

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Virus-Host Relationships

Permalink

<https://escholarship.org/uc/item/0vs9101z>

Author

Martyn, Calla

Publication Date

2022

Peer reviewed|Thesis/dissertation

Virus-Host Relationships

by
Calla Martyn

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY


in

Biological and Medical Informatics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

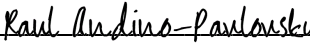
DocuSigned by:

9B0DDB91029D4F1... Katherine Pollard
Chair

DocuSigned by:

DocuSigned by:4B3... Seemay Chou

DocuSigned by:

DocuSigned by:34FD... Vasileios Ntranos

DocuSigned by:

14C20405A1FD407... Raul Andino-Pavlovsky

Committee Members

Dedication and Acknowledgement

I would like to thank my advisors Seemay Chou and Katie Pollard for their support, encouragement and mentorship over the last 4 years. Their help and advice on both my project and the scientific enterprise at large have been crucial to my success. I would also like to thank Amy Kistler, whose help was instrumental to this project and whose mentorship has been invaluable, and the Biohub at large for facilitating this collaboration. I am indebted to my committee members Raul Andino and Vasilis Ntranos for their valuable input as well as their interest and enthusiasm for my work, which reinvigorated my own excitement for science.

I would like to thank all the members of the Chou lab for their support and enthusiasm and for making the lab an overwhelmingly welcoming and positive place to work. In particular I would like to thank Beth Hayes for her willingness to tackle any scientific problem and Anne Sapiro for her supportive and steady mentorship. I would also like to thank the members of the Pollard lab for their focused attention and commitment to producing rigorous science, as well as their excellent company both in and out of the lab.

My parents and two sisters have been incredibly supportive during this time, and I thank them for their kind words and actions, for repeatedly convincing me that I could accomplish this, as well as for always picking up the tab during these lean years. I cannot emphasize enough the role my friends, both within UCSF and without, played in my success and mental health. They provided perspective when I was spiraling, and provided me with endless good conversation and entertainment over the years. Also critical to my mental health and instrumental to my success are my two cats Peter and Henry.

Contributions

Basis for Chapter 1:

Calla Martyn, Beth M. Hayes, Domokos Lauko, Edward Midthun, Gloria Castañeda, Angela Bosco-Lauth, Daniel J. Salkeld, Amy Kistler*, Katherine S. Pollard*, Seemay Chou*. “mNGS Investigation of Single Ixodes pacificus Ticks Reveals Diverse Microbes, Viruses, and a Novel mRNA-like Endogenous Viral Elements”. *bioRxiv*, August 17, 2022, 504163. <https://doi.org/10.1101/2022.08.17.504163>

*Co-corresponding authors

Abstract

Virus-Host Relationships

Calla Martyn

Ticks are increasingly important vectors of human and agricultural diseases. While many studies have focused on tick-borne bacteria, far less is known about tick-associated viruses and their roles in public health or tick physiology. To address this, I investigated patterns of bacterial and viral communities across two field populations of western black-legged ticks (*Ixodes pacificus*). In addition to commonly found tick-associated microbes, I discovered 11 novel RNA viruses from *Rhabdoviridae*, *Chuviridae*, *Picornaviridae*, *Phenuiviridae*, *Reoviridae*, *Solemoviridae*, *Narnaviridae* and 2 highly divergent RNA viruses lacking sequence similarity to any known viral families. I also unexpectedly identified numerous virus-like transcripts that are likely encoded by tick genomic DNA, and which are distinct from known endogenous viral element-mediated immunity pathways in invertebrates. Together, my work reveals that *I. pacificus* ticks carry a greater diversity of viruses than previously appreciated, in some cases resulting in evolutionarily acquired virus-like transcripts. These findings highlight how pervasive and intimate tick–virus interactions are, with major implications for both the fundamental biology and vectorial capacity of *I. pacificus* ticks. I furthermore investigated whether viral dinucleotide composition is shaped more by the host a virus infects or its phylogenetic background. I found phylogeny to be the stronger driver and identified a common source of overfitting similar models that may lead to inflated measures of accuracy.

Table of Contents

Chapter 1: Introduction	1
Chapter 2: mNGS Investigation of Single <i>Ixodes pacificus</i> Ticks	8
Chapter 3: Factors Shaping Viral Genome Composition	32
Chapter 4: Conclusion	40
References	43

List of Figures

Figure 2.1: Experimental Approach	10
Figure 2.2: Bacterial genera detected in <i>Ixodes pacificus</i>	11
Figure 2.3: Co-occurrence of tick microbes	13
Figure 2.4: Viral discovery pipeline	14
Figure 2.5: Discovery of novel viruses in <i>Ixodes pacificus</i>	16
Figure 2.6: Virus-like transcripts detected across <i>Ixodes pacificus</i> population	19
Figure 2.7: Confirmation of VLT sequence and presence in DNA	20
Figure 3.1: Unsupervised analysis of viral dinucleotide bias	33
Figure 3.2: Effects of Sequence Length on Prediction Accuracy	34
Figure 3.3: Classifier performance	37
Figure 3.4: Associations with Viral Family AUC	38

List of Abbreviations

AUC Area under the curve

DASH Depletion of abundant sequences by hybridization

EVE Endogenous viral element

LDA Linear discriminant analysis

ORF Open reading frame

PCA Principal components analysis

RdRp RNA dependent RNA polymerase

VLT Virus-like transcript

Chapter 1: Introduction

1.1 Tick metatranscriptomics

Ticks are increasingly important disease vectors for humans and livestock, particularly in the United States, where they account for more cases of vector-borne diseases than do mosquitoes. Approximately fifty thousand confirmed cases of tick-borne diseases are reported annually¹, which is likely an underestimate due to diagnostic challenges associated with Lyme disease, the most common tick-borne disease and our poor understanding of rare tick-borne diseases or diseases of unknown etiology. Currently, the majority of tick-borne microbe research focuses on the causative agent of Lyme disease *Borrelia burgdorferi* and a select number of other known human pathogens, such as *Rickettsia* (Rocky Mountain spotted fever and other Rickettsioses), *Anaplasma phagocytophilum* (Anaplasmosis), and *Babesia microti* (Babesiosis). Although the full diversity of microbes carried by ticks is much greater than those definitively linked to human disease²⁻²¹, we know strikingly little about the ecology or disease implications of most tick-associated microbes. This knowledge gap is particularly large when it comes to tick-associated viruses, known as the tick “virome”. My thesis aims to address this gap.

Historically, vector-microbe relationships have been identified through human disease. When a disease is found to have an association with vectors, the causative microbe is isolated and characterized in the lab. Studying these microbes in the lab is challenging as it often requires co-culture with either arthropod cell lines or vertebrate host factors such as blood. Some microbes cannot be cultured outside an organism at all and must be studied by maintaining populations of the relevant arthropod. In addition to the technical challenges, this type of study has been limited to a small subset of microbes. Until recently, only those microbes that are pathogenic to humans have been studied. However, there has been increasing interest in the microbiome of arthropods, that is the full community of bacteria, viruses, and protists that exist within

an organism, typically in the gut. The field of metagenomics has greatly facilitated the study of such communities. By computationally annotating all the genomes that exist in a community, as well as their relative abundance, and gene content, researchers can learn about the relationships between the host and microbes.

Metagenomics is typically performed on DNA sequencing libraries. This enables assembly of complete bacterial genomes and can capture all genes present regardless of gene expression levels. However, this approach has drawbacks for the study of viruses as it cannot detect RNA viruses, which are a major constituent of arthropod viromes. Therefore metatranscriptomics, sequencing and analyzing all the RNA in sample, is a critical approach for understanding both the bacteria and viruses in an arthropod microbiome.

There have been an increasing number of metagenomic and metatranscriptomic evaluations of arthropods in recent years. While commonly used microbiome tools such as QIIME²² or kraken²³ can be used to profile the bacterial component of arthropod microbiomes, they identify very few viral sequences, as viruses are poorly represented in reference databases. Therefore, arthropod microbiome studies typically use sequence homology searches, such as blastx²⁴ in order to identify reads of viral origin. More recently, there has been a shift to the use of hidden Markov model based tools such as HMMER²⁵, as they are able to identify divergent sequences that cannot be detected with pairwise alignment methods like blastx.

These investigations have revealed that arthropods primarily carry RNA viruses from a diversity of families such as *Rhabdoviridae*, *Flaviviridae*, *Chuviridae*, *Phenuiviridae*, *Reoviridae*, *Orthomyxoviridae* and many others. They have also helped to establish that arthropods are able to tolerate viral infections, often throughout their life, with little to no ill effect. Such studies have also led to the discovery of new viral families, the identification of human pathogens, and a deeper understanding of viral spread and segment reassortment over time and space²⁶.

However, several challenges make improvement in this field difficult. Foremost among them is the presence of abundant sequences of tick origin in the sequencing libraries. This is a common challenge in metagenomics but is exacerbated in arthropod metagenomics where nucleic acid is extracted from the entire body, as opposed to a microbial-enriched sample such as human stool. RNA libraries are even more affected as host ribosomal RNA is expressed at very high levels, dominating the RNA library and compressing the microbial sequences to less than 1% of the total library. As a result, only the highest abundance organisms can be detected.

While there are tools for depleting ribosomal RNA in libraries originating from human or murine samples, such tools do not exist for ticks or other non-model organisms. However the recent development of a CRSPR-Cas9 system, Depletion of Abundant Sequences by Hybridization (DASH)²⁷ has enabled the depletion of any sequences utilizing custom guides. This has been successfully employed in mosquito metatranscriptomics, and its extension to ticks can help further the field by increasing the effective sequencing depth and therefore profiling a greater proportion of the microbial community.

Due to the small amount of nucleic acid from an individual tick, many studies opt to pool dozens or hundreds of individuals together in libraries. While this increases the size of the population that can be surveyed it precludes other important metrics. It is not possible to determine the individual to individual variation using such methods, only a population average, which could be strongly affected by outlier individuals. This impacts both measures of relative abundance within a library as well as prevalence of individual microbes across a population.

As many of these studies identify novel taxa, and since the common players in the tick microbiome are not firmly established, it can be difficult to differentiate between those microbes that are truly tick-associated and those that may result from environmental contamination, particularly as entire

organisms are typically used. The use of control libraries generated from water, or from other biological material, while not common, can greatly aid in this distinction.

With these emerging techniques in hand, scientists are rapidly gaining understanding regarding the bacteria and viruses in tick microbiomes. We now know that ticks have a bacterial microbiome composed of diverse species from phyla such as *Proteobacteria*, *Spirochaetes*, *Tenericutes*, and *Firmicutes*. Ticks are also able to carry viruses from a broad range of families, some of which have been identified as causative agents of disease in humans, such as Tickborne Encephalitis Virus, Powassan Virus, and Severe Fever with Thrombocytopenia Syndrome Virus²⁸.

Along with these taxonomic profiles of the tick microbiome, we are beginning to appreciate the roles that microbes play in tick biology. Like many other invertebrates, tick–microbe interactions go far beyond the transmission of human pathogens. Tick–microbe interactions can be antagonistic, neutral, or beneficial^{29–32}, and have been implicated in shaping the evolution of ticks through horizontal transfer of bacterial genes³³ and endogenization of viral sequences as an immune response^{34–37}.

In order to fully understand the roles microbes play in tick physiology and human disease, it is important to know whether each bacterial or viral species is stably associated with ticks versus briefly acquired from the environment, including whether or not the species can be vertically transmitted within tick populations. Some bacteria, such as the endosymbiont *Rickettsia*, play fundamental roles in tick physiology through stable and symbiotic interactions³². It is not as well known whether viruses also form stable, symbiotic interactions with ticks and how such interactions may also influence tick biology. Some viruses have been identified not only in live ticks but also in many laboratory-passaged tick cell lines³⁸, indicating a stable association.

While field microbiome studies using metatranscriptomics have increased our catalog of tick-associated viruses²⁸, there still remain several outstanding questions. Predominant among them is whether there exist

any combinations of microbes that facilitate or exclude infection by the other. Such interactions have been identified in mosquitoes and subsequently leveraged to control human pathogens. To date no such interactions have been identified in ticks, hampered in part by the lack of metagenomic studies with individual tick resolution. Furthermore, little is known about how the native virome interacts with the tick host, particularly whether infections are quickly controlled, or maintained for long periods of time.

1.2 Viral Genome Composition

Comparative genomics often focuses on sequence similarity and nonsynonymous mutations. However there is ample evidence that genome composition also varies widely across the tree of life. For example, nucleotides are not observed at equal frequencies, with some genomes having much higher G/C content than others³⁹. While there are multiple codons that encode for each amino acid, these do not appear at equal frequencies in most genomes, leading to codon usage bias that varies by organism⁴⁰⁻⁴². Codon pair bias and dinucleotide bias, the frequency with which two codons or two nucleotides are observed together compared to their individual frequencies, also varies by organism⁴³⁻⁴⁵.

These compositional biases can exert selective pressure on organisms. CpG suppression in vertebrates is thought to arise due to epigenetic transcriptional regulation via methylation of cytosines in the CpG conformation⁴⁶. In human cells, stress-response genes exhibit a different codon utilization than non stress-response genes⁴⁷. Our understanding of the factors driving genome composition is limited, and even less is known regarding how these factors are shaped in viral genomes.

Viruses, which rely on host machinery to replicate and which must evade detection from the host immune system, are under a unique set of selective pressures with regards to genome composition. They must compete for host transfer RNAs, are susceptible to mutation by the host, and can also be recognized by the host due to their distinct genomic features⁴⁸. Therefore, it has been hypothesized that viral genome

composition should mimic that of its host as this could lead to improved translation efficiency and evasion of host defenses.

Indeed, viruses do seem to mirror the CpG content of their hosts; a subset of mammalian-infecting viruses show a significant CpG depletion, but this suppression is not observed in viruses of invertebrates⁴⁹. Furthermore, there is some evidence that “deoptimizing” viral codon usage away from that of its host can lead to viral attenuation⁴⁸. However, the inverse; recoding viral genomes to use more of their host’s common codons, does not seem to enhance viral infectivity⁵⁰. And additional research has indicated that the effects of these recoding experiments may be explained by the inadvertent introduction of CpGs. Interestingly, researchers observed that over the course of the SARS-CoV2 pandemic, the codon bias of the virus actually moved further away from that of humans, despite the virus becoming more adapted to this host⁵¹. In short, there is conflicting evidence for viral mimicry of host genome composition.

Encouraged by results showing correlation between viral and host genome composition, researchers have attempted to predict various viral characteristics from features of viral genome sequences (e.g., dinucleotide frequencies). For example, a plethora of machine learning tools exist to predict the host of phage sequences, many of them based on genome composition features⁵²⁻⁵⁷. There has also been an increase in the development of similar tools for eukaryotic viruses⁵⁸⁻⁶⁰. However, some research indicates that viral phylogeny, rather than host, is a stronger driver of dinucleotide composition⁶¹. It is difficult to reconcile these results; if viral genome composition is shaped by viral phylogeny, how can tools correctly predict viral hosts from these features? One possibility is that the tools are not accurately disentangling the two variables. Host tropism is not evenly distributed over all viruses; phylogenetically similar viruses tend to infect similar hosts. Therefore the models may actually be using dinucleotide composition to approximate viral phylogeny, which in turn is predictive of viral host.

There has been a substantial improvement in viral discovery and viral databases since the most recent paper examining viral dinucleotide composition directly. Furthermore, all of the work done to date has utilized supervised machine learning models to examine this question. Here I leverage the increase in available data as well as both unsupervised and supervised techniques to examine whether viral phylogeny or viral host are stronger drivers of viral genome composition.

Chapter 2: mNGS Investigation of Single *Ixodes pacificus* Ticks

2.1 Motivation for investigation of metatranscriptome of individual *Ixodes pacificus* ticks

A number of experimental strategies and technical hurdles have limited the scope and depth of microbiome analyses in ticks. Most tick-borne viruses discovered to date have been identified in large pools of samples, preventing quantitative examination of microbial prevalence, and per-sample relative abundance. This has also precluded analysis of co-occurrence of microbes in most cases. These metrics could greatly enable more sophisticated analyses of transmission dynamics and ecology. Furthermore, our ability to capture lower abundance microbes is hampered by the dominance of tick host sequences in genomic and metagenomic libraries. This limits our understanding to the most abundant microbes in ticks, which does not necessarily coincide with all microbes that have important roles in disease or tick physiology.

Our understanding of tick microbiota in North America is currently biased towards species historically associated with human diseases, such as *Ixodes scapularis*, the primary vector for Lyme disease in the Eastern United States. In recent years, there has been an expansion of tick-borne disease cases on the West coast of the U.S. that have been attributed to other less-studied tick vector species. *Ixodes pacificus* is a major tick species extending from Northern Mexico to British Columbia⁶². *I. pacificus* ticks are most abundant in California, where they cover 96% of all counties⁶³ and are responsible for the majority of human tick bites⁶⁴. They are vectors for a variety of well-characterized human pathogens such as *B. burgdorferi*, *Borrelia miyamotoi*, *Babesia odocoilei*, *Bartonella spp*, *A. phagocytophilum*, and *Ehrlichia spp*⁶⁵. Despite this, *I. pacificus* is substantially understudied compared to the eastern black-legged tick *I. scapularis*. For these reasons, I chose to focus this study on *I. pacificus* ticks.

To provide much-needed insight into tick-borne microbes in the Western U.S., I examined the microbiomes of *I. pacificus* ticks collected from two coastal habitats in California where humans are likely to encounter them^{64,66,67}. In order to capture lower-abundance microbes, I coupled an experimental microbial enrichment workflow with RNA sequencing to profile both bacteria and RNA viruses. Analysis of microbiomes at the level of individual ticks enabled me to also quantify patterns of microbial prevalence and abundance. We performed follow up laboratory-controlled experiments examining microbial localization within tick compartments and across tick life stages, which provided additional insights into potential transmission dynamics and the symbiotic nature of tick-virus relationships.

2.2 Establishment of an RNA-based approach to defining composition of field tick microbiomes

I set out to define the metatranscriptome of *I. pacificus* ticks collected from coastal California, focusing on two sites associated with human exposure⁶⁴⁻⁶⁸. I examined the two most developmentally advanced life stages (nymphal and adult) that are more amenable to single-tick sequencing due to greater individual biomass. These sample sets included adult ticks from Garrapata State Park and nymphal ticks from China Camp State Park. Adults were collected in the Fall (2019) and nymphs were collected in the Spring (2020) so that I could investigate seasons and life stages enriched for human contact. In total, RNA libraries were sequenced for 100 individual ticks.

The majority of whole-tick RNA libraries are composed of tick ribosomal RNA, which reduces the power to detect lower abundance bacterial and viral sequences. To address this challenge, we enriched microbial sequences by experimentally depleting abundant tick sequences through Depletion of Abundant Sequences by Hybridization (DASH) (Figure 2.1a)²⁷. For adult tick libraries, DASH-based depletion enriched non-host reads by nearly ten-fold (Figure 2.1b). Sequencing libraries generated for smaller nymphal ticks were not of sufficient concentration to effectively perform DASH. After quality filtering and host subtraction, non-host reads were first classified using the metagenomic classifier kraken2²³

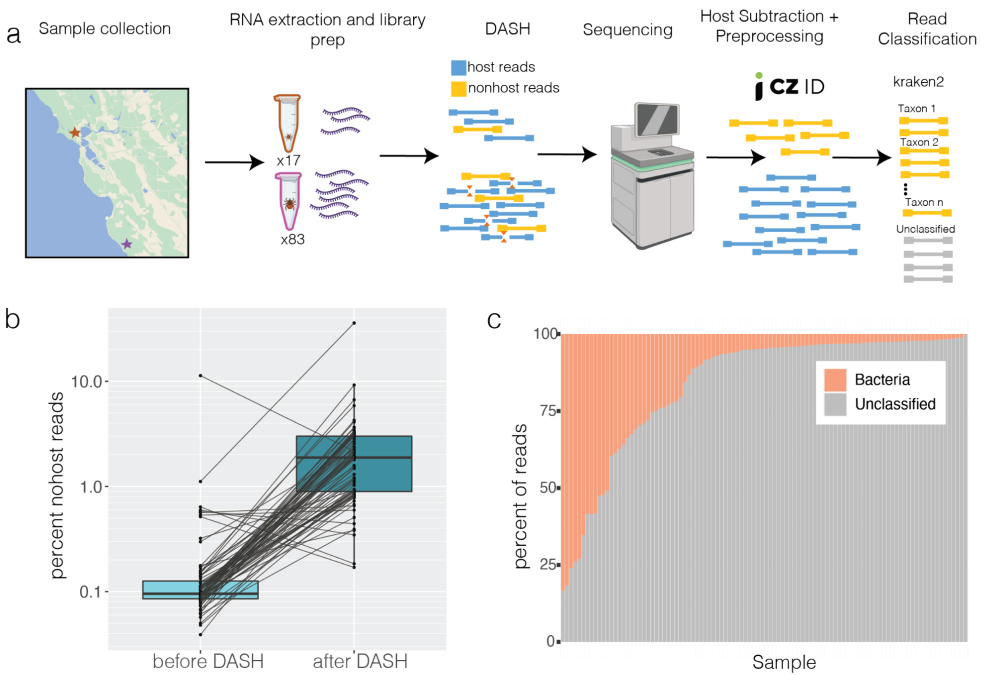


Figure 2.1: Experimental approach a) 17 nymphs and 83 adults were collected from Garrapata State Park (purple) and China Camp State Park (red) respectively. RNA was extracted from whole bodies of individual ticks, mNGS libraries were prepared and DASH was performed to deplete abundant tick sequences. After sequencing, reads were quality filtered and tick-derived reads were removed using CZID. Remaining reads were classified by kraken2. b) The percent of non-host reads as classified by CZID for matched libraries before and after DASH, line connect individual libraries. c) Percentage of nonhost reads classified per sample by kraken2.

(Figure 2.1a). Samples varied substantially in the proportion of reads able to be classified, with anywhere from 0.2% to 83% of reads being classified by the tool (Figure 2.1c). All classified reads by kraken2 were bacterial; no known viruses were identified. Because of this, viruses were classified in a custom pipeline that enabled the detection of divergent viruses (Figure 2.4a, Methods).

To assess the general validity of my approach to characterizing the microbiota of field ticks, I first quantified the bacterial component of tick metatranscriptomes. I identified 114 bacterial genera across the dataset with a median of 11 genera per tick. Larger libraries had more classified genera, indicating that sequencing depth is a limiting factor in characterizing tick microbial diversity. I compared the taxonomic composition of our samples with previously reported tick-associated human pathogens, such as *Rickettsia*, *Anaplasma*, and *Coxiella*, which are most commonly linked to *I. pacificus* ticks^{3,6,15,65}. *Borrelia*,

Borrelia, *Ehrlichia*, and *Bartonella* are also human pathogens known to circulate in this species, although typically at lower frequencies. *Ehrlichia*, *Borrelia*, and *Borrelia* were identified at rates of 2-3% across the full dataset but not in samples with a more stringent quality cut-off of at least 1 million non-host reads (Figure 2.2). I did not identify *Coxiella*, *Bartonella*, or *Francisella* in any ticks, indicating they are either absent in this population or present at levels too low to be detected.

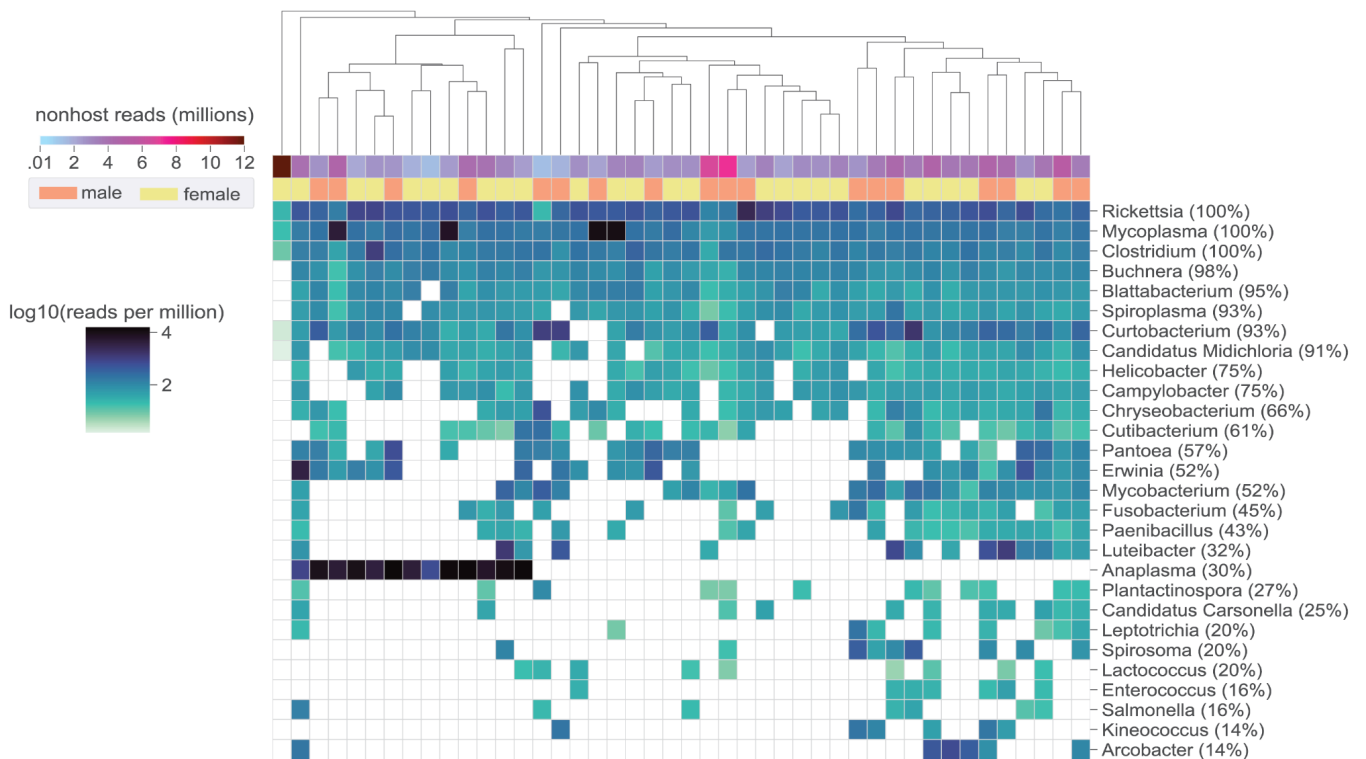


Figure 2.2: Bacterial genera detected in *Ixodes pacificus*. Heatmap displaying reads per million (rpm) of bacterial genera as classified by kraken2. Plot is limited to samples with at least 1 million nonhost reads and genera detected in at least 5 samples. Prevalence in the selected samples is shown next the genus name. Rows are ordered by decreasing prevalence and columns are hierarchically clustered by euclidean distance.

Of the samples with at least 1 million non-host reads, several genera we identified were found at frequencies similar to previous reports^{3,6,15,65}. The endosymbiont *Rickettsia* was detected in 100% and *Anaplasma* in approximately 30% of ticks (Figure 2.2). In total, results from this bacterial analyses are largely consistent with previously reported analyses of bacterial constituents of tick microbiota,

suggesting our RNA-based approach to characterizing tick-associated microbes can indeed be applied to field studies. These results provide confidence that the sequencing and analysis pipeline could reliably report on microbiomes of field *I. pacificus* ticks at nymphal and adult life stages.

2.3 Optimized workflow enables detection of low-abundance bacteria and RNA viruses

The combined approach of host depletion and RNA-sequencing opened up several unique lines of inquiry. In addition to known tick-borne human pathogens, I identified several bacterial genera previously undetected in *I. pacificus*. Although these genera had high prevalence across our samples, they were present at low relative abundances and likely escaped detection in previous studies in the absence of DASH-based microbial enrichment. *Mycoplasma*, a genus that has been linked to Lyme-like disease in patients with tick exposure^{69,70}, was identified in all of the selected libraries (Figure 2.2). *Blattabacterium*, *Buchnera*, *Spiroplasma*, and *Candidatus Midichloria* were all present in at least 90% of samples, and *Candidatus Carsonella* was identified in 25% of samples (Figure 2.2). To my knowledge, none of these known endosymbionts have been commonly identified in *I. pacificus*, and *Blattabacterium*, *Buchnera*, and *Candidatus Carsonella* represent the first report of these genera in any tick⁷¹⁻⁷³.

Sequencing individual ticks also provided sufficient resolution for co-occurrence analyses. I assessed whether presence of one microbial genus increases the statistical likelihood that another microbial genus will be present in the same tick host⁷⁴. This revealed 14 pairs of bacterial genera detected together in a statistically significant number of samples (Figure 2.3). Of note, there was a strong positive association between the endosymbiont *Candidatus Carsonella* and *Chryseobacterium*, a genus that has been shown to be pathogenic to soft ticks but tolerated by hard ticks⁷⁵. The remaining statistically significant co-occurrence relationships did not include any microbial genera known to be tick-associated, and I hypothesized that they may be environmental contaminants, such as soil bacteria. Therefore, outside of

the *Candidatus-Chryseobacterium* case, I did not find clear evidence that any microbes associated with *I. pacificus* actively promote the colonization or growth of other microbes.

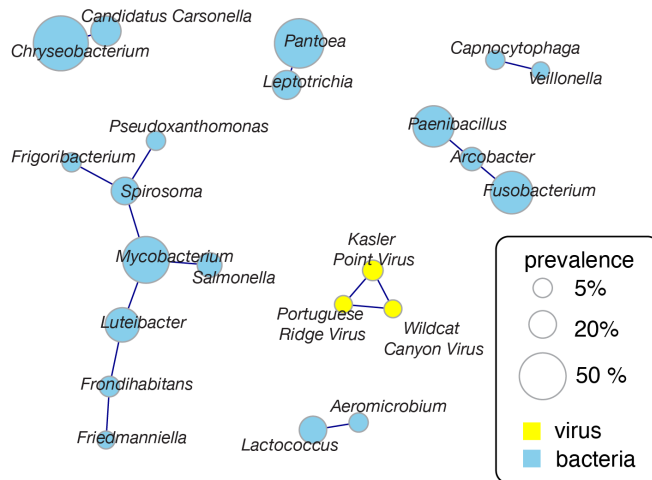


Figure 2.3: Co-occurrence of tick microbes Network representation of significant co-occurring relationships amongst all identified viruses and bacterial genera. Sizes of nodes are scaled to the prevalence in the dataset. All edges represent a positive co-occurrence of alpha value greater than or equal to 5 with a p-value less than or equal to 0.01

My RNA-based sequencing approach also enabled identification of several previously unidentified RNA viruses in *I. pacificus* ticks. Ticks are known to carry a diversity of viruses, and the majority of known transmissible arboviruses have RNA genomes^{15,76–80}. We sought evidence of these and any novel tick viruses in our metatranscriptomes. To do so, I first developed a bioinformatics strategy because standard tools for microbiome analysis (e.g., kraken2) did not detect any known tick viruses in our metatranscriptomic data²³. This is a common phenomenon in RNA virus discovery, due to the fact that the diversity of RNA viruses are not well represented in reference databases. In keeping with other viral discovery efforts^{26,81,82}, I searched for sequences containing an RNA-dependent RNA polymerase (RdRp) domain using HMMER²⁵ (Figure 2.4a).

Using this strategy, I detected a total of 13 new tick viruses in our *I. pacificus* field specimens and determined their prevalence across the dataset as well as their relative abundance within each sample (Figure 2.5a). Underscoring the novelty of these viruses, most had less than 80% amino acid identity to their nearest relative in the NCBI non-redundant protein database (Figure 2.5a). While many of these viruses (10/13) could be defined as members of viral families previously identified in tick species, none corresponded to previously described viral species. Among these, two (Kasler Point Virus and Wildcat Canyon Virus) were assigned to the recently discovered ormycovirus clade, a group of RNA viruses that has not been phylogenetically placed within any existing viral family. As all the viruses were novel isolates, the viruses were named according to geographic features in the region in which the samples were collected.

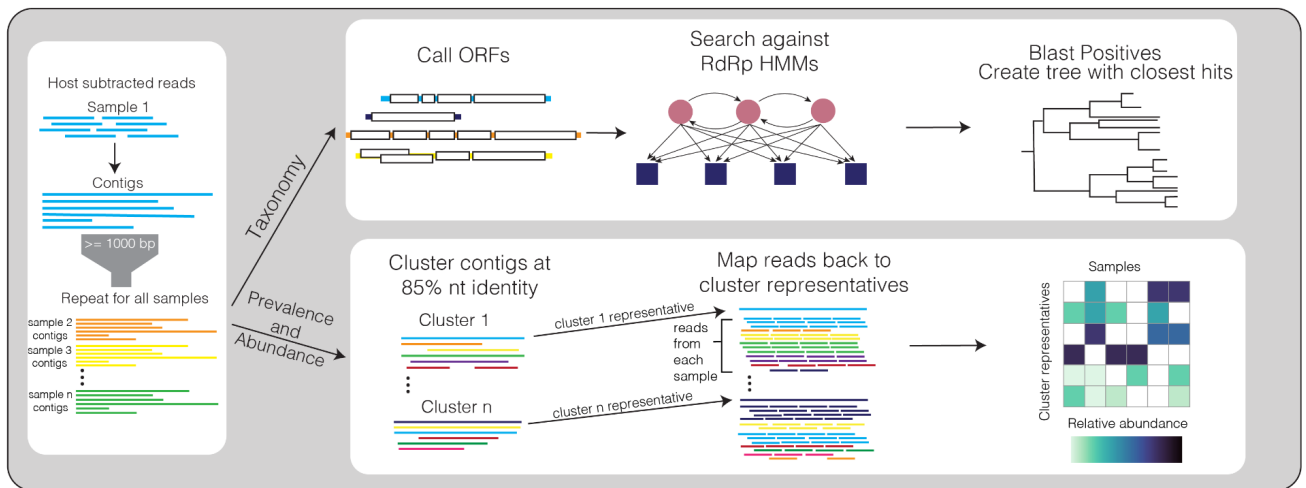


Figure 2.4: Viral discovery pipeline Analysis pipeline for identification of viruses: reads were assembled into contigs and open reading frames were predicted. Resulting proteins were scanned for the presence of an RdRp using HMMER and classified by viral family using the closest hits by blast. Prevalence and abundance was determined by clustering contigs and mapping reads back to cluster representatives.

2.4 RNA viruses are stable constituents of tick microbiota

Our RNA-based approach to tick microbiome characterization led to the discovery of several novel viruses and viral families in *I. pacificus* field ticks. To follow up on these results, I next asked if the viruses we detected were likely to represent the full virome of the sampled ticks or only the most

abundantly transcribed viruses. To do so I performed rarefaction analysis, which shows the number of new taxa discovered as a function of the number of samples sequenced. The rarefaction curve appears to be approaching an asymptote, and the estimated true number of viruses in this population using the Chao index is 14.2 (Figure 2.5b). These results indicate that we have likely discovered the majority of viruses in this population with a median of 1.7 million non-host reads after DASH host-subtraction. This is equivalent to an overall sequencing depth of 106 million reads. Relatively few samples are needed to saturate viral discovery for a given population at this sequencing depth. Hence, we propose ~100 million RNA reads with DASH and rarefaction analysis as a standard for characterizing other arthropod viromes.

I also performed co-occurrence analysis with our newly characterized tick viromes. Not only did I identify a broad diversity of viruses, but I also found evidence of co-occurring viral infections within individual ticks. Ticks had a median of two viruses present with a maximum of six in one individual. I found a statistically significant positive relationship between Portuguese Ridge Virus (*Narnaviridae*), Wildcat Canyon Virus (*ormycovirus*), and Kasler Point Virus (*ormycovirus*) (Figure 2.3). Notably, all three of these viruses contained only an RdRp; no additional segments or genes were identified. Since Narnaviruses are single-gene ribonucleoprotein complexes lacking structural proteins or capsids, it is possible that the two viruses of unknown origin replicate and transmit in a similar manner. Our findings support the model that ticks can harbor multiple viruses per individual, suggesting that RNA viruses are consistent and stable members of the tick microbiome.

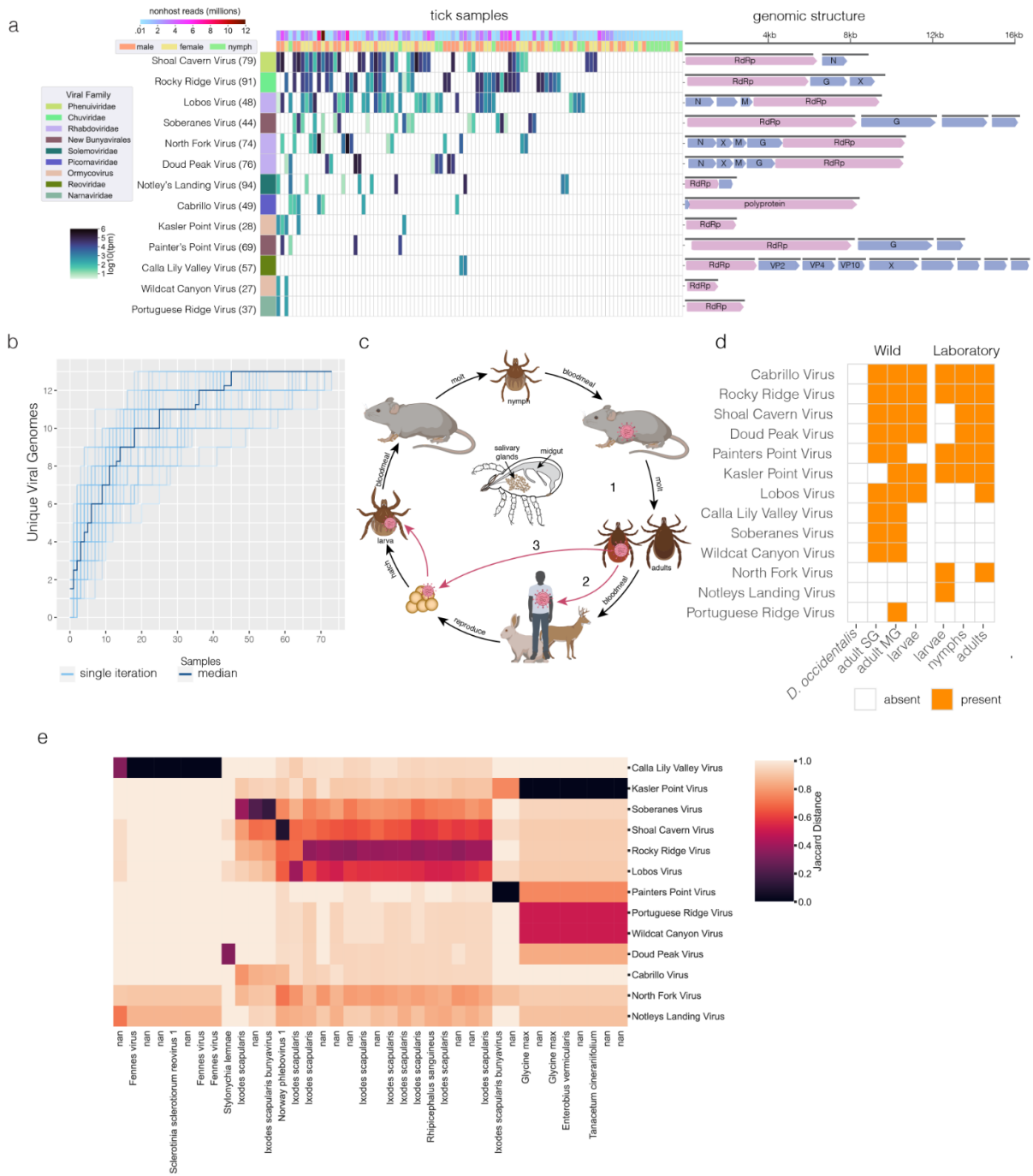


Figure 2.5: Discovery of novel viruses in *Ixodes pacificus* a) Heat map of transcripts per million (tpm) of each identified RdRp containing viral contig across the dataset. Average amino acid identity of the aligned region to closest known sequence is displayed to right of virus name. Open reading frames are displayed to the right of heatmap and annotated with identified protein (RdRp=RNA-dependent RNA polymerase, G=Glycoprotein, N=Nucleoprotein, M=Matrix protein, VP=Viral Protein). Open reading frames with "X" contain homology to known viral proteins of unknown function and open reading frames with no annotation had no identified homology to known proteins. b) Rarefaction plot showing increase in number of viral genomes for each sample analyzed, samples were randomly shuffled 50 times, median shown in dark blue. c) Schematic of *I. pacificus* three stage life cycle. Viruses can be transmitted horizontally between hosts and ticks (1 and 2) or vertically from adult female to offspring (3) d) Table summarizing detection of each viral contig across tick, SG=salivary glands, MG=midgut. e) Heatmap of Jaccard distance of rdrp-containing contigs (rows) and other contigs in the dataset (columns). Columns are annotated with the closest blastx hit or nan if no hits were found.

2.5 Life-stage and tissue tropism of viruses

Arthropods are known to tolerate viral infection more easily than vertebrates, often maintaining infections for life with no apparent ill-effects⁸³. Given that some of the viruses we identified are not only highly prevalent but also closely-related to viruses previously discovered in tick cell lines (Supplementary Note: *Rhabdoviridae*)³⁸, we next characterized several properties that could inform how these viruses are acquired and transmitted. Namely, we examined their distribution in ticks according to both life stage (larval, nymphal, and adult) and tissue (midgut and salivary glands).

To explore whether any of the viruses are stably maintained across life stages *I. pacificus* ticks, we screened cDNA generated from pools of both wild and laboratory-reared *I. pacificus* larvae by polymerase chain reaction (PCR) using primer sets specific to all 13 of the identified viruses. Nine of the 13 viruses were present in either wild-collected or laboratory-reared larvae (Figure 2.5d). Larval ticks have not yet consumed a bloodmeal, and ticks of any life stage are not known to transmit viruses directly amongst a population in the absence of a vertebrate host intermediate⁸⁴. The presence of viruses in the larval stage could be due to vertical transmission, a phenomenon known to occur with several known tick viruses⁸⁴(Figure 2.5c,d). Eight of the 9 viruses in larvae were also detected in either nymphs or adults, suggesting these viruses persist through life stages. Notably, these viruses were detected in ticks that were laboratory-reared, indicating that they can be maintained in the absence of natural bloodmeal hosts, such as the gray squirrel and the Western fence lizard (Figure 2.5c).

We also investigated whether the identified viruses localized to *I. pacificus* compartments associated with microbial transmission. *I. pacificus* ticks can transmit the microbes they harbor in their midguts or salivary glands to bloodmeal hosts during a bloodmeal. While feeding, microbes in these organs can migrate through the tick and inoculate vertebrate hosts via secreted saliva. We reasoned that viral presence in salivary glands in particular could increase the likelihood of feeding-based (horizontal) transmission

(Figure 2.5c). We used our PCR assay to screen cDNA libraries prepared from dissected salivary glands and midguts of additional field-collected *I. pacificus* ticks. The majority of the viruses identified (10/13) were detectable by PCR in tick salivary glands (Figure 2.5d). While this does not definitively demonstrate that these viruses are transmissible by feeding, we hypothesized that this could be possible and worthy of further investigation.

2.6 Identification of novel mRNA-like virus-like transcripts

In addition to the 13 viral RdRps, I identified 21 sequences with homology to an RdRp but with an open reading frame (ORF) structure inconsistent with known RNA viral genomes. Specifically, these RNA sequences encoded clusters of small ORFs with RdRp homology and large gaps (100s of bases) between their predicted ORFs (Figure 2.6a). Many also contained multiple overlapping ORFs and/or small ORFs in opposite orientations. These unusual sequences were highly prevalent (Figure 2.6a) and were independently assembled from multiple different ticks. My analysis of these sequences suggested that they originated from many of the same families as the viral genomes (Figure 2.6b), but with distinct sequences that encoded ORFs that were smaller and more numerous than would be expected for that family.

The observed irregular genomic organization of these 21 virus-like sequences thus we next sought to better understand their possible origins and functions. We first conducted experiments to eliminate potential artifactual explanations for the irregular ORF structure. To test whether these sequences could be the result of a misassembly, we selected one of the longest and most highly prevalent sequences (vlt_111) for more in-depth evaluation. We applied RACE sequencing to examine the sequence in cDNA of a Garrapata tick. Our results confirmed the accuracy of the vlt_111 sequence assembly and indicated that vlt_111 is expressed as a 3' poly-adenylated mRNA, ruling out the possibility that the non-canonical

features of this RNA are due to misassembly (Figure 2.7a).



Figure 2.6: Virus-like transcripts detected across *Ixodes pacificus* population a) Heatmap showing transcript per million value of virus-like transcripts (VLTs) across the dataset. Rows with an * were detected in DNA by PCR. Each VLT is colored by the viral family of its closest hit in blast and predicted open-reading frames (ORFs) are shown to the right. All ORFs in pink have homology to a viral RdRP, ORFs with homology to tick sequences are shown in blue and labeled by name, TRAF=TNF receptor-associated factor 6-like, TE=piggy-Bac transposable element-derived protein 4-like, X=protein of unknown function b) Viral family assignment of both exogenous viruses and virus-like sequences identified. c) Known functions of arthropod endogenous viral elements. d) Heatmap displaying length of longest perfectly matching sequence between each virus and each VLT. Rows and columns are colored by viral family.

I also considered whether the irregular ORF structures we observed could be resolved with alternative codons, such as non-standard stop and start codons. We tested whether ORF prediction with any alternative genetic codes would result in an organization more consistent with that of a viral genome. Of the 25 known genetic codes tested, none substantially changed the ORF structure of any of the sequences (Figure 2.7d), indicating that alternative genetic code alone cannot account for the observed genomic

structure. Having eliminated possible artifactual explanations for how these sequences could originate from exogenous viral genomes, I termed these sequences of unknown origin and function “virus-like transcripts” (VLTs).

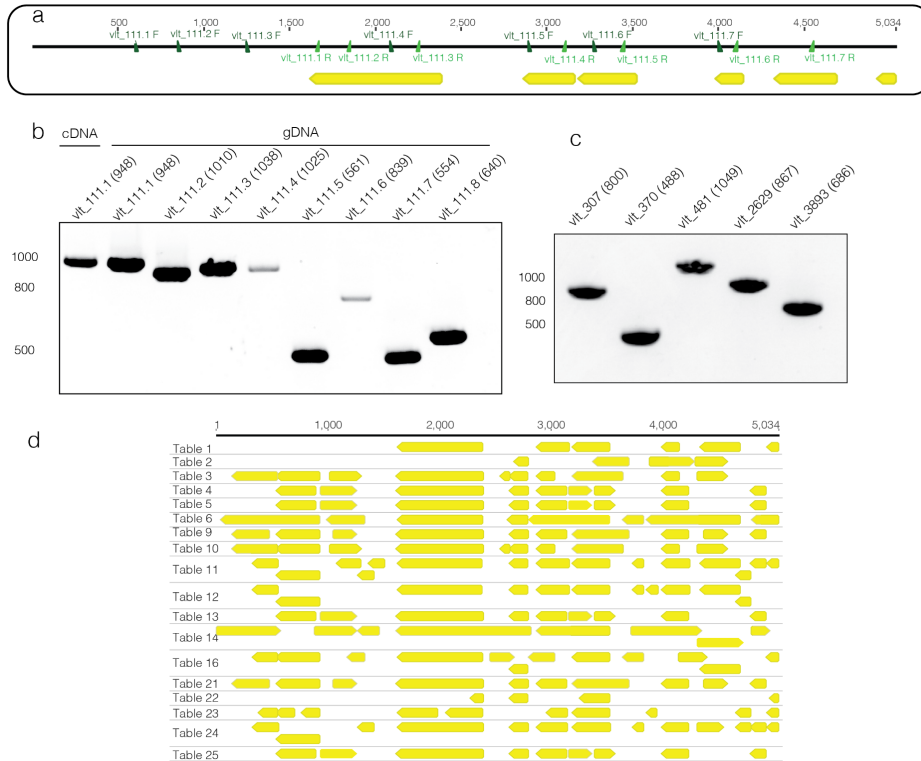


Figure 2.7: Confirmation of VLT sequence and presence in DNA a) Visual representation of c111 identified from RNA seq, including predicted open reading frames (yellow) and primer pairs b) PCR reactions amplifying the regions indicated in a, nucleic acid type indicated on lane. Expected band size in base pairs indicated in parentheses. c) PCR reactions from an additional 5 virus-like sequences amplified from gDNA, expected band size indicated in parentheses. d) Open reading frames predicted for c111 using alternative codon tables

2.7 VLTs likely serve a non-canonical function in *I. pacificus* ticks

Tick genomes are known to contain numerous endogenous viral elements (EVEs), which result from horizontal integration of RNA viral sequences into tick genomes over the course of evolution³⁴. The observed high prevalence of these VLTs lead us to hypothesize that these mysterious sequences could have similar functions or histories. In the absence of a published genome assembly for *I. pacificus*, we

could not check for corresponding sequences in a reference genome. Therefore, we developed PCR primers to screen *I. pacificus* DNA extracts isolated from the same wild-collected ticks used for metatranscriptomics for VLTs. We confirmed the presence of vlt_111 in wild-collected tick DNA (Figure 2.7b) as well as laboratory-reared tick DNA. To ensure this pattern was not specific to this particular VLT, we checked an additional 5 VLTs, all of which were present in lab-reared *I. pacificus* DNA (Figure 2.7c), suggesting a genomic origin for these VLTs.

To confirm that the presence of DNA forms of transcripts is specific to the VLTs and not a general phenomenon, we additionally screened all of the presumed exogenous viral genomes for presence in genomic DNA. Only one of the viral genomes was detected in DNA. A faint band corresponding to *Rocky Ridge Virus* was amplified from genomic DNA. This virus, a mivirus with a circular genome, could represent an intermediate between these two categories that is an exogenous RNA virus with a single or small number of recent genomic integrations into the *I. pacificus* genome. Alternatively, this could be caused by production of DNA forms of the viral genome by endogenous retrotranscriptases (either prior to genomic integration or in the absence of integration) as in Salvati et al⁸⁵. Given that its expression pattern mirrored that of the other exogenous viruses and its genome is both complete and contains the expected ORFs, we continued to classify this sequence as an exogenous virus and not a VLT.

To explore the possibility that our VLTs could have EVE-like functions, I next considered canonical pathways by which EVEs contribute to arthropod immunity (Figure 2.7c). EVEs are most commonly known to function as non-coding RNAs. Much more rarely, they are expressed and translated as proteins that can act as dominant negative viral inhibitors or serve a new function^{34,35}. As the fragmented ORF structure of the VLTs is inconsistent with expression of full-length proteins, I focused on non-coding RNA functions. Typically, arthropod EVEs play antiviral roles by serving as a template for piwi RNAs

(piRNAs), 24-31 nucleotide (nt) RNAs that target an exogenous viral RNA genome for degradation by binding to a complementary sequence within it.

To test this model, I examined *in silico* whether our VLTs could give rise to small RNAs capable of binding the exogenous viral genomes in our dataset through a matching sequence at least 24 nt long. Only one VLT contained a sequence of at least this length (vlt_307) matching one of the viral genome assemblies. Two others contained stretches longer than 20 nucleotides (Figure 2.6d). The remaining 18 VLT sequences do not contain perfect matches longer than 17 nucleotides to any of the exogenous viruses. Of the three VLT-virus combinations with perfect matches of at least 20 nucleotides, none originated from the same tick sample. In total, I did not uncover definitive evidence supporting inhibition of viral replication through a canonical piRNA pathway. These results point to either a non-canonical immunity mechanism or a different functional role entirely for the identified VLTs that could be explored in future studies.

2.8 Methods

Tick Sources

Ticks for the initial metagenomic sequencing were collected by dragging from Garrapata State Park (n=83) and China Camp State Park (n=17). Ticks for followup laboratory experiments were collected exclusively from Garrapata State Park. Adult ticks were separated by sex, surface sterilized in 1% bleach and frozen individually.

Laboratory-reared *I. pacificus* were received from the tick lab at the Center for Disease Control tick lab (Atlanta, GA) and provided through BEI Resources (a service funded by the National Institute of Allergy and Infectious Diseases and managed by ATCC). Ticks were maintained in glass jars with a relative

humidity of 95% (saturated solution of potassium nitrate) in a sealed incubator at 22°C with a light cycle of 16h/8h (light/dark).

RNA Extraction/Library Prep

RNA Extraction and library prep was performed by Amy Kistler and Gloria Castaneda. Total RNA was extracted from the wild-caught *I. pacificus* adult and nymph ticks in 2 separate batches. On ice, individual ticks were transferred to separate wells of a 96 well deepwell plate that was pre-loaded with a single 5mm steel ball bearing (OMNI International, GA, USA) and 400uL of 1X DNA/RNA shield (Zymo Research Corp., Irvine CA, USA) in each well. The plates were sealed and subjected to bead bashing (3 x 3 min, with 1 min rest on ice in between each round of bashing) on a TissueLyser II beadmill (Qiagen, Valencia, CA, USA), then clarified by centrifugation at 2000 rpm at 4°C for 5 min in a refrigerated tabletop centrifuge (Beckman Coulter, Indianapolis IN, USA) to remove large debris. 350uL of the supernatant was transferred to a fresh 96 deepwell plate and re-centrifuged under the same conditions to further clarify the homogenate. 90uL of the resulting supernatant was used as input for total RNA extractions; 110uL of supernatant was transferred to a separate plate and archived at -80°C for potential follow-up analyses.

RNA Extraction/Library Prep

Total RNA was extracted from the wild-caught *I. pacificus* adult and nymph ticks in 2 separate batches. On ice, individual ticks were transferred to separate wells of a 96 well deepwell plate that was pre-loaded with a single 5mm steel ball bearing (OMNI International, GA, USA) and 400uL of 1X DNA/RNA shield (Zymo Research Corp., Irvine CA, USA) in each well. The plates were sealed and subjected to bead bashing (3 x 3 min, with 1 min rest on ice in between each round of bashing) on a TissueLyser II beadmill (Qiagen, Valencia, CA, USA), then clarified by centrifugation at 2000 rpm at 4°C for 5 min in a refrigerated tabletop centrifuge (Beckman Coulter, Indianapolis IN, USA) to remove large debris. 350uL of the supernatant was transferred to a fresh 96 deepwell plate and re-centrifuged under the same

conditions to further clarify the homogenate. 90uL of the resulting supernatant was used as input for total RNA extractions; 110uL of supernatant was transferred to a separate plate and archived at -80°C for potential follow-up analyses.

For both the adult tick and nymph tick homogenate preps, automated RNA extraction was performed in 96 well format (Bravo automated liquid handler, Agilent Technologies, Santa Clara, CA, USA) using a modified version of the Quick DNA/RNA pathogen magbead 96 extraction kit (Zymo Research Corp., Irvine, CA, USA) to automate total nucleic acid extraction and DNase treatment. RNA extracted from 90 uL of tick homogenates was eluted in a final volume of 25uL into 96 well PCR plates. An aliquot of 3uL was used for quantitative and qualitative analysis of the total RNA for each sample via Qubit fluorometer assay (Thermo Fisher Scientific, Waltham MA, USA) and Agilent Bioanalyzer Pico 6000 total Eukaryotic RNA electrophoresis (Agilent Technologies, Santa Clara, CA, USA). A separate 5uL aliquot was used as input for RNAseq library prep, and 2 x 7uL aliquots were stamped into 2 separate daughter plates that were immediately frozen and archived at -80°C for potential follow-up studies.

RNAseq libraries preparation of the 5uL aliquots of adult tick and nymph tick total RNA preps was also performed in 96 well format on an automated liquid handler (Bravo automated liquid handler, Agilent Technologies, Santa Clara, CA, USA). Briefly, the NEBNext Ultra II Non-directional RNAseq library preparation kit (New England Biolabs, Ipswich, MA, USA) was applied, with the following modifications incorporated into the manufacturer's standard protocol: a 25pg aliquot of External RNA Controls Consortium RNA spike-in mix ("ERCC", Thermo-Fisher, Waltham, MA, USA) was added to each sample prior to RNA fragmentation; the input RNA mixture was fragmented for 8 min at 94°C prior to reverse transcription; and a total of 12 cycles of PCR for amplification of resulting individual libraries.

SPRIselect (Beckman Coulter, Indianapolis IN USA) beads were used to size-select libraries with an average total length between 450-550 bp. Library size distributions were verified by Agilent Bioanalyzer

High Sensitivity DNA electrophoresis (Agilent Technologies, Santa Clara, CA, USA) and quantified by Qubit fluorometer (Thermo Fisher Scientific, Waltham MA, USA). Paired-end 2 x 150bp sequencing runs were performed on equivolume pools of individual sequencing libraries of the adult and nymph ticks, respectively, on the Illumina MiSeq sequencing platform (Illumina, San Diego, CA, USA).

The yield of reads/uL acquired from the small scale MiSeq run of the equivolume pools of individual libraries were used to generate approximately equimolar pools of the individual adult tick and nymph tick libraries. The pooled libraries were then depleted of highly abundant sequences^{27,86}, using a previously described pool of tick gRNAs⁸⁷ complexed with in-house prep of purified recombinant Cas9 protein. Resulting DASH'd libraries were qualitatively and quantitatively analyzed by Agilent Bioanalyzer High Sensitivity DNA electrophoresis (Agilent Technologies, Santa Clara, CA, USA) and Qubit fluorometer (Thermo Fisher Scientific, Waltham MA, USA). While the DASH'd adult tick libraries provided sufficient material for large scale metatranscriptomic sequencing, insufficient material remained in the tick nymph libraries that were DASH'd. Thus for large scale metatranscriptomic sequencing, the pool of DASH'd adult tick libraries was combined with a pool of un-DASH'd nymph tick libraries. This pooled prep was subjected to PE 2x150bp format on the NextSeq2000 Illumina sequencing platform (Illumina, San Diego, CA, USA).

Host Subtraction/Pre-Processing

Fastq reads from the run were pre-processed using the CZID pipeline. Libraries underwent quality filtering and adaptor trimming. Host reads were then removed by mapping to closely related genomes. Reads from each library were then assembled into contigs using SPADEs⁸⁸ within the CZID pipeline⁸⁹. Both the nonhost reads and resulting contigs were used for downstream analysis.

Bacterial Classification

Host-subtracted reads were classified using Kraken2 (version 2.1.1)²³. The full kraken2 database was used for classification. Only libraries with at least 1000 classified reads were considered for analysis. Reads per taxon were converted to reads per million (rpm) using library size. To reduce false positives, the rpm value for each taxon was required to be at least 100 times the rpm in any of the control libraries (water and Hela cells) to be considered a positive. Additionally, at least 100 unique minimizers were required for each taxa. Taxa within each library not meeting these thresholds were excluded from analysis. Fewer genera were detected in smaller libraries, including the nymphal ticks from China Camp SP, and samples clustered primarily by library size. We therefore focused subsequent analysis on libraries of at least one million non-host reads.

Virus Identification

We focused our analysis on RNA viruses, as these tend to dominate arthropod viromes^{28,90,91}. Contigs were filtered to those of at least 1500 base pairs. Open reading frames were predicted for these contigs using prodigal⁹², and the resulting proteins were searched using HMMscan from HMMER3 (version 3.3.2)²⁵ against a collection of HMMR profiles of viral RdRPs. The following RdRP HMMs were downloaded from the pfam database⁹³ on March 4, 2021; RdRP_1, RdRP_2, RdRP_3, RdRP_4, RdRP_5, Viral, RdRp_C, Mitovir_RNA_pol, Mononeg_RNA_pol, Birna_RdRp, and Bunya_RdRp. Additionally, custom HMMs were constructed from RdRP sequences for narnaviridae and orthomyxoviridae (sequences and combined HMM available in supplement).

Any sequences with a putative RdRP hit from HMMER were then queried against the full NCBI nonredundant protein database (as of January 24, 2021) using diamond blastp version 0.9.24 to identify their closest hit²⁴. For proteins with multiple hits, the hit with the highest bitscore was reported. Sequences

covering less than 30% of their closest blast hit were initially classified as putative virus-like transcripts, and remaining sequences were classified as exogenous viral sequences. The open reading frame (ORF) structure of all sequences was then manually inspected to confirm this classification. Sequences containing significant gaps between ORFs, or multiple reading frames with RdRP homology (where one was expected) were further classified as virus-like transcripts (VLTs).

Determination of Prevalence and Abundance

Assembled contigs were clustered using cd-hit-est (CDHIT version 4.8.1) at a threshold of 85% nucleotide identity⁹⁴. Circular chuvirus genomes were rotated to a common start position using a custom python script to ensure accurate clustering. 85% identify was chosen as a cutoff to minimize multi-mapping reads between closely related sequences. However, some clusters contain significant sequence diversity and could potentially be considered to contain multiple species. The representative sequences from this clustering were used for all downstream analysis.

Reads from each library were mapped back to the collection of cluster representatives using bowtie2 version 2.4.1⁹⁵. Reads aligning to each contig were counted using samtools idxstats version 1.9⁹⁶. Aligned reads, contig length, and library size were used to calculate rpm and transcripts per million (tpm) values for each library. To consider a contig “present” in a given library, the rpm value was required to be greater than 10 times the value in any of the control libraries. This filter was designed to remove potential false positives caused by cross-contamination of high-titer species.

Identification of Additional Genomic Segments

To identify additional segments of multipartite viral genomes, I searched for contigs that were strongly co-occurring with the RdRp containing contigs. Presence/absence of each contig cluster for each library was determined by reads mapped to each contig using the filters described above. Presence was coded as

a 1 and absence as a 0 and the Jaccard distance was calculated for all pairs of contigs. I considered any sequence with Jaccard distance < 0.4 as a putative genomic segment and further considered homology of the sequence and whether additional segments are expected for the viral family in our determination of whether these sequences represent segments from the same genome (Figure 2.5e).

Rarefaction Analysis

The viral genomes in each sample were determined by the presence of a contig of at least 1000 base pairs (bp) that clustered with one of the 13 representative genomes identified. The presence of a contig rather than read mapping was used to simulate viral discovery in each sample, under the assumption that new viruses discovered may be too divergent to detect by read mapping. The samples were ordered by the number of new genomes seen (not seen in any of the previous samples). The number of new genomes was counted for the addition of each sample. This was repeated for 50 iterations and the median number of new samples at each step was determined. The Chao index was calculated using the R library fossil version 0.40^{97,98}.

Co-Occurrence of Taxa

To determine whether any pairings of taxa (either bacterial or viral) occur more or less frequently than expected given their prevalence I utilized the recently developed metric α ⁷⁴. The presence of each taxon was considered at the genus level for bacteria and at the species level for viruses. The distance α and associated p-value were determined for all pairs of taxa using the CooccurrenceAffinity R package (version 1.0)⁷⁴. Pairs were filtered to those with a p-value ≤ 0.005 . These relationships were visualized as a network with edges corresponding to α and nodes corresponding to taxa using the R package igraph (version 1.3.0)⁹⁹. Node size was scaled according to taxon prevalence in the dataset.

Phylogenetic Trees

Trees were constructed using the RdRP protein sequence of each virus identified. The top 100 closest blast hits were downloaded for each virus and clustered at 85% nucleotide identity. Sequences were aligned using MAFFT version 7.475¹⁰⁰ and maximum likelihood trees were constructed using iqtree 2.0.3¹⁰¹. Trees were visualized in iTOL version 5¹⁰².

Virus-like Sequence PCR

Virus-like sequence PCR was performed by Domokos Lauko. To verify that virus-like transcripts were present in tick genomes and expressed in wild and lab-reared ticks, we extracted tick RNA and genomic DNA and confirmed the presence of virus-like transcripts with PCR. Adult male *I. pacificus* ticks (n=3) were pooled and homogenized by beating for 2 increments of 30 seconds at 4000 bpm in a bench homogenizer (Bead Bug, Benchmark Scientific) with 1.4mm zirconium oxide ceramic beads (Fisher Scientific) in ice-cold TRIzol reagent (Thermo Fisher Scientific). RNA extraction was performed using a Zymo Research Direct-zol RNA Microprep kit (Zymo Research), and RNA was converted to cDNA using Primescript RT reagent kit (Takara Bio) in 10 uL reactions following manufacturer protocols. Genomic DNA was extracted from adult male *I. pacificus* ticks (n=3, separate individuals from RNA), which were pooled, flash frozen, and ground to powder. A DNeasy Blood & Tissue Kit (Qiagen) was used to extract genomic DNA following manufacturer protocols.

PCR experiments amplifying regions of VLTs from tick cDNA and genomic DNA were run using Platinum Superfi II Green PCR master mix (Thermo Fisher Scientific). We loaded 5uL of product onto 0.7% agarose gels and ran them at 160V for 2 hours before imaging. Primer sequences can be found in Table 2.

Virus multiplex PCR

Virus multiplex PCR was performed by Beth Hayes. Viral sequences were analyzed with Snapgene (Dotmatics) and PrimerPlex software (Premier Biosoft) to design 4 sets of multiplex primers that amplify 100-550bp regions. Platinum™ SuperFi II Green PCR Master Mix (Thermo Scientific) was used for all PCR reactions. Mixed cDNA from the original sequenced field-collected ticks was used as a positive control. No-template (water only) reactions were also included as negative controls. Primer pair sequences are listed in Table 2. PCR reactions were analyzed by electrophoresis using a 2% agarose gel containing GelRed (Biotium) and visualized using an Azure c400 imager (Azure Biosystems). At least one band corresponding to each virus was cut out of the agarose gel, purified using the QIAquick Gel extraction kit (Qiagen), and sanger sequenced (GeneWiz) to confirm correct PCR amplification of the intended virus.

Tick Dissections/Extractions

For tissue-tropism determination, wild-collected ticks (n=20) from Garrapata State Park were dissected using a micro scalpel cleaned with 70% isopropanol and a sterile needle. Ticks were dissected in batches of 3-5. The scalpel was cleaned and the needle was replaced between batches. The tick cuticle was excised and the midgut and salivary glands were removed using tweezers cleaned with 70% isopropanol. Tissues were pooled and rinsed in droplets of PBS then transferred by pipette into 300uL of Trizol. Males and females were processed separately and each pool of tissues contained material from 3-10 individuals. PCR results from males and females, as well as biological replicates are collapsed for simplicity in Figure 2.3d.

Whole adult ticks were added to 300uL of Trizol in pools of 4-5 individuals, grouped by species and sex. Nymphal ticks were added to Trizol in a pool of 3, and larval ticks were flash frozen and added to Trizol

in pools of 10-15. All ticks and tick tissues were homogenized by bead-beating with ceramic beads in increments of 30 seconds. Samples were placed on ice between cycles and cycles continued until tissue was visually homogenized.

RNA extraction was performed using Directzol RNA Extraction kits, with on-column DNase1 treatment.

RNA was converted to single stranded cDNA using Quantabio cDNA mastermix in 10uL reactions.

Chapter 3: Factors Shaping Viral Genome Composition

3.1 Viral Composition Groups by Viral Phylogeny

I first used unsupervised methods to determine whether viral genome composition is shaped more strongly by viral or host phylogeny. If viral host is truly driving genome composition, viruses that infect similar hosts should group more closely together, regardless of their phylogenetic background. The viral-host db¹⁰³ provides a valuable resource for examining this question as it contains over 12,000 annotated virus-host relationships, many of which are supported by literature curation. I calculated the dinucleotide bias of all the viral genomes in this database as the odds ratio of each of the 16 possible dinucleotides. I then performed principal components analysis and visualized the results. Bacteria-infecting versus eukaryotic-infecting viruses form two relatively distinct groups (Figure 3.1a). However, within eukaryotic viruses, there is a much clearer grouping by viral family than by host family (Figure 3.1b-c). This is supported by the mean within-group distance between sequences, which is significantly smaller for viral families than host families ($p=2.2e-06$) (Figure 3.1d). This pattern is not specific to this taxonomic level, but rather is consistent at the class and order levels as well.

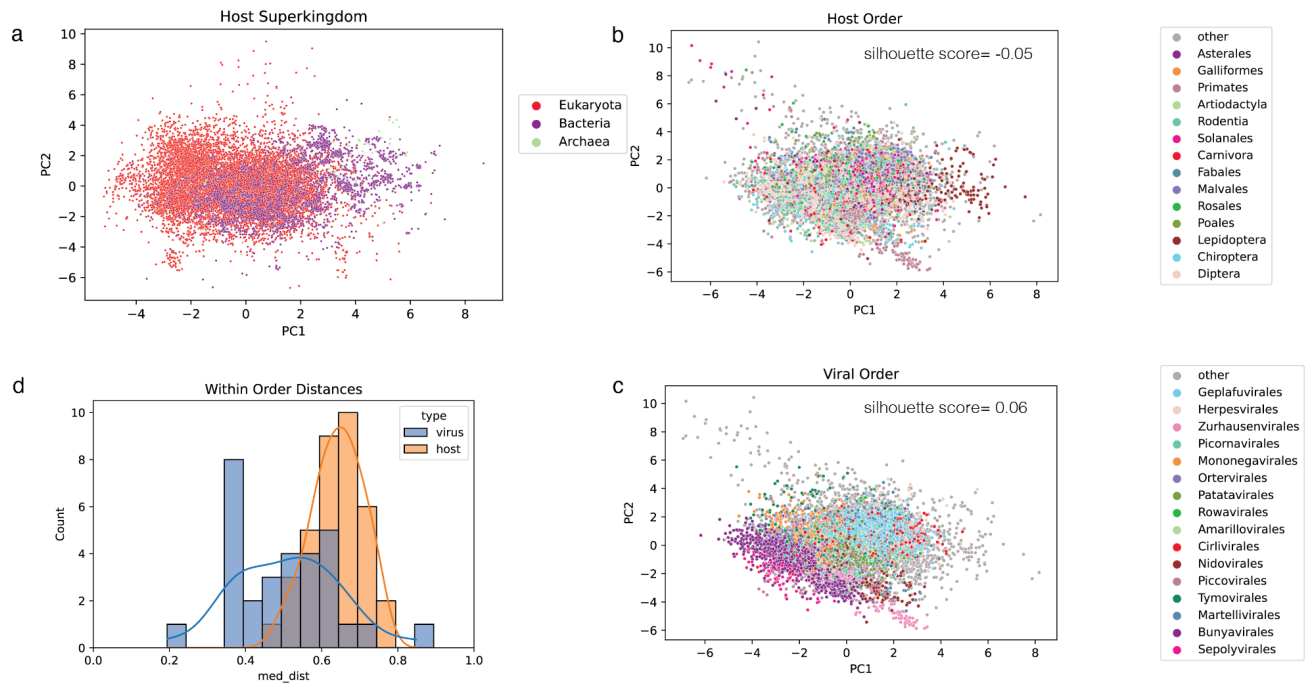


Figure 3.1: Unsupervised analysis of viral dinucleotide bias a) Principal components projection of dinucleotide bias for all sequences in dataset, with points colored by host b) Principal components projection of dinucleotide bias of eukaryotic viral sequences with each sequence colored by the order of the host it infects. c) As in b but colored by the order of the virus. d) Histogram of median euclidean distances between sequences within the same viral order and within the same host order

3.2 Viral Phylogeny is More Predictive than Host

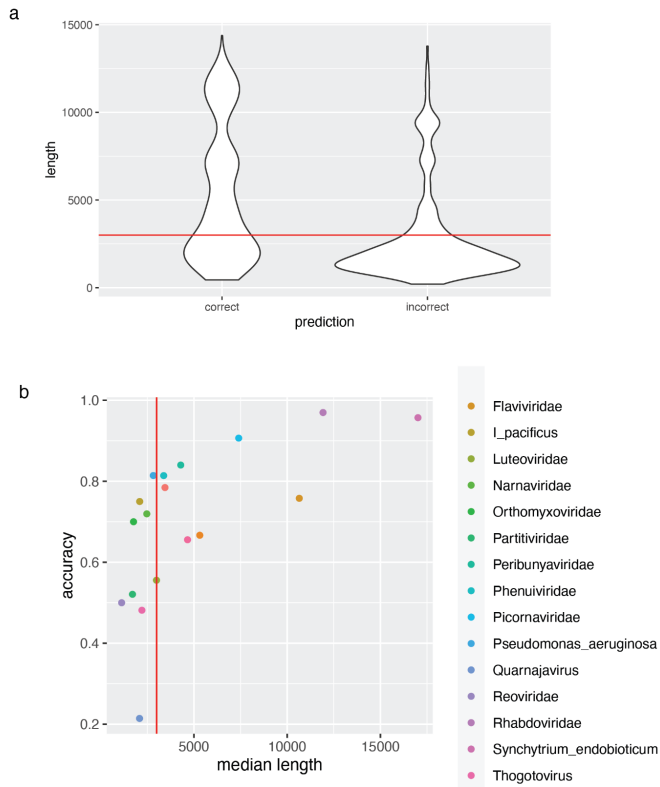


Figure 3.2: Effects of Sequence Length on Prediction Accuracy a) Violin plots displaying the distribution of sequence lengths for correctly and incorrectly classified sequences in a preliminary LDA model. Horizontal line shows length cutoff for subsequence analysis b) Scatter plot showing per-label model accuracy and median sequence length for each predicted label. Vertical line show length cutoff for subsequent analysis

I next evaluated the predictive power of dinucleotide bias for viral phylogeny and viral host. Using Linear Discriminant Analysis (LDA), a model previously used for this task⁶¹, I first trained a model on a subset of viral families that were present in *I. pacificus* ticks as well as two non-viral organisms (*I. pacificus* and *Synchytrium endobioticum*) for differentiation. I found that accuracy of predictions was highly dependent on the length of the sequence. Shorter sequences were more likely to be misclassified than longer ones (Figure 3.2a), and viral families with smaller median lengths had lower accuracy than those with higher

medians (Figure 3.2b). For this reason, in subsequent models I limited sequences to those of at least 3 kilobases (kb). Unfortunately this also excludes entire viral families with segmented genomes.

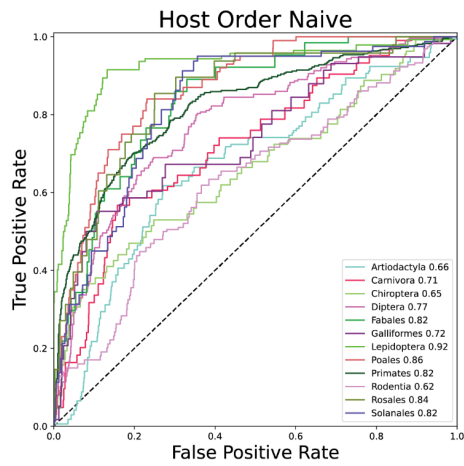
I then trained two LDA models on all of the sequences; one using viral families as the labels and one using host class as the labels. I used a random 80%/20% train/test split. The overall accuracy of the host-predicting model was 38%, with substantial variation in performance amongst the classes as measured by the area under the receiver operating characteristic curve (AUC) (Figure 3.3a), indicating a weaker signal in dinucleotide composition for some classes. In contrast, the accuracy for the model predicting viral family was 82%. Similar to the unsupervised analysis, this is strongly indicative that dinucleotide bias is more strongly linked to viral phylogeny than by the host the virus infects.

However, a random train-test split could lead to bias by including highly similar sequences in both the train and test set. This could artificially raise the accuracy of both models. Accordingly, I repeated the analysis but held out entire viral genera from each viral family. This reduced the accuracy of the host model to 25% and the viral model to 41%. Interestingly, there was a wide range of AUC values in this model for individual viral families; some such as *Coronaviridae*, *Rhabdoviridae*, and *Potyviridae* maintained AUC values of .99 or higher, whereas others such as *Adenoviridae*, and *Flaviviridae* dropped precipitously (Figure 3.3d).

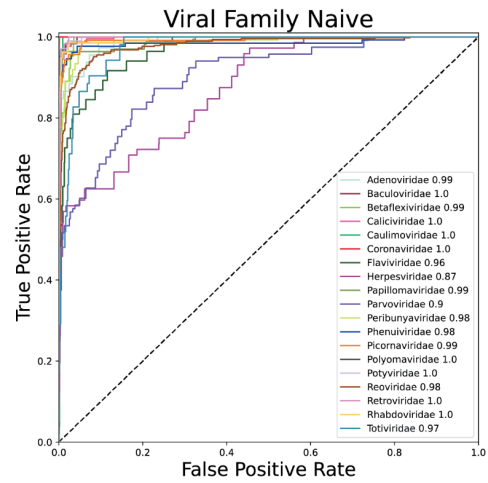
I was curious about what could cause this variation in different families. So I next compared the AUC of each class in the phylogenetically aware model to several other factors. The AUC does not seem to be associated with the number of sequences analyzed per family, nor by genome length or number of hosts (Figure 3.4). I also noted that some viral families contained more genera, and many had highly imbalanced genera (e.g., one genus comprising 80% of the sequences and 10 genera comprising the remaining 20%) which could make it difficult to make a generalizable model. However, neither the number of genera per family nor the proportion of the largest genus correlated well with family AUC

(Figure 3.4). More investigation is needed to understand this underlying variability amongst families, particularly whether this is due to differing amounts of sequence variability (such as difference between well-studied families with highly similar sequences deposited and newer families with more divergent sequences), modes of transmission (vector-borne, airborne, contact dependent), or other features of viral replication and transmission. However it is clear that viral phylogeny is more strongly associated with dinucleotide composition than viral host.

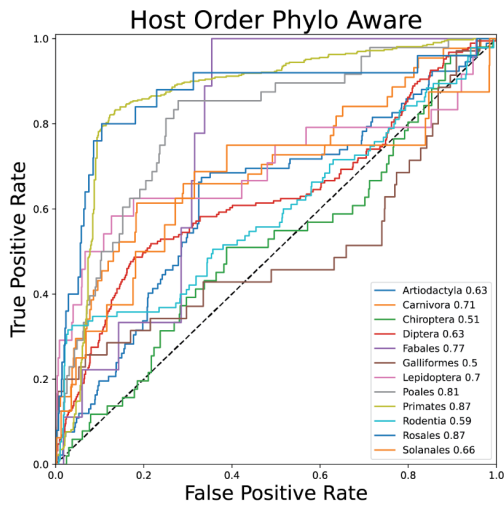
a



b



c



d

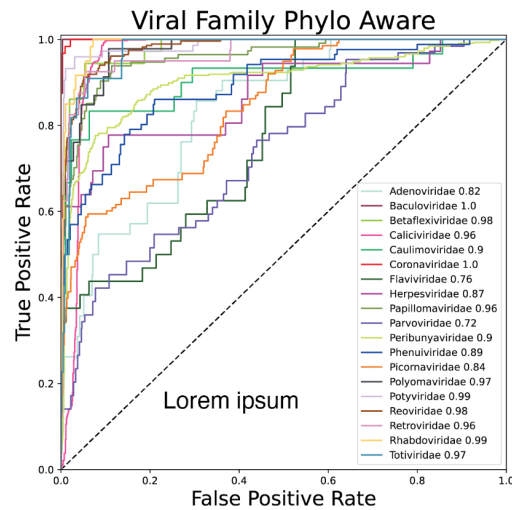


Figure 3.3: Classifier performance a-b) Receiver operating curves (ROC) for LDA classifiers predicting viral host (a) and viral family (b) from dinucleotide bias using randomly split train-test data. Area under the ROC (AUC) for each label is displayed in legend. c-d) As in a-b but using train-test split with regards to viral phylogeny

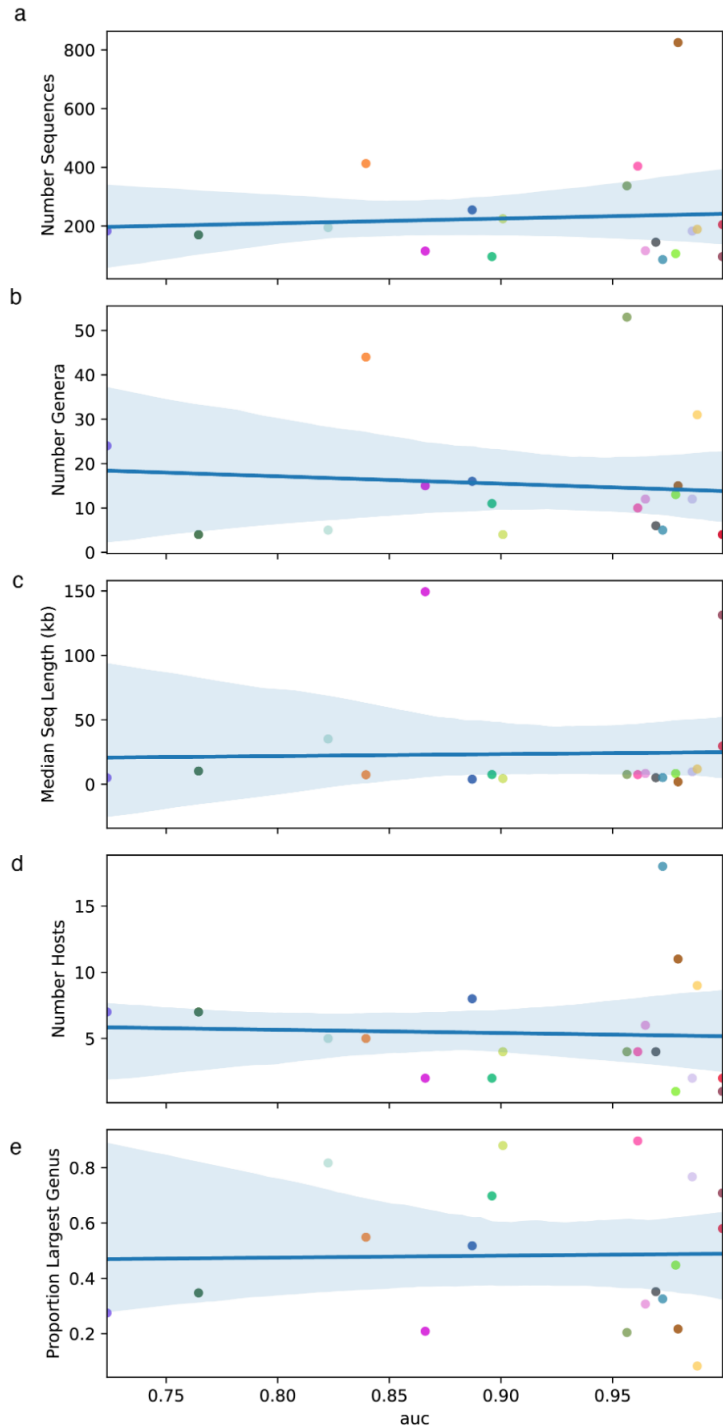


Figure 3.4: Associations with Viral Family AUC a) scatterplot displaying relationship between AUC value of viral family and number of sequences analyzed in family. Line shows linear regression model fit. b-e) as in a but displaying association with the number of genera per family, the median sequence length of the family, the number of hosts infected by the family, and the proportion of sequences accounted for by the largest genera in the family

3.3 Methods

Virus-host associations, including accession numbers for viral genomes were obtained from virus-host db¹⁰³. Viral genomes were obtained from NCBI nt and dinucleotide bias was calculated as the odds ratio of each pair of nucleotides compared to the underlying frequencies of each individual nucleotide (16 total). PCA and LDA analysis was performed in python using the scikit-learn package¹⁰⁴. Matplotlib and seaborn were used for visualization. An initial LDA model was trained on a subset of viral families, sequences for which were obtained from NCBI nt. Families were selected based on their presence in ticks in Chapter 2.

For LDA analysis, the sequences were first filtered to sequences with only one annotated host. The data were further filtered to include only labels with at least 75 sequences. The data was split first into a random 80% of viruses for training and 20% for testing. For the phylogenetically-aware analysis, viral genera were ranked from lowest to highest percentage of the family (in terms of number of sequences). Genera were then added to the held out set in order, until at least 20% of the data was held out. This split was utilized for both the viral family and host order prediction models. Both random and phylogenetically-aware splits were performed one time. The taxonomic level at which sequences were predicted (viral family and host order) was selected to maximize the total number of sequences in the dataset (those with no annotation at a given taxonomic level were dropped) as well as to balance classes.

Chapter 4: Conclusion

4.1 Tick Metagenomics

In this thesis, I show the power of combining experimental enrichment of microbial sequences with single-tick metatranscriptomics for identification of bacteria and viruses in field *I. pacificus* tick communities. The ability to deeply sequence the non-host fraction allowed me to identify several genera of bacteria in *I. pacificus* previously unidentified in any tick species. Further investigation is warranted into the consequences and mechanisms underlying these symbiotic tick–bacteria partnerships.

This approach also enabled the discovery of many novel viruses, which we further investigated in a series of laboratory experiments. I found that several of these viruses were not only highly prevalent but were also present in tick salivary glands. This has important implications for public health as bloodmeal hosts (including humans) are very likely to be exposed to viruses present in the salivary glands during feeding. Currently, tick-borne disease surveillance in California is focused on Lyme disease (*Borrelia*) and a small number of other bacterial pathogens, but my results indicate humans in the region may also be exposed to diverse viruses¹. While our approach cannot determine whether any of these viruses are human pathogens, screening of clinical samples for exposure to them is a critical first step to understanding their disease-causing potential.

In addition to providing comprehensive and quantitative insights into the *I. pacificus* microbiome, one of the most exciting and unexpected themes that emerged from my work relates to how viruses fit into the broader framework of the tick microbiome and tick biology. We found evidence that several viruses persisted in ticks across multiple life stages, including juvenile naïve larvae, as well as across wild and laboratory-reared populations, suggesting they are stable constituents of the tick microbiome. Certain viruses in *I. pacificus* may be able to establish and maintain independent niches within their tick hosts.

These findings lay the groundwork for future work aimed at understanding tick–virus dynamics and how such relationships fit into our understanding of tick physiology.

Further underscoring the critical importance of tick–virus interactions for *I. pacificus* biology was the discovery that numerous VLTs may originate from EVEs in the tick genome. Closer experimental evaluation of our VLTs pointed to a non-canonical mechanism that is distinct from known antiviral pathways. While we cannot eliminate the possibility that VLTs are spuriously expressed without serving an adaptive function, EVEs are an underexplored feature of tick genomes and future studies could determine whether VLTs represent a new mechanistic class of EVEs with adaptive contributions to tick immunity and biology.

The integration of RNA viral genomes into the tick genome as EVEs also provides a unique historical footprint for viruses that may have infected that tick host in the near or distant past. Interestingly, the VLTs identified in *I. pacificus* appear to derive from several viral families from which no exogenous viral genomes were found in this study, including the recently-discovered segmented flaviviruses that cause febrile illness^{105–107}. Future field studies that expand on our *I. pacificus* virome analyses will help determine whether VLTs stem from ancient tick–virus interactions or contemporaneous interactions that were not captured in this study due to low abundance, our limited sample size, and our focus on two collection locations.

Altogether, these results highlight the need for more studies such as this in order to capture the full range of tick–associated microbes that could represent critical components of tick physiology or poorly understood pathogenic threats to human health. This work provides an improved experimental and computational framework with increased sensitivity for low-abundance bacterial and viral taxa present in this increasingly important class of arthropod disease vectors.

4.3 Viral Genome Composition

This work provides an important basis for understanding the variability in genome composition across viruses. My findings indicate that viral phylogeny is a stronger driver of dinucleotide composition than is viral host, a fact with important implications for predictive tools. However, my analysis is limited in scope both by the sequences included and the features selected. It is possible that including additional features such as codon usage bias, or tri- and tetra-nucleotide bias could help better resolve viral hosts. However, these features are difficult to estimate accurately on short sequences like viral genome assemblies. Furthermore, this study is limited to the set of genomes contained in a single well-annotated database. These sequences are likely biased toward viruses that have been better studied and therefore have phylogenetically well-placed genomes and well-annotated hosts. There have been several large-scale viral discovery efforts which have increased the catalog of known viruses by over an order of magnitude^{108,109}. However, many of these viruses do not have a known host. Nonetheless, extension of my methods to these more comprehensive datasets would help provide a more accurate understanding of the factors driving viral genome composition.

However, even within this limited analysis, it is clear that there are several important considerations to account for when developing machine learning models to make predictions based on viral sequence. Firstly, eukaryotic viral phylogeny is tightly correlated with host tropism; phylogenetically similar viruses infect similar hosts. Therefore, any model predicting host type from genomic features should control for phylogenetic background, especially if the goal is to determine how much of viral genome composition can be explained by the host specifically. Secondly, randomly partitioning phylogenetically related sequences into training and test sets can lead to biases that artificially inflate the performance of such models. Therefore, researchers should take care to hold out phylogenetically novel data. Performing group cross-validation (i.e., blocking) is one potential solution.

References

1. CDC. Tickborne disease surveillance data summary | CDC. *Centers for Disease Control and Prevention* <https://www.cdc.gov/ticks/data-summary/index.html> (2021).
2. Tokarz, R. *et al.* Microbiome analysis of Ixodes scapularis ticks from New York and Connecticut. *Ticks Tick-Borne Dis.* (2019) doi:10.1016/j.ttbdis.2019.04.011.
3. Ross, B. D. *et al.* Ixodes scapularis does not harbor a stable midgut microbiome. *ISME J.* **12**, 2596–2607 (2018).
4. Aivelo, T., Norberg, A. & Tschirren, B. Bacterial microbiota composition of Ixodes ricinus ticks: the role of environmental variation, tick characteristics and microbial interactions. *PeerJ* **7**, e8217 (2019).
5. Chandra, S., Harvey, E., Emery, D., Holmes, E. C. & Šlapeta, J. Unbiased Characterization of the Microbiome and Virome of Questing Ticks. *Front. Microbiol.* **12**, (2021).
6. Socarras, K. M. *et al.* Species-Level Profiling of Ixodes pacificus Bacterial Microbiomes Reveals High Variability Across Short Spatial Scales at Different Taxonomic Resolutions. *Genet. Test. Mol. Biomark.* **25**, 551–562 (2021).
7. Adegoke, A. *et al.* Tick-Borne bacterial and protozoan animal pathogens shape the native microbiome within Hyalomma anatolicum anatolicum and Rhipicephalus microplus tick vectors. <http://biorxiv.org/lookup/doi/10.1101/2020.01.20.912949> (2020) doi:10.1101/2020.01.20.912949.
8. Cross, S. T. *et al.* Co-Infection Patterns in Individual Ixodes scapularis Ticks Reveal Associations between Viral, Eukaryotic and Bacterial Microorganisms. *Viruses* **10**, (2018).
9. Bonnet, S. I., Binetruy, F., Hernández-Jarguín, A. M. & Duron, O. The Tick Microbiome: Why Non-pathogenic Microorganisms Matter in Tick Biology and Pathogen Transmission. *Front. Cell. Infect. Microbiol.* **7**, (2017).
10. Chicana, B., Couper, L. I., Kwan, J. Y., Tahiraj, E. & Swei, A. Comparative Microbiome Profiles of

- Sympatric Tick Species from the Far-Western United States. *Insects* **10**, 353 (2019).
11. Díaz-Sánchez, S. *et al.* Characterization of the bacterial microbiota in wild-caught *Ixodes ventralloi*. *Ticks Tick-Borne Dis.* **10**, 336–343 (2019).
 12. Estrada-Peña, A., Cabezas-Cruz, A., Pollet, T., Vayssier-Taussat, M. & Cosson, J.-F. High Throughput Sequencing and Network Analysis Disentangle the Microbial Communities of Ticks and Hosts Within and Between Ecosystems. *Front. Cell. Infect. Microbiol.* **8**, 236 (2018).
 13. Lee, S. *et al.* Comparative microbiomes of ticks collected from a black rhino and its surrounding environment. *Int. J. Parasitol. Parasites Wildl.* **9**, 239–243 (2019).
 14. Lejal, E. *et al.* Temporal patterns in *Ixodes ricinus* microbial communities: an insight into tick-borne microbe interactions. *Microbiome* **9**, 153 (2021).
 15. Bouquet, J. *et al.* Metagenomic-based Surveillance of Pacific Coast tick *Dermacentor occidentalis* Identifies Two Novel Bunyaviruses and an Emerging Human Rickettsial Pathogen. *Sci. Rep.* **7**, 12234 (2017).
 16. Couper, L. I., Kwan, J. Y., Ma, J. & Sweit, A. Drivers and patterns of microbial community assembly in a Lyme disease vector. *Ecol. Evol.* **9**, 7768–7779 (2019).
 17. Ravi, A. *et al.* Metagenomic profiling of ticks: Identification of novel rickettsial genomes and detection of tick-borne canine parvovirus. *PLoS Negl. Trop. Dis.* **13**, e0006805 (2019).
 18. Tufts, D. M. *et al.* A metagenomic examination of the pathobiome of the invasive tick species, *Haemaphysalis longicornis*, collected from a New York City borough, USA. *Ticks Tick-Borne Dis.* **11**, 101516 (2020).
 19. Kurilshikov, A. *et al.* Comparative Metagenomic Profiling of Symbiotic Bacterial Communities Associated with *Ixodes persulcatus*, *Ixodes pavlovskyi* and *Dermacentor reticulatus* Ticks. *PLoS One* **10**, e0131413 (2015).
 20. Xia, H. *et al.* Metagenomic profile of the viral communities in *Rhipicephalus* spp. ticks from Yunnan,

- China. *PloS One* **10**, e0121609 (2015).
21. Zakham, F. *et al.* Viral RNA Metagenomics of Hyalomma Ticks Collected from Dromedary Camels in Makkah Province, Saudi Arabia. *Viruses* **13**, 1396 (2021).
 22. QIIME allows analysis of high-throughput community sequencing data | Nature Methods. <https://www.nature.com/articles/nmeth.f.303>.
 23. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
 24. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
 25. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
 26. Single mosquito metatranscriptomics identifies vectors, emerging pathogens and reservoirs in one assay | eLife. <https://elifesciences.org/articles/68353v1>.
 27. Gu, W. *et al.* Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol.* **17**, 41 (2016).
 28. Vandegrift, K. J. & Kapoor, A. The Ecology of New Constituents of the Tick Virome and Their Relevance to Public Health. *Viruses* **11**, (2019).
 29. Duron, O. *et al.* Tick-Bacteria Mutualism Depends on B Vitamin Synthesis Pathways. *Curr. Biol. CB* **28**, 1896-1902.e5 (2018).
 30. Guizzo, M. G. *et al.* A Coxiella mutualist symbiont is essential to the development of Rhipicephalus microplus. *Sci. Rep.* **7**, 17554 (2017).
 31. Hofer, U. Tick symbiont promotes blood feeding. *Nat. Rev. Microbiol.* **19**, 744–744 (2021).
 32. Hunter, D. J. *et al.* The Rickettsia Endosymbiont of Ixodes pacificus Contains All the Genes of De Novo Folate Biosynthesis. *PLOS ONE* **10**, e0144552 (2015).

33. Chou, S. *et al.* Transferred interbacterial antagonism genes augment eukaryotic innate immune function. *Nature* **518**, 98–101 (2015).
34. Ter Horst, A. M., Nigg, J. C., Dekker, F. M. & Falk, B. W. Endogenous Viral Elements Are Widespread in Arthropod Genomes and Commonly Give Rise to PIWI-Interacting RNAs. *J. Virol.* **93**, (2019).
35. Palatini, U. *et al.* Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors *Aedes aegypti* and *Aedes albopictus*. *BMC Genomics* **18**, 512 (2017).
36. Russo, A. G., Kelly, A. G., Enosi Tuipulotu, D., Tanaka, M. M. & White, P. A. Novel insights into endogenous RNA viral elements in *Ixodes scapularis* and other arbovirus vector genomes. *Virus Evol.* **5**, (2019).
37. Bell-Sakyi, L. & Attoui, H. Endogenous tick viruses and modulation of tick-borne pathogen growth. *Front. Cell. Infect. Microbiol.* **3**, (2013).
38. Alberdi, M. P. *et al.* Detection and identification of putative bacterial endosymbionts and endogenous viruses in tick cell lines. *Ticks Tick-Borne Dis.* **3**, 137–146 (2012).
39. Huttener, R. *et al.* GC content of vertebrate exome landscapes reveal areas of accelerated protein evolution. *BMC Evol. Biol.* **19**, 144 (2019).
40. Jenkins, G. M. & Holmes, E. C. The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Res.* **92**, 1–7 (2003).
41. Liu, Y. A code within the genetic code: codon usage regulates co-translational protein folding. *Cell Commun. Signal.* **18**, 145 (2020).
42. Parvathy, S. T., Udayasuriyan, V. & Bhadana, V. Codon usage bias. *Mol. Biol. Rep.* **49**, 539–565 (2022).
43. Coleman, J. R. *et al.* Virus Attenuation by Genome-Scale Changes in Codon Pair Bias. *Science* **320**, 1784–1787 (2008).

44. Kunec, D. & Osterrieder, N. Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias. *Cell Rep.* **14**, 55–67 (2016).
45. Mazumdar, P., Binti Othman, R., Mebus, K., Ramakrishnan, N. & Ann Harikrishna, J. Codon usage and codon pair patterns in non-grass monocot genomes. *Ann. Bot.* **120**, 893–909 (2017).
46. Cooper, D. N. & Krawczak, M. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Hum. Genet.* **83**, 181–188 (1989).
47. Torrent, M., Chalancon, G., de Groot, N. S., Wuster, A. & Madan Babu, M. Cells alter their tRNA abundance to selectively regulate protein synthesis during stress conditions. *Sci. Signal.* **11**, eaat6409 (2018).
48. Gaunt, E. R. & Digard, P. Compositional biases in RNA viruses: Causes, consequences and applications. *Wiley Interdiscip. Rev. RNA* **13**, e1679 (2022).
49. Simmonds, P., Xia, W., Baillie, J. K. & McKinnon, K. Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla--selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. *BMC Genomics* **14**, 610 (2013).
50. Martrus, G., Nevot, M., Andres, C., Clotet, B. & Martinez, M. A. Changes in codon-pair bias of human immunodeficiency virus type 1 have profound effects on virus replication in cell culture. *Retrovirology* **10**, 78 (2013).
51. Mogro, E. G., Bottero, D. & Lozano, M. J. Analysis of SARS-CoV-2 synonymous codon usage evolution throughout the COVID-19 pandemic. *Virology* **568**, 56–71 (2022).
52. Zielezinski, A., Deorowicz, S. & Gudyś, A. PHIST: fast and accurate prediction of prokaryotic hosts from metagenomic viral sequences. *Bioinforma. Oxf. Engl.* **38**, 1447–1449 (2021).
53. Lu, C. *et al.* Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC Biol.* **19**, 5 (2021).
54. Galiez, C., Siebert, M., Enault, F., Vincent, J. & Söding, J. WISH: who is the host? Predicting

- prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* **33**, 3113–3114 (2017).
55. Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A. & Sun, F. Alignment-free $\$d_2^*$ oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.* **45**, 39–53 (2017).
 56. J, V. *et al.* HostPhinder: A Phage Host Prediction Tool. *Viruses* **8**, (2016).
 57. Liu, D., Ma, Y., Jiang, X. & He, T. Predicting virus-host association by Kernelized logistic matrix factorization and similarity network fusion. *BMC Bioinformatics* **20**, 594 (2019).
 58. Zhang, Z. *et al.* Rapid identification of human-infecting viruses. *Transbound. Emerg. Dis.* **66**, 2517–2522 (2019).
 59. Gałan, W., Bąk, M. & Jakubowska, M. Host Taxon Predictor - A Tool for Predicting Taxon of the Host of a Newly Discovered Virus. *Sci. Rep.* **9**, 3436 (2019).
 60. Mock, F., Viehweger, A., Barth, E. & Marz, M. VIDHOP, viral host prediction with deep learning. *Bioinformatics* **37**, 318–325 (2021).
 61. Giallonardo, F. D., Schlub, T. E., Shi, M. & Holmes, E. C. Dinucleotide Composition in Animal RNA Viruses Is Shaped More by Virus Family than by Host Species. *J. Virol.* **91**, e02381-16 (2017).
 62. Hahn, M. B., Jarnevich, C. S., Monaghan, A. J. & Eisen, R. J. Modeling the Geographic Distribution of *Ixodes scapularis* and *Ixodes pacificus* (Acari: Ixodidae) in the Contiguous United States. *J. Med. Entomol.* **53**, 1176–1191 (2016).
 63. Eisen, R. J., Eisen, L. & Beard, C. B. County-Scale Distribution of *Ixodes scapularis* and *Ixodes pacificus* (Acari: Ixodidae) in the Continental United States. *J. Med. Entomol.* **53**, 349–386 (2016).
 64. Salkeld, D. J., Porter, W. T., Loh, S. M. & Nieto, N. C. Time of year and outdoor recreation affect human exposure to ticks in California, United States. *Ticks Tick-Borne Dis.* **10**, 1113–1117 (2019).
 65. McVicar, M., Rivera, I., Reyes, J. B. & Gulia-Nuss, M. Ecology of *Ixodes pacificus* Ticks and Associated Pathogens in the Western United States. *Pathogens* **11**, 89 (2022).

66. Porter, W. T. *et al.* Predicting the current and future distribution of the western black-legged tick, *Ixodes pacificus*, across the Western US using citizen science collections. *PLoS One* **16**, e0244754 (2021).
67. Forum: Reported Distribution of *Ixodes scapularis* and *Ixodes pacificus* (Acari: Ixodidae) in the United States | Journal of Medical Entomology | Oxford Academic.
<https://academic.oup.com/jme/article/35/5/629/2221705>.
68. Xu, G., Pearson, P., Dykstra, E., Andrews, E. S. & Rich, S. M. Human-Biting *Ixodes* Ticks and Pathogen Prevalence from California, Oregon, and Washington. *Vector-Borne Zoonotic Dis.* **19**, 106–114 (2019).
69. Eskow, E., Adelson, M. E., Rao, R.-V. S. & Mordechai, E. Evidence for disseminated *Mycoplasma fermentans* in New Jersey residents with antecedent tick attachment and subsequent musculoskeletal symptoms. *J. Clin. Rheumatol. Pract. Rep. Rheum. Musculoskelet. Dis.* **9**, 77–87 (2003).
70. Straub, R. K. & Powers, C. M. Chronic Fatigue Syndrome: A Case Report Highlighting Diagnosing and Treatment Challenges and the Possibility of Jarisch–Herxheimer Reactions If High Infectious Loads Are Present. *Healthcare* **9**, 1537 (2021).
71. Sasser, D. *et al.* ‘*Candidatus Midichloria mitochondrii*’, an endosymbiont of the tick *Ixodes ricinus* with a unique intramitochondrial lifestyle. *Int. J. Syst. Evol. Microbiol.* **56**, 2535–2540 (2006).
72. Ogata, S. *et al.* Spiroplasma Infection among Ixodid Ticks Exhibits Species Dependence and Suggests a Vertical Pattern of Transmission. *Microorganisms* **9**, 333 (2021).
73. Beliauskaja, A. *et al.* Spiroplasma Isolated From Third-Generation Laboratory Colony *Ixodes persulcatus* Ticks. *Front. Vet. Sci.* **8**, (2021).
74. Mainali, K. P., Slud, E., Singer, M. C. & Fagan, W. F. A better index for analysis of co-occurrence and similarity. *Sci. Adv.* **8**, eabj9204.
75. Buresová, V., Franta, Z. & Kopáček, P. A comparison of *Chryseobacterium indologenes*

- pathogenicity to the soft tick *Ornithodoros moubata* and hard tick *Ixodes ricinus*. *J. Invertebr. Pathol.* **93**, 96–104 (2006).
76. Blomström, A.-L. *et al.* Novel Viruses Found in Antricola Ticks Collected in Bat Caves in the Western Amazonia of Brazil. *Viruses* **12**, 48 (2020).
77. Brinkmann, A. *et al.* A metagenomic survey identifies Tamdy orthonairovirus as well as divergent phlebo-, rhabdo-, chu- and flavi-like viruses in Anatolia, Turkey. *Ticks Tick-Borne Dis.* **9**, 1173–1183 (2018).
78. Cholleti, H. *et al.* Viral metagenomics reveals the presence of highly divergent quaranjavirus in *Rhipicephalus* ticks from Mozambique. *Infect. Ecol. Epidemiol.* **8**, 1478585 (2018).
79. Harvey, E. *et al.* Extensive Diversity of RNA Viruses in Australian Ticks. *J. Virol.* **93**, e01358-18 (2019).
80. Kobayashi, D. *et al.* RNA virome analysis of questing ticks from Hokuriku District, Japan, and the evolutionary dynamics of tick-borne phleboviruses. *Ticks Tick-Borne Dis.* **11**, 101364 (2020).
81. Reyes, A., Alves, J. M. P., Durham, A. M. & Gruber, A. Use of profile hidden Markov models in viral discovery: current insights. *Adv. Genomics Genet.* **7**, 29–45 (2017).
82. Skewes-Cox, P., Sharpton, T. J., Pollard, K. S. & DeRisi, J. L. Profile Hidden Markov Models for the Detection of Viruses within Metagenomic Sequence Data. *PLOS ONE* **9**, e105067 (2014).
83. Oliveira, J. H., Bahia, A. C. & Vale, P. F. How are arbovirus vectors able to tolerate infection? *Dev. Comp. Immunol.* **103**, 103514 (2020).
84. Maqbool, M. *et al.* Potential Mechanisms of Transmission of Tick-Borne Viruses at the Virus-Tick Interface. *Front. Microbiol.* **13**, (2022).
85. Salvati, M. V. *et al.* Virus-derived DNA forms mediate the persistent infection of tick cells by Hazara virus and Crimean-Congo hemorrhagic fever virus. *J. Virol.* JVI0163821 (2021)
doi:10.1128/JVI.01638-21.


86. Lyden, A., Crawford, E., Quan, J., Caldera, S. & Dynerman, D. DASH Protocol. *protocols.io*
<https://www.protocols.io/view/dash-protocol-6rjhd4n> (2019).
87. Ring, K. *et al.* Host blood meal identity modifies vector gene expression and competency. *Mol. Ecol.* **31**, 2698–2711 (2022).
88. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
89. Kalantar, K. L. *et al.* IDseq-An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. *GigaScience* **9**, giaa111 (2020).
90. Wu, H. *et al.* Abundant and Diverse RNA Viruses in Insects Revealed by RNA-Seq Analysis: Ecological and Evolutionary Implications. *mSystems* (2020) doi:10.1128/mSystems.00039-20.
91. Li, C.-X. *et al.* Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *eLife* **4**,.
92. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
93. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
94. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
95. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
96. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).
97. Chao, A. Nonparametric Estimation of the Number of Classes in a Population. *Scand. J. Stat.* **11**, 265–270 (1984).

98. MJ. fossil: palaeoecological and palaeogeographical analysis tools. (2011).
99. Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal Complex Systems*, 1695 (2006).
100. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
101. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
102. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
103. Mihara, T. *et al.* Linking Virus Genomes with Host Taxonomy. *Viruses* **8**, 66 (2016).
104. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Mach. Learn. PYTHON* 6.
105. Kholodilov, I. S. *et al.* Isolation and Characterisation of Alongshan Virus in Russia. *Viruses* **12**, (2020).
106. Vandegrift, K. J. *et al.* Presence of Segmented Flavivirus Infections in North America. *Emerg. Infect. Dis.* **26**, 1810–1817 (2020).
107. Kumar, A. *et al.* Early Release - Virome Analysis for Detection of Segmented Flaviviruses in Wild Rodents, Pennsylvania, USA - Volume 26, Number 8—August 2020 - Emerging Infectious Diseases journal - CDC. doi:10.3201/eid2608.190986.
108. Edgar, R. C. *et al.* Petabase-scale sequence alignment catalyses viral discovery. *Nature* **602**, 142–147 (2022).
109. Neri, U. *et al.* A five-fold expansion of the global RNA virome reveals multiple new clades of RNA bacteriophages. 70.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:

052A90EE04BF4C7... Author Signature

12/13/2022
Date