers for downstream analysis. From there, all vContigs were combined into a single fasta file. They were then clustered into viral operational taxonomic units (vOTUs) at 95% average nucleotide identity (ANI) over 85% of the alignment fraction of the shorter sequence using the greedy centroid algorithm (anicalc.py and aniclust.py) from CheckV (v x.x.x). vOTUs were then ran through CheckV for quality scores and lifestyle predictions. In addition to CheckV, BACPHLIP (**V**) was also used in addition to finer scale resolution of viral lifestyle of the vOTUs. Viral lifestyle (lytic or lysogenic) was determined by integrating predictions from CheckV and BACPHLIP. The CheckV “provirus” annotation was considered, where a value of “Yes” indicated the presence of a provirus and suggested a lysogenic lifestyle. BACPHLIP outputs included probabilities for “Virulent” and “Temperate” lifestyles, each ranging from 0 to 1. A vOTU was classified as lysogenic if CheckV annotated it as a provirus (“Yes”), or if the BACPHLIP “Temperate” probability was ≥0.5. A contig was classified as lytic if CheckV did not detect a provirus (“No”) and the BACPHLIP “Virulent” probability was ≥0.75, or ≥0.5 if the genome was estimated to be 100% complete. If neither tool confidently predicted the lifestyle, the contig was labeled as “unclear.” The method used for each final prediction was also recorded, indicating whether the lifestyle assignment was based on CheckV, BACPHLIP, or a combination of both.

### vOTU and MAG coverage

Quantification of dereplicatied MAGs and vOTUs across all samples was calculated using the anvi’o tool. Briefly, quality controlled reads were mapped to the vOTUs and representative MAGs using bowtie (v x.x.x) and sam files were converted to indexed and sorted bam files using samtools (v x.x.x) (**CITATION**). The resulting BAM files for the MAGs and vOTUs were then used by the anvi’o tool kit to calculate the depth of coverage across nucleotide positions in which coverage was within the interquartile range (Q2Q3). Q2Q3 coverage was used to filter out outliers in coverage caused highly similar sequences shared among different taxa. MAG abundance was quantified in units of weight mean transcripts per million reads (wtTPM) for contigs within each MAG. First, we calculated the transcripts per million reads (TPM) for each contig at each site (**equation 1**).

**Equation 1:**

In this equation, denotes the Transcripts Per Million value for contig , representing the normalized abundance of that contig after adjusting for both contig length and total sequencing depth. The variable refers to the number of sequencing reads mapped to contig , while is the length of contig in base pairs. We then multiply by one million () to convert the values into TPM units.To then quantify MAG abundance across samples, we then calculated the wtTPM using the below equation 2

**Equation 2:**

In this expression, denotes the length-weighted mean TPM for the total contigs within metagenome-assembled genome (MAG) . vOTU abundance was quantified as transcripts per million reads (TPM; Equation 1).

### vOTU and MAG metabolic predictions

Metacerberus (**CITATION**) was used for open reading frame prediction, gene annotation, KEGG orthology identifier linkage, and metabolic predictions using the Functional Ontology Assignments for Metagenomes (FOAM; (**CITATION**)), KEGG (**CITATION**), CAZy/dbCAN (**CITATION**), VOG (**CITATION**), pVOG (**CITATION**), PHROG (**CITATION**), and COG (**CITATION**) databases. An in house script was usied to identify AMR and CAZY genes as well as calculate their coverage/abundance.

### Linking virus to host

In addition to the host associations predicted through proximity ligation by ProxiMeta, iPHoP (Integrated Phage HOst Prediction; **Roux et al. 2023**) was also used to identify additional potential host associations. iPHoP uses six host prediction methods to determine phage-host interactions (typically at a genus or family rank): Blast, CRISPR spacer matching (**CITATION**), three different oligonucleotide frequency (ONF) comparison methods (VirHostMatcher (**CITATION**), WIsH (**CITATION**), and PHP (**CITATION**)), and Random Forest Assignment of Hosts (RaFAH; (**CITATION**)). First, iPHoP was ran with default parameters to predict host of a vOTUs based on iPHoP’s default database (**DATA BASE VERSION**). Then, the original 441 MAGs from ProxiMeta were added to the default database using (using the “add\_to\_db” function) to create a custom database. This custom database was used to run a final host prediction on the same vOTUs. Finally, phage-host pairings were filtered using a minimum cutoff pairing score of 75. Only host paring predictions to MAGs were considered for down stream analysis. If a vOTU was paired to a MAG that was not a representative MAG via proximity ligation or one of iPHoPs methods, then that pairing was given to the representative MAG for downstream analysis.

### Statistical analysis

Results were imported into R for analysis and visualization using the R packages data.table (**CITATION**), tidyverse (**CITATION**), ggplot (**CITATION**), vegan (**CITATION**), hillR (**CITATION**), phyloseq (**CITATION**), ggtext (**CITATION**), and ggpubr (**CITATION**). mOTUs identified by SingleM were used to assess alpha and beta diversity. Differences in prokaryotic community composition among bat species were analyzed using distance-based redundancy analysis (db-RDA) on a Hellinger distance matrix. Statistical significance was evaluated with 9,999 permutations using the R package vegan (**V**) (**Oksanen et al. 2015**). Richness (Chao1), evenness (Pielou’s index), and alpha diversity (Hill numbers at q = 1, corresponding to the inverse Shannon index) were calculated. Comparisons of alpha diversity across bat species were performed using two-way ANOVA. To identify microbial mOTUs associated with each tree species, we performed an indicator species analysis following the approach of **De Cáceres and Legendre (2009)** and **De Cáceres et al. (2010)**. Group-equalized correlation coefficients (r₉) were calculated for each bat species to account for differences in taxon abundance and enable direct comparisons among taxa with varying relative abundances. Statistical significance was assessed through 9,999 permutations. All analyses were carried out using the R package indicspecies (**V**) (**De Cáceres et al. 2016**). Subsequently, the relative abundances of indicator mOTUs were calculated for each tree species.

# Results

## SingleM analysis

Bat gut microbiomes were profiled based on metagenome operational taxonomic units (mOTU) using 59 single copy conserved genes identified by SingleM. After rarefaction, and removal of Eukaryota sequences, a total of 32233 mOTUs were retained. tb-RDA on the Hellinger distance matrix revealed significant differences in the gut prokaryotic community composition across the three bat species (**Fig. 1**, *adj. R2*=0.27, *F*=2.3, *P* < 0.01). However, pairwise comparisons were unable to detect differences across the bat species.

![Figure 1. Transformation based redundancy analysis of the bat gut prokaryotic community mOTUs.](data:application/pdf;base64,)

**Figure 1. Transformation based redundancy analysis of the bat gut prokaryotic community mOTUs.**

### Prokaryotic community differences across the gut of bat species

Across bat species, 34% of the mOTUs belonged to *Actinomycetota*, 19% belonged to *Pseudomonadota*, 18% belonged to *Bacillota*, 7% belonged to *Bacillota\_A*, 2% belonged to *Campylobacterota*, 2% belonged to *Patescibacteria*, 2% belonged to *Bacteroidota*, 2% belonged to *Planctomycetota*, 2% belonged to *Acidobacteriota*, 2% belonged to *Desulfobacterota*, 2% belonged to *Chloroflexota* and 1% belonged to *Thermoproteota*. The remaining phylum constituted < 1% each. Prokaryotic -diversity (inverse-Shannon) did not significantly differ across bat species. However, Prokaryotic community evenness was significantly lower in *Phyllostomus discolor* than in *P. hastus* (**Fig. 2C**, P<0.05). Even so, no significant difference was found in the prokaryotic richness (Chao1). A total of 0Indicator species analysis, based on rg, identified specific mOTUs common in some bat species, but in lower relative abundance in others (**Table SX**).

For example, members of the bacteria pylum *Bacillota*, *Loactococcus lactis* and unidentified taxa in the genera *Loactococcus* and *Streptococcus*, had high relative abundances in *Mimon crenulatum* (10%, 10%, and 6% respectively) while they were not found at all in *P. discolor* nor *P. hastus*. Unidentified taxa in the bacteria genus *Arachnia* and another in the domain *Archean* were abundant in *P. discolor* (.03% and .01% respectively) but were not present in *P. hastus* nor *M. crenulatum*. Conversely, members of the bacteria *Bacillota*, an unidentified taxa in the genus *Gemella* and the species *Gemella sanguinis*, had higher abundances in *P. hastus* (35%, and 3% respectively) when compared to *M. crenulatum* or *P. discolor*.

![Figure 2. Bat gut prokaryotic community diversity of the mOTUs.](data:application/pdf;base64,)

**Figure 2. Bat gut prokaryotic community diversity of the mOTUs.**

## Metagenomic analysis

### MAG diversity and community representation

Assembly, binning, and dereplication at the species level resulted in the recovered of 5 Archaeal and 234 Bacterial medium to high quality MAGs with 50 % and 10 % redundancy. Of these 239 MAGs, 135 were 90% complete and 10% redundant, while **X** meet the MIMAG standard of high quality draft genomes ((**CITATION**), **Table**). **X%** (n=**X**) of the MAGs were prepredicted to be novel by GTDB-Tk (**what tax levels**).

These MAGs represented between 16% and 78% of the whole community, with a median of 61.4% based on the genus-level recovery estimentes ((**Singlem Citation**), **TABLE X**). *Bacillota*, *Actinomycetota*, and *Pseudomonadata* were the predominate phylum across the gut microbiome MAGs within all 3 of the bat species (**Fig 3**).

![Figure 3. Top 10 most abundant MAG phylum from across the bat gut metagenomes](data:application/pdf;base64,)

**Figure 3. Top 10 most abundant MAG phylum from across the bat gut metagenomes**

### vOTUs

A total of **17702** potential viral sequences were ID from the 6 individual assemblies. Clustering at 95% ANI over 85% of the shortest sequence identified 16289 viral operational taxanomic units (vOTUs) that were 1 kbp in length. Further filtering for sequences 2.5 kpb resulted in a final set of 6235 vOTUs. After rarefaction, 5582 vOTUs were retained. CheckV was used to assess the quality of these sequences, revealing that 361 (6%) were 50% complete including 0 complete vOTUs that identified on the bases of direct terminal repeats (DTR), 0 high quality vOTUs that were identified on the bases of AAI (0 vOTUs) and HMM (0 vOTUs), 0 medium quality vOTUs that were idenified on the bases on AAI (0 vOTUs) and HMM (0 vOTUs), and 0 low quality vOTUs that were identified based on AAI (0 vOTUs) and HMM (0 vOTUs). The reset of the vOTUs (5221) were of low quality (3298) or the quality was undetermined (1923). An unclassified order of the class *Caudoviricetes*, families *Retroviridae*, *Adintoviridae*, and *Iridoviridae*, an unclassified family of *Kyanoviridae*, families *Inoviridae* and *Bornaviridae*, an unclassified family of *Herelleviridae*, families *Mimiviridae* and *Parvoviridae* were the most predominate viral taxa across the gut virome with all 3 of the bat species (**Figure 4**). No statistically significant differences were found in the viral richness, envenness, -diversity, or -diversity nor did indicator analysis reveal any indicator viral spices for the bat species.

![Figure 4. Top 10 most abundant viral taxa from across the bat gut metagenomes](data:application/pdf;base64,)

**Figure 4. Top 10 most abundant viral taxa from across the bat gut metagenomes**

### Virus-host predicitons

Phage-host detected by Hi-C were compared to those made by the bioinformatics pipeline iPHoP. Using both Hi-C and iPHoP methods, a total of 1299 phage-host predictions were made using phase HiC and IPHoP methods. These predictions included 78% (n= 186) of the total MAGs and \*\*x/y\*100\*\*% (n=953) of the total vOTUs.

Hi-C analysis allowed for the identification of 352 unique phage-host links comprised of 130 MAGs and 349 vOTUs. Using the six methods employed by iPHoP, a total of 875 unique phage-host links, comprised of 137 MAGs and 581 vOTUs, were identified. We also identified 81 unique phage-host links that were identified by both methods (**Figure 5**). These were comprised of 38 MAGs and 80 vOTUs. Of those predictions made specifically by IPHoP, 510 are based on blastn and 365 are based on a combination of the 6 methods that iPHoP uses (called iPHoP-RF).

![Figure 5. Venn diagram of phage-host predictions made with the Hi-C and iPHoP methods](data:application/pdf;base64,)

**Figure 5. Venn diagram of phage-host predictions made with the Hi-C and iPHoP methods**

Of the 352 phage-host predictions made by phase, 18% of the pairings were with lysogenic phage, 75% were with lytic phage, and 7% were with phages whose lifestyle was not able to be predicted. Of the 875 phage-host predictions made by iphop, 32% of the pairings were with lysogenic phage, 65% were with lytic phage, and 3% were with phages whose lifestyle was not able to be predicted. Of the 81 phage-host predictions made by both phase and iphop, 32% of the pairings were with lysogenic phage, 63% were with lytic phage, and 5% were with phages whose lifestyle was not able to be predicted (**Fig. 6**).

MAGs within the bacteria phyla *Actinomycetota*, *Bacillota*, *Bacillota\_A*, *Pseudomonadota* and *Desulfobacterota* had the highest frequency of being targeted by viruses (84, 43, 19, 18 and 13 MAGs respectively). Host families most often targeted within *Actinomycetota* include *Mycobacteriaceae*, *Nocardioidaceae*, *Dermatophilaceae*, *Propionibacteriaceae*, *Actinomycetaceae* and *JAAYBP01* (20, 10, 9, 8, 6 and 5 MAGs respectively). The host family most often targeted within *Bacillota* is *Enterococcaceae*, *Streptococcaceae*, *Staphylococcaceae*, *Gemellaceae* and *Lactobacillaceae* (15, 9, 6, 5 and 5 MAGs respectively). The host family most often targeted within *Bacillota\_A* is *Ruminococcaceae* (4 MAGs). The host family most often targeted within *Pseudomonadota* is *Enterobacteriaceae* (6 MAGs). The host family most often targeted within *Desulfobacterota* is *Desulfovibrionaceae* (9 MAGs).

Based on the singleM genus-level recovery estimates, we were able to determine that 59% (n=297) of the 507 mOTU indicator species are represented by 16% (n=29) of the 186 MAGs that were paired to phages (Fig. 6). 18 of these MAGs belong to the phylum Bacillota (in the genre Streptococcus, Gemella, Enterococcus, Enterococcus\_A, Abiotrophia, Lactococcus, Lentilactobacillus and Staphylococcus (5, 4, 3, 2, 1, 1, 1 and 1 respectively)), 7 belong to the phylum Pseudomonadota (in the genre Aggregatibacter, HY038, Lautropia, Morganella and Scandinavium (2, 2, 1, 1 and 1 respectively)), and 4 belong to the phylum Actinomycetota (in the genre Corynebacterium, Actinomyces and Streptomyces (2, 1 and 1 respectively)).

20 MAGs were indicators for *P. hastatus*, 8 were indicators for *M. crenulatum*, and 1 was an indicator for both *P. hastatus* and *M. crenulatum*.

These 29 indicator MAGs were paired with 40% (n=382) of the 953 phage paired to host MAGs. 20 phase-host prediction involving indicator MAGs were made with Hi-C, 28 were made by IPHoP, and 12 were made by both the Hi-C and IPHoP methods. 246 of vOTUs matched to indicator MAGs were predicted to be lytic, 115 were lysogenic, while the lifestyle of 21 was unable to be predicted.

![Figure 6. Phage-host network analysis.](data:application/pdf;base64,)

**Figure 6. Phage-host network analysis.**

# Discussion

Our novel study is the first to offer empirical evidence of phage-host interactions in the gut microbiome of bats using metagenomics in conjunction with Hi-C techniques. (**Talk about previous attempts of phage-host studies in bat guts and the methods they used**).

(**Need an introduction to this paragraph here that leads to the hypothesis in the following sentence**). Therefore, we hypothesized that the bat gut virome acts as a phage-mediated immune system against pathogenic prokaryotes. In support of this hypothesis, those prokaryotic families that had more frequent pairings with phage were those that contained known pathogenic members. For example, the families *Mycobacteriaceae*, *Nocardioidaceae*, *Dermatophilaceae*, *Propionibacteriaceae*, *Actinomycetaceae* and *JAAYBP01* within the bacteria phylum *Actinomycetota*, the family *Enterococcaceae*, *Streptococcaceae*, *Staphylococcaceae*, *Gemellaceae* and *Lactobacillaceae* within the bacteria phylum *Bacillota*, the and family *Ruminococcaceae* within the bacteria phylum *Bacillota\_A* all contain members of known human pathogens (**Citations**).

While the family *Enterobacteriaceae* within the bacteria phylum *Pseudomonadota* contain no known pathogenic members, members are associated with the human gut (**Citation**).

Macro Eukarotic \* (AND) *Iridoviridae* \* (AND) *Bornaviridae* \* (AND) *Paroviruses* Micro Eukarotic \* (AND) *Mimiviridae* Bacterial phage \* (AND) an unclassified family of *Kyanoviridae* \* (AND) *Inoviridae* \* (AND) an unclassified family of *Herelleviridae*

# Conclusion

# Acknowledgement