DISSERTATION

EXPLORING THE EVOLUTION OF A PERSISTENT VIRAL INFECTION OF *DROSOPHILA MELANOGASTER* ACROSS TIME AND SPACE

Submitted by

Alexandra Harrington Keene-Snickers

Department of Cell and Molecular Biology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2025

Doctoral Committee:

Advisor: Mark Stenglein

Zaid Abdo
Brian Foy
Meena Balgopal

ABSTRACT

EXPLORING THE EVOLUTION OF A PERSISTENT VIRAL INFECTION OF *DROSOPHILA MELANOGASTER* ACROSS TIME AND SPACE

Efforts to characterize the viral communities of organisms across the tree of life have uncovered diversity that is magnitudes greater than previously thought. It is now understood that organisms can be infected with many viruses at once and that many of these infections have limited or no impact on the hosts health. Some of these infections are persistent, lasting the lifetime of the host. The basic biology and evolution that allow these viruses to persist are not well understood but are important to create a more realistic picture of host-virus interactions. In this dissertation I have employed the model organism *Drosophila melanogaster* and its viral symbiont, galbut virus, to study the evolution of persistent viral infections. In Chapter 1, I provided a brief background on current methods to study viruses over time and space. I also discussed the characteristics of persistent viral infections as they are currently understood. Finally, I provided rationale for the use of the *D. melanogaster*-galbut virus model as an appropriate system to study persistent viral infections and their evolution. In Chapter 2, I utilized old *Drosophila* museum specimens to study viral and host RNA from samples over 100-years-old. This work shed light on the evolutionary history of galbut virus and revealed that RNA survives in dried samples better than prevailing dogma would suggest. In Chapter 3, I developed two fly traps for the effective capture of wild *D. melanogaster* by untrained participatory scientists. These traps were instrumental to the

collection of wild specimens for my work in Chapter 4. In Chapter 4, I broadly sampled flies across the USA and generated several hundred new galbut virus sequences. This enabled me to characterize galbut virus diversity, evolutionary patterns, and genotype-phenotype associations. In summary, my dissertation explores the evolution of a persistent viral infection using virus sequences diverse in time and space. The methods employed provide insight into the forces at work in persistent viral infections.

ACKNOWLEDGEMENTS

I want to express my sincerest gratitude to all of those who have made this PhD possible. This has been an incredible journey filled with learning and self-discovery, and I am incredibly grateful for the support, patience and guidance from so many people over the past five years. Without them, I truly would not have been able to reach this milestone.

First, I'd like to thank Dr. Mark Stenglein for allowing me to join his lab during the Covid-19 pandemic and trusting that I was a good fit for his lab despite my short rotation. Your guidance has been exactly what I needed to make it through this phase of my life. Thank you for giving me the freedom to pursue my own questions and letting me take on a project we had no idea would be successful. Thank you for the countless hours working on my first publication and for not failing to see the bright side when we received criticism. Thank you for reminding me to maintain a growth mindset and to treat each experience as a learning opportunity not only as a scientist but also as a person. I am confident that your mentorship has made me a more thoughtful, involved and confident scientist.

I would also like to thank the entire Stenglein lab both past and present. Thank you to Ali Brehm for keeping me on my toes and for commiserating with me through the highs and lows. Thank you, Tillie Dunham for being a light in the lab and for providing countless laughs especially when I first joined the lab. I'd also like to thank Taylor Gelpi for letting me learn how to be a good mentor through you. It was a pleasure to see you grow as a scientist. I would also like to thank Victoria Nieciecki for the

countless hours spent at coffee shops working and/or chatting. I treasure your friendship and am so happy that we found each other through CMB.

I am grateful to the rest of my committee, Dr. Zaid Abdo, Dr. Brian Foy and Dr. Meena Balgopal. Dr. Balgopal, thank you for asking the big picture questions and keeping me thinking about the “so, what” of my projects. Thank you, Dr. Foy for asking tough questions and preparing me to advocate for my science. And, thank you, Dr. Abdo for your help learning how to interpret phylogenetic data and being an easy to approach member of my committee.

This PhD would not have been possible without the support of all those outside my academic circle. First, I want to thank my husband Brandon Snickers for all his love and unrelenting commitment to helping me through this process. Through good days and bad days, he has kept me fed, watered and walked. This journey was a big step not only for me but also Brandon and I am thankful every day that I have someone so caring, down to earth and generous by my side to take on the world with. Thank you to my parents, Lisa Harrington and Paul Keene for always believing me and being there when I needed a shoulder to cry and/or scream into. I could not have done this without your positive words and ability to listen to me rant about all sorts of things. You have ensured that I believe in myself as much as you believe in me. Thank you to my in-laws Karne and John Snickers. Your love and acceptance into your family has made a tremendous impact on me. And finally, I’d like to thank my best friend Elle Irons for not only keeping me filled with delicious baked goods but keeping me grounded and aware that the world exists beyond my PhD. Thank you for inviting me to all the things, laughing with me and crying with me. I am so happy to have found you.

TABLE OF CONTENTS

CHAPTER 1: AN INTRODUCTION TO THE STUDY OF PERSISTENT VIRUS EVOLUTION AND THE *DROSOPHILA MELANOGASTER*-GALBUT VIRUS MODEL

This dissertation is focused on galbut virus, a remarkably successful viral symbiont of *Drosophila melanogaster*. There are two main parts to this dissertation. The first focuses on galbut virus evolution over a century. The second focuses on galbut virus evolution across the United States. My hope is that the research I contributed to during my PhD is informative on the biology and evolution of galbut virus. I also hope that this dissertation leads to an increased use of museum specimens to study old RNA and challenges the paradigm of virus evolution.

1.0 Why do we study virus evolution? Why does virus evolution matter?

Disease causing viruses have been infecting human, animal and plant cells since there has been cellular life, but the first virus identified, tobacco mosaic virus, was not described until 1898[1,2]. Since then, there has been considerable effort to characterize viral infections of human, animal and plant importance. With the rise of next-generation sequencing technology, it has become clear that viral abundance and diversity extends beyond disease-causing agents and is far higher than previously thought[3–7].

Viral genomes evolve rapidly. This happens due to errors during replication, response to pressures exacted by the host or after host switching[8,9]. It is important to study these events because understanding where and when viruses emerge or have emerged is of concern for the wellbeing of humans, animals and plants[10,11]. Studying virus evolution also contributes to our understanding of the mechanisms that lead to

disease. This includes how viruses evade the host immune response, particularly in reinfection events or when the host has vaccine-induced immunity[12–14]. The study of viral evolution is also informative for understanding how drug resistance occurs and how viruses become more or less virulent[15–17].

1.1 Why study virus evolution on expanded temporal and spatial scales?

The study of viruses generally relies on samples collected in the present. This is partly because there is a need to characterize viruses of immediate concern and because of the belief that viral genomes, particularly RNA ones, do not survive long in the environment. Evolution is fundamentally a time-dependent phenomenon. The study of old viral sequences is valuable but has been mostly limited to samples stored in fixed tissues, frozen in biobanks or in natural environments like permafrost[18–21]. Despite their limited availability, studies of such samples have led to new models of virus evolution[22].

To get a more comprehensive view of virus evolution, it is also important to collect virus sequences from broad geographic ranges. Sequences obtained from diverse geographic areas provide information on the emergence of new strains, virus movement into populations and can even be used to inform therapeutic and vaccine development[12,21,23]. A concerted effort to sample from diverse locations and into the past can elucidate the evolutionary forces at work in virus populations.

1.1.1 Viral genomic epidemiology

Viral genomic epidemiology leverages genomic data and metadata to track a virus's spread temporally and spatially[24]. Its use in response to viral outbreaks has been instrumental to understanding how a virus is moving through a population and whether a particular public health intervention could be effective. The combination of genomic

data along with metadata makes it possible to tease apart how host, virus and environmental factors impact the spread of a virus within a population[23,25]. The use of sequences recovered from long timescales and in different regions is important to inform viral genomic epidemiology efforts.

1.1.2 Context for the current study of viral evolution.

Viruses that cause disease in humans and organisms of economic importance are highly studied and are thus often used to model viral evolution. Current tools and technologies have made it easy to acquire, test and sequence many viral genomes quickly and cheaply[26,27]. This has led to an abundance of data on viruses of particular concern. For example, there are millions of SARS-CoV-2 genomes available through repositories such as The National Center for Biotechnology Information (NCBI) and the Global Initiative on Sharing All Influenza Data (GISAID). These sequences have been instrumental to inform public health responses and therapy development and as a real-time study of viral evolution after a recent host jump[28–30].

1.1.3 Limitations to studying viruses on short time scales

Although sequencing many genomes from as many samples as possible during a pandemic or local outbreak is helpful, these short-term studies miss valuable insights into overall virus evolution. They capture mutations that may be short-lived on longer evolutionary time scales[31,32]. Additionally, most sequences derive from few viruses (e.g., SARS-CoV-2, influenza viruses, and human immunodeficiency viruses), meaning that evolutionary rates extrapolated to other viruses may be distorted and not represent general virus biology. Selection pressures are different in viruses that cause acute disease, persistent infection or infections with minimal impact on the host[8,33–36]. Thus,

more sequences from diverse viral families and from different points in time are needed to study overall viral evolution.

Short-term evolutionary estimates suggest that viruses, particularly RNA ones evolve rapidly[37,38]. However, estimates that include virus sequences from old samples such as those found as endogenous retrovirus elements (ERVs) and endogenous viral elements (EVEs) have found that long-term evolutionary rates are orders of magnitudes lower than short-term rates[39,40]. This concept has been labeled the time-dependent rate phenomenon (TDRP)[41,42]. These data have shown that viruses we once thought had only been around for a couple hundred years, could be traced back thousands or millions of years[22,41,43]. For the purposes of the virus studied in this thesis, it is important to note that previous studies of the TDRP did not include rate estimates for double-stranded (ds) RNA viruses[22].

The prisoner of war hypothesis, developed in parallel with the TDRP, seeks to explain slow long-term virus evolutionary rates. Simmonds et. al. suggest that the constraints put on a virus by its host are what lead to steep declines of evolutionary rates when old sequences are included in analyses[44]. These constraints can be more easily seen in viruses that have been infecting their host for a long time and are thus host adapted. For some persistent viral infections this hypothesis makes sense. For example, viruses that rely on vertical transmission to pass to new hosts need their host to survive and reproduce to be successful. Any change that leads to the loss of fitness in the host, particularly in reproduction, will decrease viral fitness. Only studying the evolution of a few viruses on short timescales does not reveal the whole picture.

1.1.4 Tools to study virus evolution on an expanded temporal scale

To increase our overall understanding of the evolutionary forces at work in viruses it is necessary to sample virus sequences from the past. The validity of ancient DNA (aDNA) has been generally accepted by the scientific community and great effort has been made to study the evolution of organisms across the tree of life using old samples[45–48]. Even with the success of aDNA, there is less evidence that old RNA might be similarly useful.

DNA from old or ancient specimens can generally be studied because it is expected to survive over long periods, and methods exist to work with highly fragmented DNA[49–52]. DNA from historic and ancient specimens have been used to elucidate the origins and evolution of variola viruses, hepatitis B viruses and viral communities in great ape museum specimens[53–55]. EVEs, endogenized viral sequences, that become incorporated into the hosts genome, are another way to study old virus sequences. Viruses that have been studied this way include parvovirus 19 and herpes simplex virus-1[49,56]. Finally, endogenized retroviral sequences are common in eukaryotic genomes and, ancient retroviruses have been studied using DNA obtained from museum specimens[57]. Retrovirus evolution can also be studied using ERVs such as in the case of foamy viruses and lentiviruses using contemporary specimens[39,58,59]. If the ERV is located in a region of the genome that results in replication, the substitution rate will now align with the hosts evolutionary rate.

When it comes to studying RNA viruses using old RNA, researchers are more cautious. Techniques to sequence old RNA virus genomes generally rely on rare specimens that have been frozen or fixed[18,20,60,61]. Specimens from herbaria have also

been used to study historic and ancient RNA viruses[61–64]. Many of these studies rely on RNA stored in seeds because RNA needed for germination is thought to be protected from degradation[62,65]. Although these studies are important to understanding viral evolution, there is a significant gap in representation of diverse viral families.

There is also a general misconception about the extent to which RNA survives over long periods. With the success of herbaria specimens in studying old RNA, it is worth exploring other museum specimen types as potential reservoirs of RNA. Museum collections hold millions of specimens and writing off these specimens as not suitable for RNA extraction and sequencing limits progress[66]. There are characteristics of RNA that permit survival over time including RNA secondary and tertiary structure, specimen type, location of the RNA in the specimen and the environment the specimen was stored[67]. Retrieving RNA from specimens that are 100 or more years old would give insight into viral evolution over longer periods.

1.1.5 Tools to study virus evolution on an expanded spatial scale

It is useful to study viruses from diverse regions. Specimen collection often requires workers who are highly trained and, in the area, where the virus of interest is found. These specimens then rely on complex transport chains to make their way back to laboratories for diagnostic and research purposes[68,69].

Some ways that researchers have begun to address the difficulties of transporting specimens between local, national or international borders are tools such as blood spot cards and storage in RNA later reagents[70–72]. Although these tools are effective, they require collectors to be trained on sample procurement and handling. There has been some movement promoting sample collection by the infected individual,

including SARS-CoV-2, human immunodeficiency virus (HIV) and human papillomavirus (HPV)[73,74].

Making specimen collection more accessible, particularly by non-scientists makes it easier for specimens to be collected from many places at large scale. One option that has continued to gain traction is the use of participatory scientists[75,76]. Volunteer scientists have been crucial for ecological studies and websites like SciStarter.org and CitSci.org aim to make participatory science easy and fun[77,78]. Participatory scientists could be used to collect samples from all over the world. Not all viruses could be studied this way due to health and safety concerns, but whole organisms such as ticks have been sent to scientists before [79].

1.2 Persistent viral infections

1.2.1 What is a persistent viral infection?

Persistent viral infections are classically defined as acute viral infections that were not cleared by the infected host[80,81]. The virus goes from an active state of replication, assembly and spread during acute infection to an inactive state through endogenization in the hosts genome or by sequestration into episomes. Persistent viral infections are classified as latent, asymptomatic or pathogenic and can shift between types depending on the circumstances of the infected individual[80,81]. There is emerging evidence of a fourth type of persistent viral infection, one that bypasses the acute phase, continually replicating and lasting the lifetime of the host[82–84].

Persistent viral infections such as human immunodeficiency virus and herpes viruses in humans are widely studied due to their impact on human health and their abundance within human populations[85]. The impact these infections have on their host

varies widely and is a factor of host genetics, host biology, virus biology, interactions with other microbiota and the environment[85–87].

1.2.2 How do persistent arthropod viral infections impact their host?

Efforts to characterize the virome of many organisms including arthropods has uncovered an unexpected diversity of viruses[4,88–93]. The discovery of these infections in ticks, mosquitoes and fruit flies have led to questions of what these viruses are doing in their host and how they interact with other microbiota in their host. Through comparative studies it is becoming clear that at least some of these viruses play a role in the abundance of and transmission of viruses of human and agricultural importance[94–100]. Some viral infections have been shown to modulate or are modulated by other viruses present in the host[101–103]. And the presence of bacterial endosymbionts such as Wolbachia, can modulate the interactions between the host and virus, the host and Wolbachia and the virus and Wolbachia[104–106]. The fact that arthropods do not have an adaptive immune response is also an important factor in understanding the relationship between virus and host[107,108].

1.2.3 Evolution of persistent viruses

Overall, persistent viral infections follow the evolutionary patterns of acute viral infections. On the short-term the evolutionary rate estimates follow estimates based on their Baltimore Classification. On the long-term they appear to be constrained by the TDRP and prisoner of war hypothesis. These viruses balance replication and dissemination while evading the immune system or by purposefully inhibiting the immune response. However, persistently infecting double-stranded (ds) RNA viruses of fungi, plants, protozoa and arthropods show lower than expected evolutionary

rates[109,110]. The factors that lead to these different patterns are still unknown. It could be that the close relationship between host and virus has constrained the ability of the virus to change without conferring fitness costs to the virus itself or the host that ultimately leads to decreased virus spread.

1.3 *Drosophila melanogaster* and galbut virus as a model to study persistent viral infections.

1.3.1 Partitiviruses are common infections of arthropods

Members of the *Partitiviridae* family are small (3-4.8 kbp), non-enveloped, segmented, dsRNA viruses that infect a wide range of hosts[111–113]. Their genome is characterized by at least two segments, RNA 1 encodes the RNA-dependent RNA polymerase protein and RNA 2 encodes the coat protein[112,113]. Some partitiviruses have been described with a third or fourth segment for which functional roles have yet to be described[113]. Replication occurs in the cytoplasm and is monocistronic with translation occurring on the positive-sense strand of the dsRNA[112].

Partitiviruses have been most widely studied as infections of fungi, plants and protozoa[113–115]. However, the number of partitiviruses associated with arthropods is increasing[116–118]. Partitiviruses or partiti-like viruses have now been described in *Hemiptera*, *Thysanoptera*, *Lepidoptera*, *Acari* and *Diptera*[89,94,96,119,120]. Arthropod infecting partitiviruses are thought to have limited impact on their host. However, three partiti-like viruses described from *Spodoptera exempta* were shown to slow growth and decrease reproduction. In non-natural hosts, they led to increased mortality in larvae and pupae and susceptibility to pathogenic viruses[116,121]. Additionally, male-killing has

been reported in *D. melanogaster* and *Homona magnanima* by members of the *Partitiviridae*[118,122].

1.3.2 Galbut virus as a model partitivirus and model persistent virus

First described in 2015, galbut virus is a partitivirus that infects several members of the *D. melanogaster* subgroup[100,123]. Galbut virus is a three segmented partitivirus for which the role of RNA 3 is unknown[117,123]. Chaq virus, described at the same time as galbut virus is thought to be an optional fourth segment or satellite virus as it is only found with galbut virus but is not required for infection[117]. Galbut virus is ubiquitous in wild *D. melanogaster* populations: there is yet to be a wild population found without it. Our lab's recent sampling of ~1000 wild-caught flies from across the USA found that ~67% of flies were infected by galbut virus.

Like other partitiviruses, galbut virus is vertically transmitted from either parent to offspring at near 100% efficiency[117]. It also appears to have minimal impact on the host. One study found no significant difference in life span, fecundity or the microbiome consitituents[124]. That same study found that in at least one laboratory strain, galbut virus infected flies had a decrease in time to pupation and eclosion but that there was some effect on the ability of infected flies to fight off fungal infections[124]. These results were highly strain and sex dependent[124]. Another study found that galbut virus was associated with a 23% average decrease in lifespan but agreed that there was no effect on fecundity[125].

This data and additional unpublished data from our lab suggest that there is conflict between galbut virus and *D. melanogaster*. In stocks that we have collected and maintained since 2017, there is a decline in galbut virus prevalence. Additionally, the

mechanism of galbut virus transmission suggests that infection should reach 100% prevalence. However, when we have sampled the same populations over several years, the prevalence is variable. Additionally, there is evidence of positive selection acting on regions of the galbut virus genome, something that has been found in other partitiviruses[123,126]. A paradox my thesis seeks to explore is, how galbut virus can be so common despite the understanding that infection produces a fitness cost?

Knowing when galbut virus first began infecting flies is of evolutionary interest. One study has suggested that galbut virus has been infecting *D. melanogaster* for at least 200 years[123]. However, given its ubiquitous range, presence in *D. simulans* and *D. suzukii* and the endogenization of a near complete EVE of the RNA 1 genome in some European populations, it is likely galbut virus has been infecting flies for much longer[100,123,127]. These conflicting findings prompt the question, what are the evolutionary forces at work that make it appear that the virus is simultaneously only 200 years old and under selective pressure but also globally distributed and endogenized in some populations? Additionally, as a dsRNA virus, the study of galbut virus evolution on a large timescale would fill in gaps missing in the TDRP[22].

1.3.3 The *Drosophila melanogaster* and galbut virus model system

D. melanogaster has been used as a model organism since the early 1900s[128,129]. This fruit fly has been instrumental to our understanding of genetics, evolution and infectious disease[99,130–132]. It has even been used to recapitulate neurodegenerative diseases, cancers and metabolic diseases of humans[133–135]. The biology of wild populations is also of interest and efforts to fully understand its natural biology have been increasing. The natural *D. melanogaster* microbiome including

bacteria, viruses and fungi are well-characterized[90,123,127,136–139]. Efforts have also been made to understand how the interactions of pathogenic microbiota and standard microbiota interact with the host and other microbiota[98,140–144].

One powerful feature of *D. melanogaster* is that wild flies are easy to capture[145,146]. The most basic traps require just rotting fruit, yeast and a bucket[147]. Wild caught flies can be brought to the lab and maintained on standard fly food and/or used for downstream experiments. The molecular and genetic tools available due to *D. melanogaster's* use as a model organism and the abundance of galbut virus in wild populations make this system excellent for studying the biology and evolution of persistent viral infections.

1.4 Summary of the work contained in this dissertation

In this thesis, I used the galbut virus – *D. melanogaster* model to study the evolutionary forces at work in a persistent viral infection on expanded spatial and temporal scales. I first leveraged the use of old, insect museum specimens as a source of RNA to study persistent viral evolution over more than a century. I then developed easy to use and cost-effective fly traps for the collection of wild *D. melanogaster*. Finally, I collaborated with volunteers to collect flies from local and national locations and used them to study galbut virus diversity across a large geographic range.

CHAPTER 1: REFERENCES

1. Taylor, M. W. Introduction: A Short History of Virology. in *Viruses and Man: A History of Interactions* 1–22 (Springer International Publishing, 2014). doi:10.1007/978-3-319-07758-1_1.
2. Rowlands, D. J. A Brief History of Virology. in *Encyclopedia of Virology* 3–13 (Elsevier, 2021). doi:10.1016/b978-0-12-814515-9.00022-9.
3. Wu, Q. *et al.* Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. *Proc Natl Acad Sci U S A* **107**, 1606–1611 (2010).
4. Li, C.-X. *et al.* Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. doi:10.7554/eLife.05378.001.
5. Abeles, S. R. & Pride, D. T. Molecular Bases and Role of Viruses in the Human Microbiome. *J Mol Biol* **426**, 3892–3906 (2014).
6. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat Microbiol* **6**, 960–970 (2021).
7. Zeigler Allen, L. *et al.* The Baltic Sea Virome: Diversity and Transcriptional Activity of DNA and RNA Viruses. *mSystems* **2**, (2017).
8. Duffy, S., Shackelton, L. A. & Holmes, E. C. Rates of evolutionary change in viruses: Patterns and determinants. *Nature Reviews Genetics* vol. 9 267–276 Preprint at https://doi.org/10.1038/nrg2323 (2008).
9. Jenkins, G. M., Rambaut, A., Pybus, O. G. & Holmes, E. C. Rates of molecular evolution in RNA viruses: A quantitative phylogenetic analysis. *J Mol Evol* **54**, 156–165 (2002).
10. Holmes, E. C. What Does Virus Evolution Tell Us about Virus Origins? *J Virol* **85**, 5247–5251 (2011).
11. Neumann, G. & Kawaoka, Y. Which Virus Will Cause the Next Pandemic? *Viruses* **15**, 199 (2023).
12. Hannoun, C. The evolving history of influenza viruses and influenza vaccines. *Expert Rev Vaccines* **12**, 1085–1094 (2013).
13. Yoo, S. J. *et al.* Genetic evolution of classical swine fever virus under immune environments conditioned by genotype 1-based modified live virus vaccine. *Transbound Emerg Dis* **65**, 735–745 (2018).
14. Fitch, W. M., Leiter, J. M., Li, X. Q. & Palese, P. Positive Darwinian evolution in human influenza A viruses. *Proceedings of the National Academy of Sciences* **88**, 4270–4274 (1991).
15. Irwin, K. K., Renzette, N., Kowalik, T. F. & Jensen, J. D. Antiviral drug resistance as an adaptive process. *Virus Evol* **2**, vew014 (2016).
16. Moscona, A. Global Transmission of Oseltamivir-Resistant Influenza. *New England Journal of Medicine* **360**, 953–956 (2009).

17. Geoghegan, J. L. & Holmes, E. C. The phylogenomics of evolving virus virulence. *Nat Rev Genet* **19**, 756–769 (2018).
18. Taubenberger, J. K., Reid, A. H., Krafft, A. E., Bijwaard, K. E. & Fanning, T. G. Initial Genetic Characterization of the 1918 “Spanish” Influenza Virus. *Science (1979)* **275**, 1793–1796 (1997).
19. Castello, J. D. *et al. Detection of Tomato Mosaic Tobamovirus RNA in Ancient Glacial Ice*.
20. Smith, A. W., Skilling, D. E., Castello, J. D. & Rogers, S. O. Ice as a reservoir for pathogenic human viruses: Specifically, caliciviruses, influenza viruses, and enteroviruses. *Med Hypotheses* **63**, 560–566 (2004).
21. Faria, N. R. *et al.* The early spread and epidemic ignition of HIV-1 in human populations. *Science (1979)* **346**, 56–61 (2014).
22. Ghafari, M., Simmonds, P., Pybus, O. G. & Katzourakis, A. A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses. *Current Biology* **31**, 4689-4696.e5 (2021).
23. Pollett, S. *et al.* Genomic Epidemiology as a Public Health Tool to Combat Mosquito-Borne Virus Outbreaks. *J Infect Dis* **221**, S308–S318 (2020).
24. Hill, V., Ruis, C., Bajaj, S., Pybus, O. G. & Kraemer, M. U. G. Progress and challenges in virus genomic epidemiology. *Trends in Parasitology* vol. 37 1038–1049 Preprint at https://doi.org/10.1016/j.pt.2021.08.007 (2021).
25. Uddin, M. *et al.* SARS-CoV-2/COVID-19: Viral Genomics, Epidemiology, Vaccines, and Therapeutic Interventions. *Viruses* **12**, 526 (2020).
26. Svraka, S. *et al.* Metagenomic sequencing for virus identification in a public-health setting. *Journal of General Virology* **91**, 2846–2856 (2010).
27. Rahimian, M. & Panahi, B. Next generation sequencing-based transcriptome data mining for virus identification and characterization: Review on recent progress and prospects. *Journal of Clinical Virology Plus* **4**, 100194 (2024).
28. Williams, T. C. & Burgers, W. A. SARS-CoV-2 evolution and vaccines: cause for concern? *The Lancet Respiratory Medicine* vol. 9 333–335 Preprint at https://doi.org/10.1016/S2213-2600(21)00075-8 (2021).
29. Day, T., Gandon, S., Lion, S. & Otto, S. P. On the evolutionary epidemiology of SARS-CoV-2. *Current Biology* vol. 30 R849–R857 Preprint at https://doi.org/10.1016/j.cub.2020.06.031 (2020).
30. Markov, P. V. *et al.* The evolution of SARS-CoV-2. *Nature Reviews Microbiology* vol. 21 361–379 Preprint at https://doi.org/10.1038/s41579-023-00878-2 (2023).
31. Romano, C. M. *et al.* Inter- and Intra-Host Viral Diversity in a Large Seasonal DENV2 Outbreak. *PLoS One* **8**, e70318 (2013).
32. Lythgoe, K. A. *et al.* SARS-CoV-2 within-host diversity and transmission. *Science (1979)* **372**, (2021).

33. Drummond, A., Pybus, O. G. & Rambaut, A. Inference of viral evolutionary rates from molecular sequences. *Adv Parasitol* **54**, 331–58 (2003).
34. Shi, M., Zhang, Y.-Z. & Holmes, E. C. Meta-transcriptomics and the evolutionary biology of RNA viruses. *Virus Res* **243**, 83–90 (2018).
35. Harmon, L. J. *et al.* Causes and Consequences of Apparent Timescaling Across All Estimated Evolutionary Rates. (2021) doi:10.1146/annurev-ecolsys-011921.
36. Sanjuán, R. From molecular genetics to phylodynamics: Evolutionary relevance of mutation rates across viruses. *PLoS Pathog* **8**, (2012).
37. Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. Viral Mutation Rates. *J Virol* **84**, 9733–9748 (2010).
38. Peck, K. M. & Lauring, A. S. Complexities of Viral Mutation Rates. (2018) doi:10.1128/JVI.
39. Aiewsakun, P. Avian and serpentine endogenous foamy viruses, and new insights into the macroevolutionary history of foamy viruses. *Virus Evol* **6**, (2020).
40. Holmes, E. C. The evolution of endogenous viral elements. *Cell Host and Microbe* vol. 10 368–377 Preprint at https://doi.org/10.1016/j.chom.2011.09.002 (2011).
41. Aiewsakun, P. & Katzourakis, A. Time-Dependent Rate Phenomenon in Viruses. *J Virol* **90**, 7184–7195 (2016).
42. Duchêne, S., Holmes, E. C. & Ho, S. Y. W. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proceedings of the Royal Society B: Biological Sciences* **281**, 20140732 (2014).
43. Ho, S. Y. W. *et al.* Time-dependent rates of molecular evolution. *Molecular Ecology* vol. 20 3087–3101 Preprint at https://doi.org/10.1111/j.1365-294X.2011.05178.x (2011).
44. Simmonds, P., Aiewsakun, P. & Katzourakis, A. Prisoners of war — host adaptation and its constraints on virus evolution. *Nat Rev Microbiol* **17**, 321–328 (2019).
45. Pääbo, S. *et al.* Genetic analyses from ancient DNA. *Annual Review of Genetics* vol. 38 645–679 Preprint at https://doi.org/10.1146/annurev.genet.37.110801.143214 (2004).
46. Leonardi, M. *et al.* Evolutionary Patterns and Processes: Lessons from Ancient DNA. *Syst Biol* **66**, syw059 (2016).
47. Orlando, L. & Cooper, A. Using ancient DNA to understand evolutionary and ecological processes. *Annu Rev Ecol Evol Syst* **45**, 573–598 (2014).
48. Kefi, R. Ancient DNA investigations: A review on their significance in different research fields. *International Journal of Modern Anthropology* **1**, (2011).
49. Ávila-Arcos, M. C. *et al.* Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. *Sci Rep* **1**, (2011).
50. Taylor, G. M., Mays, S. A. & Huggett, J. F. Ancient DNA (aDNA) studies of man and microbes: General similarities, specific differences. *International Journal of Osteoarchaeology* vol. 20 747–751 Preprint at https://doi.org/10.1002/oa.1077 (2010).
51. Orlando, L. *et al.* Ancient DNA analysis. *Nature Reviews Methods Primers* vol. 1 Preprint at https://doi.org/10.1038/s43586-020-00011-0 (2021).

52. Dabney, J., Meyer, M. & Pääbo, S. Ancient DNA damage. *Cold Spring Harb Perspect Biol* **5**, a012567 (2013).
53. Mühlemann, B. *et al.* Diverse variola virus (smallpox) strains were widespread in northern Europe in the Viking Age. *Science (1979)* **369**, (2020).
54. Kocher, A. *et al.* Ten millennia of hepatitis B virus evolution. *Science (1979)* **374**, 182–188 (2021).
55. Hämmerle, M. *et al.* Screening great ape museum specimens for DNA viruses. *Sci Rep* **14**, (2024).
56. Mühlemann, B. *et al.* Ancient human parvovirus B19 in Eurasia reveals its long-term association with humans. *Proceedings of the National Academy of Sciences* **115**, 7557–7562 (2018).
57. Ávila-Arcos, M. C. *et al.* One hundred twenty years of koala retrovirus evolution determined from museum Skins. *Mol Biol Evol* **30**, 299–304 (2013).
58. Cui, J. & Holmes, E. C. Endogenous Lentiviruses in the Ferret Genome. *J Virol* **86**, 3383–3385 (2012).
59. Wertheim, J. O. & Worobey, M. Dating the Age of the SIV Lineages That Gave Rise to HIV-1 and HIV-2. *PLoS Comput Biol* **5**, e1000377 (2009).
60. Zhang, G. *et al.* Evidence of Influenza A Virus RNA in Siberian Lake Ice. *J Virol* **80**, 12229–12235 (2006).
61. Castello, J. D. *et al. Detection of Tomato Mosaic Tobamovirus RNA in Ancient Glacial Ice*. (1999).
62. Smith, O. *et al.* A complete ancient RNA genome: Identification, reconstruction and evolutionary history of archaeological Barley Stripe Mosaic Virus. *Sci Rep* **4**, 4003 (2014).
63. Rollo, F. *Characterisation by Molecular Hybridization of RNA Fragments Isolated from Ancient (1400 B.C.) Seeds*. *Theor Appl Genet* vol. 71 (1985).
64. Peyambari, M., Warner, S., Stoler, N., Rainer, D. & Roossinck, M. J. A 1,000-Year-Old RNA Virus. *J Virol* **93**, e01188-18 (2019).
65. Fordyce, S. L. *et al.* Deep Sequencing of RNA from Ancient Maize Kernels. *PLoS One* **8**, e50961 (2013).
66. Johnson, K. R., Owens, I. F. P. & null null. A global approach for natural history museum collections. *Science (1979)* **379**, 1192–1194 (2023).
67. Friedländer, M. R. & Gilbert, M. T. P. How ancient RNA survives and what we can learn from it. *Nature Reviews Molecular Cell Biology* vol. 25 417–418 Preprint at https://doi.org/10.1038/s41580-024-00726-y (2024).
68. Spackman, E., Pedersen, J. C., McKinley, E. T. & Gelb, J. Optimal specimen collection and transport methods for the detection of avian influenza virus and Newcastle disease virus. *BMC Vet Res* **9**, 35 (2013).
69. Zaman, F. A. *et al.* Exploratory study on the operational issues faced in collection, transportation, and laboratory testing related to COVID-19 in remote areas of selected

EAG states of North East and East India. *J Family Med Prim Care* **10**, 1443–1452 (2021).
70. Heneghan, N., Fu, J., Pritchard, J., Payton, M. & Allen, R. W. The effect of environmental conditions on the rate of RNA degradation in dried blood stains. *Forensic Sci Int Genet* **51**, (2021).
71. Fauver, J. R. *et al.* Xenosurveillance reflects traditional sampling techniques for the identification of human pathogens: A comparative study in West Africa. *PLoS Negl Trop Dis* **12**, (2018).
72. Mutter, G. L. *et al.* Comparison of frozen and RNALater solid tissue storage methods for use in RNA expression microarrays. *BMC Genomics* **5**, (2004).
73. Saberi, P. *et al.* Feasibility and Acceptability of Home-Collected Samples for Human Immunodeficiency Virus Preexposure Prophylaxis and Severe Acute Respiratory Syndrome Coronavirus 2 Laboratory Tests in San Francisco Primary Care Clinics. *Open Forum Infect Dis* **9**, (2022).
74. Garg, S. *et al. TYPE Systematic Review PUBLISHED December Rrrr DOI OPEN ACCESS EDITED BY REVIEWED BY Does Self-Sampling for Human Papilloma Virus Testing Have the Potential to Increase Cervical Cancer Screening? An Updated Meta-Analysis of Observational Studies and Randomized Clinical Trials*. www.mendeley.com.
75. Fraisl, D. *et al.* Citizen science in environmental and ecological sciences. *Nature Reviews Methods Primers* **2**, 64 (2022).
76. Newman, G., Graham, J., Crall, A. & Laituri, M. The art and science of multi-scale citizen science support. *Ecological Informatics* vol. 6 217–227 Preprint at https://doi.org/10.1016/j.ecoinf.2011.03.002 (2011).
77. Hoffman, C., Cooper, C. B., Kennedy, E. B., Farooque, M. & Cavalier, D. SciStarter 2.0. in 50–61 (2017). doi:10.4018/978-1-5225-0962-2.ch003.
78. Wang, Y., Kaplan, N., Newman, G. & Scarpino, R. CitSci.org: A New Model for Managing, Documenting, and Sharing Citizen Science Data. *PLoS Biol* **13**, e1002280 (2015).
79. Porter, W. T. *et al.* Citizen science informs human-tick exposure in the Northeastern United States. *Int J Health Geogr* **18**, 9 (2019).
80. Fields, B. N., Knipe, D. M. & Howley, P. M. *Fields Virology*. (Wolters Kluwer Health/Lippincott Williams & Wilkins, c2013., Philadelphia, 2013).
81. Boldogh, I., Albrecht, T. & Porter, D. D. *Persistent Viral Infections*. (Galveston, TX, 1996).
82. Bhagchandani, T., Nikita, Verma, A. & Tandon, R. Exploring the Human Virome: Composition, Dynamics, and Implications for Health and Disease. *Curr Microbiol* **81**, 16 (2024).
83. Yan, M. & Yu, Z. Viruses contribute to microbial diversification in the rumen ecosystem and are associated with certain animal production traits. *Microbiome* **12**, 82 (2024).

84. Alfonso, P., Butković, A., Fernández, R., Riesgo, A. & Elena, S. F. Unveiling the hidden viromes across the animal tree of life: insights from a taxonomic classification pipeline applied to invertebrates of 31 metazoan phyla. *mSystems* **9**, (2024).
85. Kane, M. & Golovkina, T. Common Threads in Persistent Viral Infections. *J Virol* **84**, 4116–4123 (2010).
86. Sullivan, B. M., Teijaro, J. R., de la Torre, J. C. & Oldstone, M. B. A. Early Virus-Host Interactions Dictate the Course of a Persistent Infection. *PLoS Pathog* **11**, 1–14 (2015).
87. Virgin, H. W., Wherry, E. J. & Ahmed, R. Redefining Chronic Viral Infection. *Cell* vol. 138 30–50 Preprint at https://doi.org/10.1016/j.cell.2009.06.036 (2009).
88. Identification of diverse viruses associated with grasshoppers unveils phylogenetic congruence between hosts and viruses. doi:10.1101/2021.11.16.468783.
89. Qi, Y. H., Ye, Z. X., Zhang, C. X., Chen, J. P. & Li, J. M. Diversity of RNA viruses in agricultural insects. *Computational and Structural Biotechnology Journal* vol. 21 4312–4321 Preprint at https://doi.org/10.1016/j.csbj.2023.08.036 (2023).
90. Webster, C. L., Longdon, B., Lewis, S. H. & Obbard, D. J. Twenty-Five New Viruses Associated with the Drosophilidae (Diptera). *Evolutionary Bioinformatics* **12s2**, EBO.S39454 (2016).
91. Wu, H. *et al. Abundant and Diverse RNA Viruses in Insects Revealed by RNA-Seq Analysis: Ecological and Evolutionary Implications*. https://journals.asm.org/journal/msystems (2020).
92. de Oliveira Ribeiro, G. *et al.* Aedes aegypti from amazon basin harbor high diversity of novel viral species. *Viruses* **12**, (2020).
93. Shi, M. *et al.* Redefining the invertebrate RNA virosphere. *Nature* **540**, 539–543 (2016).
94. Harvey, E. *et al.* Extensive Diversity of RNA Viruses in Australian Ticks. *J Virol* **93**, (2019).
95. Shi, C. *et al.* Stable distinct core eukaryotic viromes in different mosquito species from Guadeloupe, using single mosquito viral metagenomics. *Microbiome* **7**, 121 (2019).
96. Ternovoi, V. A. *et al.* The Viromes of Mosquitoes from the Natural Landscapes of Western Siberia. *Viruses* **15**, (2023).
97. Coatsworth, H. *et al.* Intrinsic variation in the vertically transmitted core virome of the mosquito *Aedes aegypti*. *Mol Ecol* (2022) doi:10.1111/mec.16412.
98. Cogni, R., Ding, S. D., Pimentel, A. C., Day, J. P. & Jiggins, F. M. Wolbachia reduces virus infection in a natural population of Drosophila. *Commun Biol* **4**, 1327 (2021).
99. Hughes, T. T. *et al.* Drosophila as a genetic model for studying pathogenic human viruses. *Virology* vol. 423 1–5 Preprint at https://doi.org/10.1016/j.virol.2011.11.016 (2012).
100. Ortiz-Baez, A. S., Shi, M., Hoffmann, A. A. & Holmes, E. C. RNA virome diversity and Wolbachia infection in individual Drosophila simulans flies. *J Gen Virol* **102**, 001639 (2021).

101. Liang, G. & Bushman, F. D. The human virome: assembly, composition and host interactions. *Nature Reviews Microbiology* vol. 19 514–527 Preprint at https://doi.org/10.1038/s41579-021-00536-5 (2021).
102. Liu, X., Tang, Y., Chen, H., Liu, J.-X. & Sun, H.-Z. Rumen DNA virome and its relationship with feed efficiency in dairy cows. *Microbiome* **13**, 14 (2025).
103. Yu, M. *et al.* Diversity and potential host-interactions of viruses inhabiting deep-sea seamount sediments. *Nat Commun* **15**, 3228 (2024).
104. Martinez, J. *et al.* Virus evolution in wolbachia-infected drosophila. *Proceedings of the Royal Society B: Biological Sciences* **286**, (2019).
105. Serbus, L. R., Casper-Lindley, C., Landmann, F. & Sullivan, W. The Genetics and Cell Biology of Wolbachia-Host Interactions. *Annu Rev Genet* **42**, 683–707 (2008).
106. Martinez, J. *et al.* Symbionts Commonly Provide Broad Spectrum Resistance to Viruses in Insects: A Comparative Analysis of Wolbachia Strains. *PLoS Pathog* **10**, (2014).
107. Swevers, L., Broeck, J. Vanden & Smagghe, G. The possible impact of persistent virus infection on the function of the RNAi machinery in insects: A hypothesis. *Front Physiol* **4 NOV**, (2013).
108. Lee, W. S., Webster, J. A., Madzokere, E. T., Stephenson, E. B. & Herrero, L. J. Mosquito antiviral defense mechanisms: A delicate balance between innate immunity and persistent viral infection. *Parasit Vectors* **12**, (2019).
109. Botella, L., Vainio, E. J., Hantula, J., Diez, J. J. & Jankovsky, L. Description and prevalence of a putative novel mycovirus within the conifer pathogen Gremmeniella abietina. *Arch Virol* **160**, 1967–1975 (2015).
110. Thapa, V. *et al.* Using a Novel Partitivirus in Pseudogymnoascus destructans to Understand the Epidemiology of White-Nose Syndrome. *PLoS Pathog* **12**, e1006076 (2016).
111. Ghabrial, S. A., Ochoa, W. F., Baker, T. S. & Nibert, M. L. Partitiviruses: General Features. in *Encyclopedia of Virology (Third Edition)* (eds. Mahy, B. W. J. & Van Regenmortel, M. H. V) 68–75 (Academic Press, Oxford, 2008). doi:https://doi.org/10.1016/B978-012374410-4.00573-2.
112. Vainio, E. J. *et al.* ICTV virus taxonomy profile: Partitiviridae. *Journal of General Virology* **99**, 17–18 (2018).
113. Nibert, M. L. *et al.* Taxonomic reorganization of family Partitiviridae and other recent progress in partitivirus research. *Virus Res* **188**, 128–41 (2014).
114. Xiao, X. *et al.* A Novel Partitivirus That Confers Hypovirulence on Plant Pathogenic Fungi. *J Virol* **88**, 10120–10133 (2014).
115. Shimura, H., Kim, H., Matsuzawa, A., Akino, S. & Masuta, C. Coat protein of partitiviruses isolated from mycorrhizal fungi functions as an RNA silencing suppressor in plants and fungi. *Sci Rep* **12**, (2022).

116. Xu, P. *et al.* Novel partiti-like viruses are conditional mutualistic symbionts in their normal lepidopteran host, African armyworm, but parasitic in a novel host, Fall armyworm. *PLoS Pathog* **16**, (2020).
117. Cross, S. T. *et al.* Partitiviruses Infecting Drosophila melanogaster and Aedes aegypti Exhibit Efficient Biparental Vertical Transmission. *J Virol* **94**, (2020).
118. Fujita, R. *et al.* Late Male-Killing Viruses in Homona magnanima Identified as Osugoroshi Viruses, Novel Members of Partitiviridae. *Front Microbiol* **11**, (2021).
119. Oluwayelu, D. O., Karim, S., Du, J. & Sun, Y. *Diversity of RNA Viruses of Three Dominant Tick Species in North China*. www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi.
120. Fauver, J. R. *et al.* West African Anopheles gambiae mosquitoes harbor a taxonomically diverse virome including new insect-specific flaviviruses, mononegaviruses, and totiviruses. *Virology* **498**, 288–299 (2016).
121. Xu, P. *et al.* Partiti-like viruses from African armyworm increase larval and pupal mortality of a novel host: the Egyptian cotton leafworm. *Pest Manag Sci* **78**, 1529–1537 (2022).
122. Kageyama, D. *et al.* A male-killing gene encoded by a symbiotic virus of Drosophila. *Nat Commun* **14**, 1357 (2023).
123. Webster, C. L. *et al.* The Discovery, Distribution, and Evolution of Viruses Associated with Drosophila melanogaster. *PLoS Biol* **13**, e1002210 (2015).
124. Cross, S. T. *et al.* Galbut Virus Infection Minimally Influences Drosophila melanogaster Fitness Traits in a Strain and Sex-Dependent Manner. *Viruses* **15**, 539 (2023).
125. Wallace, M. A. & Obbard, D. J. Naturally occurring viruses of *Drosophila* reduce offspring number and lifespan. *Proceedings of the Royal Society B: Biological Sciences* **291**, (2024).
126. Petrzik, K. Evolutionary forces at work in partitiviruses. *Virus Genes* **55**, 563–573 (2019).
127. Wallace, M. A. *et al.* The discovery, distribution, and diversity of DNA viruses associated with *Drosophila melanogaster* in Europe. *Virus Evol* **7**, veab031 (2021).
128. Morgan, T. H. Sex Limited Inheritance in *Drosophila*. *Science (1979)* **32**, 120–122 (1910).
129. CASTLE, W. E. INBREEDING, CROSS-BREEDING AND STERILITY IN DROSOPHILA. *Science (1979)* **23**, 153–153 (1906).
130. Yamaguchi, M. & Yoshida, H. Drosophila as a Model Organism. in *Advances in Experimental Medicine and Biology* vol. 1076 1–10 (Springer New York LLC, 2018).
131. Markow, T. A. The secret lives of Drosophila flies. *Elife* **4**, (2015).
132. Beckingham, K. M., Armstrong, J. D., Texada, M. J., Munjaal, R. & Baker, D. A. Drosophila melanogaster--the model organism of choice for the complex biology of multi-cellular organisms. *Gravit Space Biol Bull* **18**, 17–29 (2005).

133. Bolus, H., Crocker, K., Boekhoff-Falk, G. & Chtarbanova, S. Modeling Neurodegenerative Disorders in Drosophila melanogaster. *Int J Mol Sci* **21**, 3055 (2020).
134. Mirzoyan, Z. *et al.* Drosophila melanogaster: A Model Organism to Study Cancer. *Front Genet* **10**, (2019).
135. Hoffmann, J., Romey, R., Fink, C. & Roeder, T. Drosophila as a Model to Study Metabolic Disorders. in 41–61 (2013). doi:10.1007/10_2013_196.
136. Kang, D. & Douglas, A. E. Functional traits of the gut microbiome correlated with host lipid content in a natural population of *Drosophila melanogaster*. *Biol Lett* **16**, (2020).
137. Adair, K. L., Wilson, M., Bost, A. & Douglas, A. E. Microbial community assembly in wild populations of the fruit fly *Drosophila melanogaster*. *ISME J* **12**, 959–972 (2018).
138. Wang, Y. *et al.* Common structuring principles of the <scp> *Drosophila melanogaster* </scp> microbiome on a continental scale and between host and substrate. *Environ Microbiol Rep* **12**, 220–228 (2020).
139. Palmer, W. H., Medd, N. C., Beard, P. M. & Obbard, D. J. Isolation of a natural DNA virus of Drosophila melanogaster, and characterisation of host resistance and immune responses. *PLoS Pathog* **14**, (2018).
140. Magwire, M. M. *et al.* Genome-Wide Association Studies Reveal a Simple Genetic Basis of Resistance to Naturally Coevolving Viruses in Drosophila melanogaster. *PLoS Genet* **8**, (2012).
141. Longdon, B., Wilfert, L. & Jiggins, F. M. *The Sigma Viruses of Drosophila from: Rhabdoviruses: Molecular Taxonomy, Evolution, Genomics, Ecology, Host-Vector Interactions, Cytopathology and Control*. (Caister Academic Press, U.K., 2012).
142. Shi, M. *et al.* No detectable effect of Wolbachia wMel on the prevalence and abundance of the RNA virome of Drosophila melanogaster. *Proceedings of the Royal Society B: Biological Sciences* **285**, (2018).
143. Mongelli, V. *et al.* Innate immune pathways act synergistically to constrain RNA virus evolution in Drosophila melanogaster. *Nat Ecol Evol* (2022) doi:10.1038/s41559-022-01697-z.
144. Cao, C., Magwire, M. M., Bayer, F. & Jiggins, F. M. A Polymorphism in the Processing Body Component Ge-1 Controls Resistance to a Naturally Occurring Rhabdovirus in Drosophila. *PLoS Pathog* **12**, e1005387 (2016).
145. Markow, T. A. & O'Grady, P. M. Collecting Drosophila in the wild. in *Drosophila* 145–153 (Elsevier, 2006). doi:10.1016/B978-012473052-6/50004-4.
146. Carson, H. L., Ashburner, M. & Thompson, J. N. Methods of Collecting Drosophila. in *The Genetics and Biology of Drosophila* vol. 3d 1–28 (Academic Press, London, 1983).
147. Markow, T. A. & O'Grady, P. M. Chapter 4 - Collecting Drosophila in the wild. in *Drosophila* (eds. Markow, T. A. & O'Grady, P. M.) 145–153 (Academic Press, San Diego, 2006). doi:https://doi.org/10.1016/B978-012473052-6/50004-4.

CHAPTER 2: SEQUENCING RNA FROM OLD, DRIED SPECIMENS REVEALS PAST VIROMES AND PROPERTIES OF LONG-SURVIVING RNA[1]

2.1 Summary

Recovery of virus sequences from old samples provides an opportunity to study virus evolution and reconstruct historic virus-host interactions. Studies of old virus sequences have mainly relied on DNA or RNA from fixed or frozen samples. The millions of specimens in museums represent a large underutilized source of historical virus sequences, but it is not clear how well RNA survives in old samples. We experimentally assessed the stability of RNA in insects stored dry at room temperature over 72 weeks. Although RNA molecules grew fragmented, RNA yields remained surprisingly constant. RT-qPCR of host and virus RNA showed minimal differences between dried and frozen specimens. To assess RNA survival in much older samples we acquired *Drosophila* specimens from North American entomological collections. We recovered sequences from known and novel viruses including coding complete virus genomes from a 125-year-old fly. We found that the virome of *D. melanogaster* has changed little over the past century. Galbut virus, the most prevalent virus infection in contemporary *D. melanogaster*, was also the most common in historic samples. Finally, we investigated the genomic and physical features of surviving RNA. RNA that survived was fragmented, chemically damaged, and protected by RNA-RNA and RNA-protein interactions. We ensured that our results are based on sample-derived sequences. This

[1] This work has been submitted for publication: Keene-Snickers, A.H. and Stenglein, M.D. (2024) Sequencing RNA from old, dried specimens reveals past viromes and properties of long-surviving RNA. Submitted to J. Virol. doi: https://doi.org/10.1101/2024.10.03.616531.

confirms that RNA - especially certain types - survives in biological specimens over extended periods in the absence of fixation or freezing and highlights the utility of dried specimens to provide a clearer understanding of virus evolution.

2.2 Importance

RNA from old specimens has been instrumental to study the origin and evolution of viruses. Most of these studies rely on rare fixed or frozen samples. Using insect museum specimens, we found that RNA can persist for 100+ years in the absence of fixation. We first assessed RNA stability over time using dried or frozen *Drosophila melanogaster*. We reliably detected host and virus RNA over 72-weeks. Next, we sequenced *Drosophila* museum specimens stored for 9-125 years. We found that the virome of *D. melanogaster* has remained relatively stable over the last 100-years and novel virus sequences were also described. Finally, we looked at the features of surviving host and viral RNA. We found that double-stranded RNA or RNA-protein complexes remained the most stable. This work expands current RNA sample resources to the millions of archived specimens stored in museums which can be used in future virus evolution studies.

2.3 Introduction

The study of ancient DNA has been a useful way to better understand the evolutionary history of cellular organisms and their pathogens[1–4]. RNA viruses cause many important diseases, but the study of ancient or even old RNA has mainly relied on relatively rare fixed or frozen specimens[5,6]. Nevertheless, such specimens have been useful for reconstructing the early history of pandemic pathogens[7,8]. Other studies have taken advantage of endogenized RNA virus sequences, which capture the sequences

of viruses that existed long ago [9–11]. The understanding of virus evolution would benefit from the recovery of additional old RNA virus sequences. But how well RNA survives in the absence of freezing or fixation is unclear.

There are reasons to believe that RNA may not survive well over extended periods. For one thing, RNA is short-lived during normal cellular function. The half-lives of messenger and ribosomal RNAs are measured in minutes or hours in a living cell[12]. RNA molecules are shorter than genomic DNA and more prone to hydrolysis at physiological conditions[13]. Marketing of RNA-stabilizing reagents further reinforces the concept that most RNA is unstable. Studies of historic RNA may therefore be uncommon simply because researchers assume that RNA does not survive well in old specimens.

However, there are an increasing number of studies that indicate that RNA may in fact survive well over long periods[13]. Examples include recovery of a single stranded (ss) RNA virus genome from 750-year-old barley, sequencing of a double stranded (ds) RNA virus genome from 1000-year-old corn, and profiling of RNA from a 130-year-old extinct Tasmanian tiger[14–16]. These studies indicate that RNA can survive for long periods and hint at the potential broad utility of old specimens to study RNA viruses from hundreds or thousands of years ago.

There are nearly 300 million archived arthropod specimens in North American museums alone[17]. Dried beetles, butterflies, ants and fruit flies have yielded useful DNA but to our knowledge RNA has not been sampled from such specimens[18–24]. Amongst the diverse organisms contained in entomological collections are flies in the *Drosophilidae* family[24]*. Drosophila melanogaster* has been instrumental to our

understanding of genetics, evolution, and disease, and the virome of contemporary *D. melanogaster* has been described in detail[25–29].

In this study we used two approaches to investigate RNA survival. First, we experimentally dried and froze flies and mosquitoes and measured RNA yield, length, and detectability over 72 and 52 weeks. Next, we obtained *Drosophila* specimens from museum collections. Using next generation sequencing we characterized surviving viral and host RNA. We recovered viral genome sequences from over 100 years ago and identified properties of RNA associated with survival. We followed strict lab protocols to minimize contamination by contemporary nucleic acid and present many lines of evidence to support the idea that our conclusions are based on RNA from the actual samples. Our findings challenge the assumption that RNA does not survive well and confirm the usefulness of dried biological specimens as a source of RNA to study past viral infection.

2.4 Results

2.4.1 Viral and host RNA persist in dried biological specimens

Entomological collections contain millions of dried specimens, but it's not clear to what extent RNA survives in dried insects, or more generally in dried biological specimens. To assess RNA survival in such samples, we recovered RNA from flies and mosquitoes that had been pinned and stored at room temperature or that had been frozen. We used adult *D. melanogaster* and *Aedes aegypti* from outbred colonies maintained in our insectary[30,31]. Both populations contained individuals persistently infected by one or more viruses. We sampled *D. melanogaster* over 72 weeks and *Ae.*

aegypti over 52 weeks. We quantified the concentration and size distribution of recovered RNA and used RT-qPCR to quantify levels of specific host and viral RNAs.

RNA concentrations recovered from individual flies were variable but we were able to recover RNA from all dried samples (**Fig. 1A**). We used a multiple linear regression model (MLR) to assess the relationship between concentration, time and storage condition. There was a small but statistically significant decrease in RNA concentration over time ($F = 4.9$, $p = 0.03$) but surprisingly RNA yields from dried and frozen samples did not significantly differ ($F = 0.84$, $p = 0.36$). We assessed the size distribution of recovered RNA and found that RNA grew increasingly fragmented in both dried and frozen flies (**Fig. 1B, Supp. Fig. 1**). In a multiple linear regression model with the same predictors as above, there was a statistically significant decrease in RNA length over time ($F = 19.8$, $p = 4.3 \times 10^{-5}$) but no significant difference in length between dried and frozen samples ($F = 0.4$, $p = 0.53$).

Host messenger RNA (mRNA) remained reliably detectable by RT-qPCR in both dried and frozen samples. In *D. melanogaster*, we detected ribosomal protein L32 (RpL32) mRNA in 72 of 72 frozen flies and 68 of 72 dried flies (**Fig. 1C**). When we used shorter-range primers that amplified a 56 bp region of RpL32 instead of the 110 bp region amplified by our standard primers, 71 of 72 dried flies were positive (**Fig. 1C**).

Although there was no statistically significant difference in RNA concentration and length between dried and frozen flies, there was variability in detected levels of specific host and viral RNA targets (**Fig. 1D; Supp. Fig. 2**). Levels of host and viral RNA measured by RT-qPCR decreased as storage time increased. Levels of RpL32 mRNA were higher in frozen flies than they were in the dried flies (**Fig. 1D**).

Each qPCR run included one positive control: RNA from known infected flies, and three negative controls: an extraction blank (a no sample extraction), a no-sample RT control, and a no template qPCR control. The use of negative controls corresponding to each step of the extraction and detection process would have allowed us to determine when cross-contamination had been introduced if it had been. All controls behaved as expected, confirming the absence of cross-contamination during extraction and RT-qPCR (**Supp. Fig. 3A**).

The FoCo-17 *D. melanogaster* population we sampled contained individual flies variably infected by galbut virus, La Jolla virus, Nora virus, and Thika virus[32,33]. The prevalence of each virus varied in this outbred population as did viral RNA levels in infected individuals. As with RpL32 mRNA, average levels of detected viral RNAs decreased over time in both dried and frozen flies (**Fig. 1D; Supp. Fig. 2A**). Levels of galbut virus dropped by 69x in frozen samples and 181x in dried samples between 0 and 72 weeks. After 24 weeks, there was no significant difference between galbut virus RNA levels in dried and frozen flies. The overall pattern was similar for the three other viruses (**Fig. 1D; Supp Fig. 2**). Shorter range primers detected galbut virus RNA in fewer qPCR cycles, unsurprisingly confirming that short PCR products work better for fragmented RNA (**Fig. 1E**).

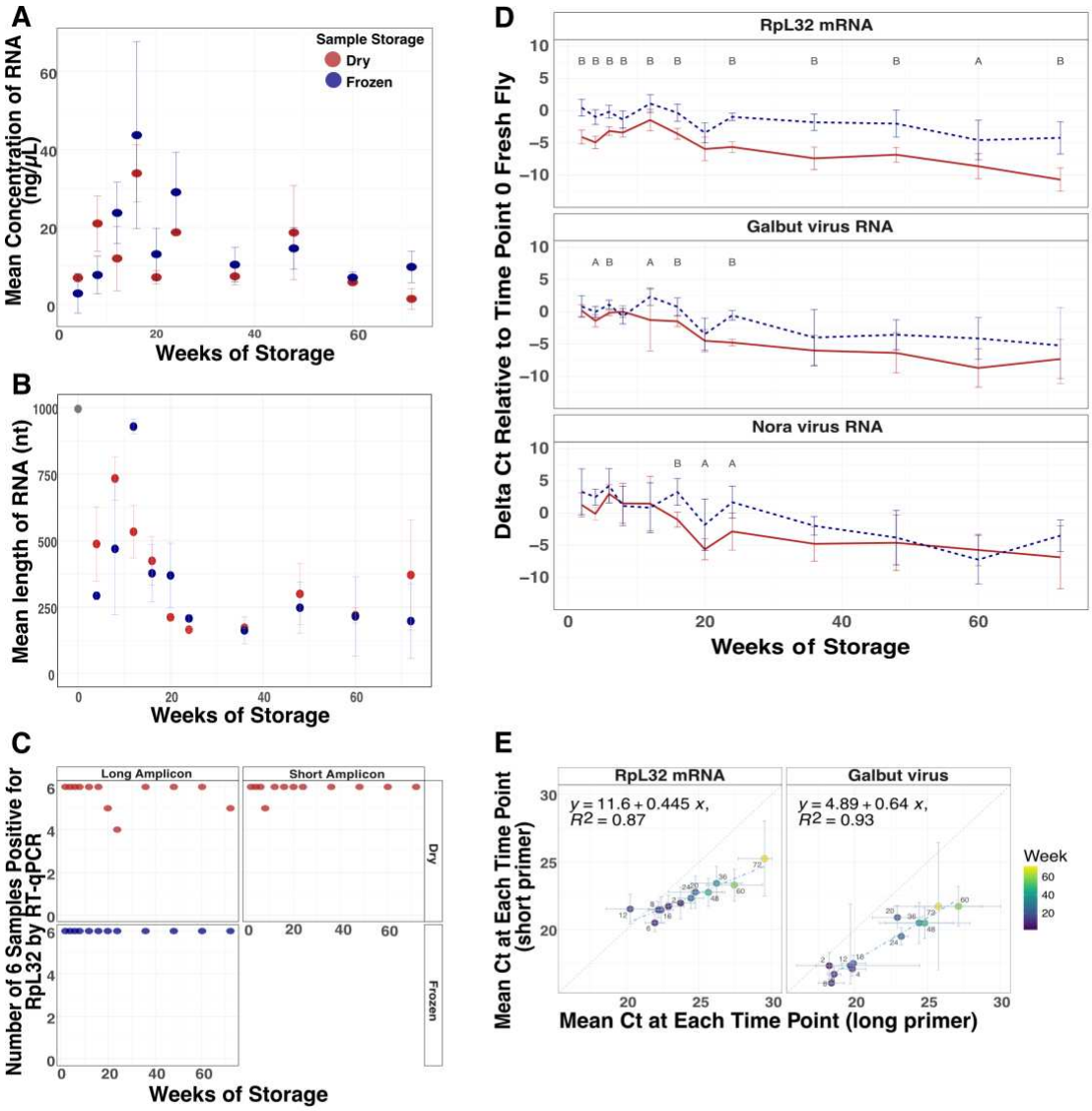


Figure 1: RNA survived in dried flies over 72 weeks. (A) Average RNA concentration and (B) length of RNA from dried and frozen flies over 72 weeks. N = 3 female *D. melanogaster* from each storage condition at each timepoint. (C) Number of samples positive of 3 male and 3 female flies at each timepoint for RpL32 mRNA via RT-qPCR using primers that target long (110 nt) and short (56 nt, dried only) regions of RpL32 mRNA. (D) Average difference of RpL32 mRNA, galbut virus and Nora virus RNA levels relative to fresh *D. melanogaster* collected at the initial timepoint. A two-tailed t-test was used to determine the statistical significance between dried and frozen means at every time point (A < 0.05, B < 0.001). (E) Comparison of mean Ct value at each time point of dried *D. melanogaster* using primers that target a long (x-axis) or

short (y-axis) region of RpL32 mRNA and galbut virus. Error bars represent mean plus and minus the standard deviation. N = 3 males and 3 females at each time point.

RNA also survived well in dried mosquitoes (**Supp. Fig 4**). An MLR was used to assess whether time and/or storage condition impacted mosquito RNA concentration. Over time ($F = 2.09$, $p = 0.15$) and between dried and frozen mosquitoes ($F = 2.37$, $p = 0.13$) there was no statistically significant difference in mosquito RNA concentrations (**Supp. Fig. 4A**). A caveat is that we neglected to save fresh mosquitoes at time 0 as we did with flies, so RNA levels were normalized to levels in the 4-week frozen samples, which had likely already experienced degradation. There was no statistically significant difference in RNA length over time after 4 weeks ($F = 0.75$, $P = 0.39$) but there was by storage condition ($F = 79.3$, $p = 2.5x10^{-13}$; **Supp. Fig. 4B**).

We used RT-qPCR to assess levels and detectability of specific host and viral RNAs in stored mosquitoes. We detected actin mRNA in all 78 dried and all 78 frozen mosquitoes (**Supp. Fig. 4C**). Actin mRNA levels remained consistent over time with no difference between 4 weeks and 52 weeks in the frozen samples and a 6x decrease in the dried samples. This mosquito population harbored verdadero virus and Guadeloupe mosquito virus[30,34]. Viral RNAs decreased over time in the dried and frozen specimens. In the dried mosquitoes, virus levels dropped between 13x for verdadero virus and 18x for Guadeloupe mosquito virus between 4 and 52 weeks. In frozen mosquitoes there was no difference in verdadero virus levels between 4 weeks and 52 weeks but there was a 5x decrease for Guadeloupe mosquito virus (**Supp. Fig. 4D-E**).

The mosquito positive and negative controls behaved as expected except for the cDNA positive control sample at week 28, which was negative for Guadeloupe mosquito

virus. This may reflect a true negative infection status, given that our cDNA positive control samples were individual mosquitoes from a variably infected outbred population confirmed to have a verdadero virus infection (**Supp. Fig. 3B**).

2.4.2 *Drosophila* obtained for RNA metagenomic sequencing were 10-125 years old and from entomological collections across the United States and Canada

Having determined that RNA survives well in experimentally dried insects, we turned our attention to much older samples. We obtained 46 specimens from entomological collections across the United States and Canada (**Table 1**). Thirty-two had been assigned as *D. melanogaster* based on morphology, nine as *D. simulans*, and five as unknown *drosophilids.* Collection dates ranged from 1896 to 2011 (**Fig. 2A; Table 1; Supp. Table 1**). We extracted RNA from 29 specimens following a standard destructive protocol, which yielded RNA concentrations from 0 to 152 ng/μL (**Fig. 2B; Table 1; Supp. Table 1**). Some samples were loaned on the condition that they not be destroyed, so we followed a non-destructive RNA isolation protocol for the remaining samples[35]. This used an overnight incubation with proteinase to liberate internal contents of the specimen while leaving morphology intact. The non-destructive extractions yielded more RNA than the standard protocol ($p = 0.03$), with concentrations ranging from 4 to120 ng/μL (**Fig. 2B; Table 1; Supp. Table 1**). However, when samples from Hawai'i, which had unusually low yields, were removed there was no statistically significant difference between RNA concentration obtained from the two methods ($p = 0.59$; **Supp. Fig. 5**). Four samples that had been stored in 70% ethanol required rehydration prior to extraction (**Supp. Fig 5**).

Table 1: Museum specimen metadata.

Sequencing Based Species ID	Date Collected	Location (1)	Donating Institution	BioSample (2)	Sample ID (3)	Museum Accession	Storage Type	Extraction Method	Concentration (ng/µl) (4)
D. melanogaster	1896	Urbana, IL	Illinois Natural History Survey	SAMN38071661	Illinois_1896	1004284	Dried	Non-Destructive	4
D. melanogaster	1908	Albany, NY	New York State Museum	SAMN38071662	NewYork_1908	-	Dried	Destructive	152
D. melanogaster	1908	Urbana, IL	Illinois Natural History Survey	SAMN38071654	Illinois_1908	1004277	Dried	Non-Destructive	54.6
D. melanogaster	1915	Urbana, IL	Illinois Natural History Survey	SAMN38071656	Illinois_1915_1	1004279	Dried	Non-Destructive	84
D. putrida	1915	Urbana, IL	Illinois Natural History Survey	SAMN38071657	Illinois_1915_2	1004280	Dried	Non-Destructive	6.8
D. melanogaster	1915	Urbana, IL	Illinois Natural History Survey	SAMN38071658	Illinois_1915_3	1004281	Dried	Non-Destructive	12.7
D. melanogaster	1915	Urbana, IL	Illinois Natural History Survey	SAMN38071659	Illinois_1915_4	1004282	Dried	Non-Destructive	16.9
D. melanogaster	1919	St. Paul, MN	University of Minnesota, Insect Collection	SAMN38090828	Minnesota_1919_1	-	Dried	Destructive	0
D. melanogaster	1919	St. Paul, MN	University of Minnesota, Insect Collection	SAMN38071687	Minnesota_1919_2	-	Dried	Destructive	21.6
D. melanogaster	1919	St. Paul, MN	University of Minnesota, Insect Collection	SAMN38071688	Minnesota_1919_3	-	Dried	Destructive	18.9
D. melanogaster	1927	Albany, NY	New York State Museum	SAMN38071663	NewYork_1927	-	Dried	Destructive	21
D. melanogaster	1930	Urbana, IL	Illinois Natural History Survey	SAMN38071660	Illinois_1930	1004283	Dried	Non-Destructive	13.6
D. melanogaster	1942	Glassboro, NJ	Illinois Natural History Survey	SAMN38071655	NewJersey_1942	1004278	Dried	Non-Destructive	10.2
D. melanogaster	1950	Oahu, HI	University of Hawaii, Insect Museum	-	Hawaii_1950_1	-	Dried	Destructive	0
D. melanogaster	1950	Oahu, HI	University of Hawaii, Insect Museum	-	Hawaii_1950_2	-	Dried	Destructive	0
D. melanogaster	1951	Honolulu, HI	University of Hawaii, Insect Museum	-	Hawaii_1951_1	-	Dried	Destructive	0
D. melanogaster	1951	Oahu, HI	University of Hawaii, Insect Museum	-	Hawaii_1951_2	-	Dried	Destructive	0
D. melanogaster	1952	Kauai, HI	University of Hawaii, Insect Museum	-	Hawaii_1952_1	-	Dried	Destructive	0
D. melanogaster	1953	Kauai, HI	University of Hawaii, Insect Museum	SAMN38071677	Hawaii_1953_1	-	Dried	Destructive	0
D. melanogaster	1963	Lebanon Co, PA	Frost Entomological Museum	SAMN38071668	Pennsylvania_1963_1	-	Dried	Non-Destructive	66.4
D. melanogaster	1963	Lebanon Co, PA	Frost Entomological Museum	SAMN38071669	Pennsylvania_1963_2	-	Dried	Non-Destructive	24.4
D. melanogaster	1968	St. Cloud, MN	University of Minnesota, Insect Collection	SAMN38071684	Minnesota_1968	-	Dried	Destructive	13
D. melanogaster	1969	St. Cloud, MN	University of Minnesota, Insect Collection	SAMN38071685	Minnesota_1969	-	Dried	Destructive	38.6
D. melanogaster	1991	St. Cloud, MN	University of Minnesota, Insect Collection	SAMN38071686	Minnesota_1991	-	Dried	Destructive	24.4
D. putrida	2003	Lebanon Co, PA	Frost Entomological Museum	SAMN38071673	Pennsylvania_2003_1	4908	Dried	Non-Destructive	71.2
D. putrida	2003	Lebanon Co, PA	Frost Entomological Museum	SAMN38071672	Pennsylvania_2003_2	4887	Dried	Non-Destructive	13.9
S. pallida	2003	Lebanon Co, PA	Frost Entomological Museum	SAMN38071675	Pennsylvania_2003_3	5867	Dried	Non-Destructive	42
S. pallida	2003	Lebanon Co, PA	Frost Entomological Museum	SAMN38071676	Pennsylvania_2003_4	5873	Dried	Non-Destructive	120
S. pallida	2003	Lebanon Co, PA	Frost Entomological Museum	SAMN38071674	Pennsylvania_2003_5	5863	Dried	Non-Destructive	25.6
D. melanogaster	2006	Davidson, NC	The Davidson College Entomology Collection	SAMN38071666	NorthCarolina_2006_1	-	Dried	Destructive	19.6
D. melanogaster	2006	Davidson, NC	The Davidson College Entomology Collection	SAMN38071667	NorthCarolina_2006_2	-	Dried	Destructive	56.8
D. melanogaster	2006	Davidson, NC	The Davidson College Entomology Collection	-	NorthCarolina_2006_3	-	Dried	Destructive	0
D. melanogaster	2010	Puslinch Township, Ontario, Canada	University of Guelph, Centre for Biodiversity Genomics	SAMN38071680	Canada_2010_1	PHDIP1072-11	Ethanol	Destructive	0
D. simulans	1952	Kauai, HI	University of Hawaii, Insect Museum	-	Hawaii_1952_2	-	Dried	Destructive	0
D. simulans	1952	Kauai, HI	University of Hawaii, Insect Museum	-	Hawaii_1952_3	-	Dried	Destructive	0
D. simulans	1953	Kilauea, HI	University of Hawaii, Insect Museum	SAMN38071678	Hawaii_1953_2	-	Dried	Destructive	0
D. simulans	1958	Maui, HI	University of Hawaii, Insect Museum	-	Hawaii_1958	-	Dried	Destructive	0
D. simulans	1960	Oahu, HI	University of Hawaii, Insect Museum	-	Hawaii_1960	-	Dried	Destructive	0
D. simulans	1963	Lebanon Co, PA	Frost Entomological Museum	SAMN38071670	Pennsylvania_1963_3	-	Dried	Non-Destructive	84.6
D. simulans	1963	Lebanon Co, PA	Frost Entomological Museum	SAMN38071671	Pennsylvania_1963_4	-	Dried	Non-Destructive	48
D. simulans	1965	Oahu, HI	University of Hawaii, Insect Museum	SAMN38071679	Hawaii_1965	-	Dried	Destructive	0
D. simulans	2000	Santa Barbara, CA	Santa Barbara Museum of Natural History	SAMN38071683	California_2000_1	-	Dried	Destructive	17.8
D. simulans	2000	Santa Barbara, CA	Santa Barbara Museum of Natural History	SAMN38071682	California_2000_2	-	Dried	Destructive	55.4
D. simulans	2010	Puslinch Township, Ontario, Canada	University of Guelph, Centre for Biodiversity Genomics	SAMN38071681	Canada_2010_2	PHDIP946-11	Ethanol	Destructive	17.8
D. simulans	2011	Ventura Co, CA	University of Guelph, Centre for Biodiversity Genomics	SAMN38071664	California_2011_1	BBDIV1549-12	Ethanol	Destructive	74.2
D. simulans	2011	Ventura Co, CA	University of Guelph, Centre for Biodiversity Genomics	SAMN38071665	California_2011_2	BBDIV1550-12	Ethanol	Destructive	15.9

(1) All specimens from USA unless otherwise noted
(2) Specimens that did not yield sufficient sequencing library were not sequenced and thus have no BioSample ID
(3) Our study sample ID
(4) RNA concentrations below the limit of detection on the Qubit were assigned a concentration of 0

RNA from samples collected after or before 1960 had average lengths of 159 and 60 nt, respectively (**Fig. 2C**). Although there was variability in RNA quantity and quality, the RNA obtained from these specimens was of suitable length and concentration for downstream analyses.

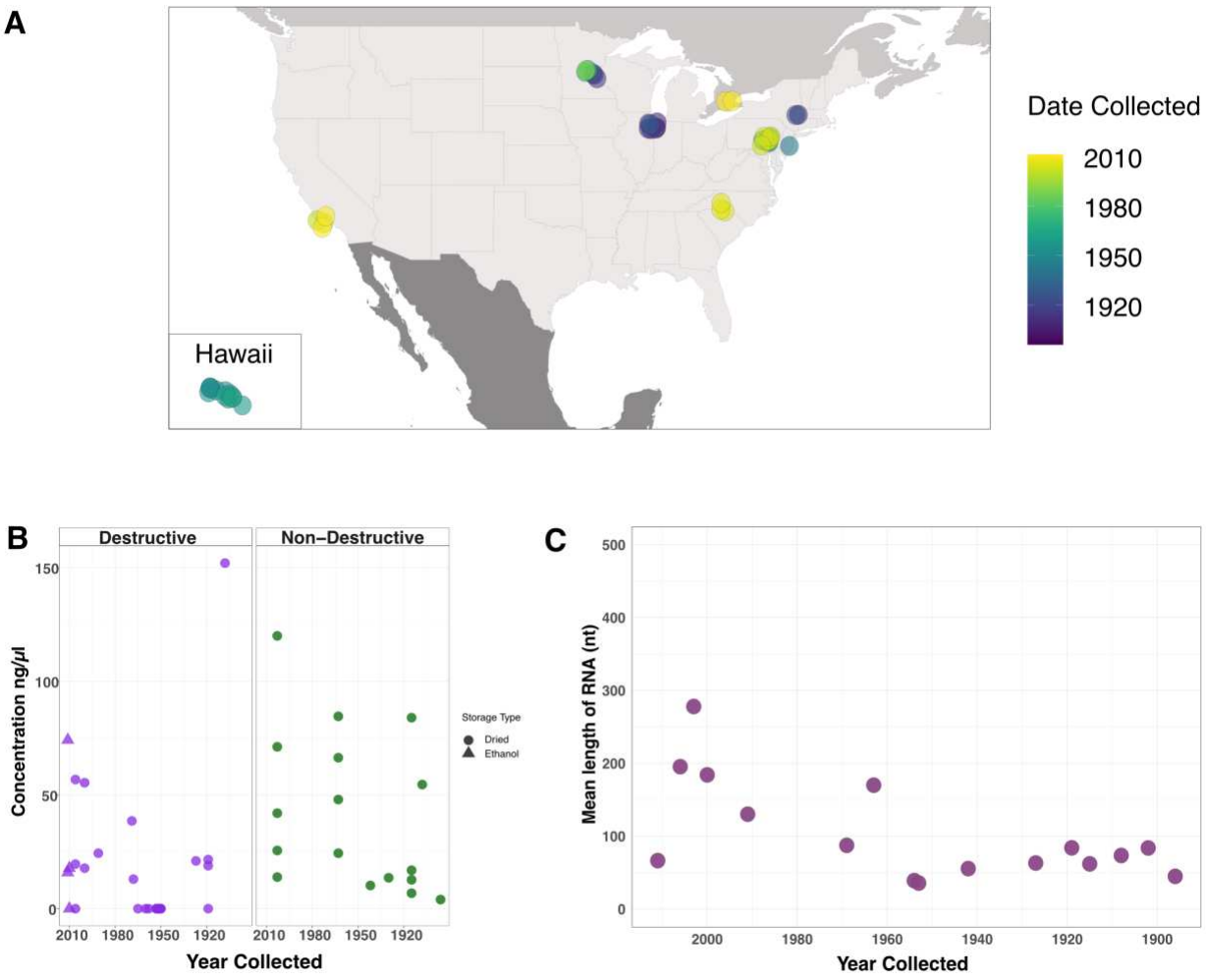


Figure 2: Museum sample collection locations and characteristics of extracted RNA. (A) Map of the United States and Canada indicating sample collection locations. (B) RNA concentration from individual museum flies using either the destructive (purple) or non-destructive (green) extraction method. (C) Mean length of RNA molecules from a subset of museum specimens encompassing the range of years specimens obtained for this study had been collected.

We screened museum samples for galbut virus RNA and RpL32 mRNA using our short-range primers. Sixteen of 46 samples (35%) were positive for galbut virus and 21 (46%) were positive for RpL32 mRNA (**Supp. Fig. 6A, Supp. Table 1**). Positive status was confirmed using melting temperature and agarose gel electrophoresis. The

small size of PCR products made them difficult to distinguish from primer dimers on agarose gels, so it is possible that there were false negative calls. All negative control samples were negative as expected (**Supp. Fig. 6B**).

2.4.3 Metagenomic sequencing of museum samples and molecular species assignment

We prepared shotgun sequencing libraries from museum sample RNA. Thirty-six samples yielded libraries suitable for sequencing. Sequencing yielded an average of 48 million reads per sample (range: 0.4-238 million). Several samples were sequenced multiple times to increase recovery of complete virus sequences.

The records for some museum samples did not include morphological taxonomic assignments below the family level, so we used competitive mapping to generate molecular taxonomic assignments. We mapped quality and adapter trimmed reads to a set of 545 representative cytochrome oxidase subunit-1 (CO1) sequences from across the Drosophilidae family to generate a molecular species identification for each sample. This molecular species assignment corroborated the museum morphology-based assignment for most samples (**Supp. Table 2)**. Four samples labeled as *D. melanogaster* in museum records were revealed to be *D. simulans* by CO1-mapping. *D. melanogaster* and *D. simulans* are common, sympatric, and difficult to distinguish morphologically, so such misassignment is expected[36]. Five samples did not have species-level taxonomic assignments from the museums. Of these, CO1-mapping indicated that 2 were *Drosophila putrida* and the remaining three were *Scaptomyza pallida*. One fly collected in Illinois in 1915 was also assigned as *D. putrida* (**Supp. Table 1-2**). These assignments are consistent with the known range of *D. putrida* and

S. pallida[37]. All fresh frozen and experimental dried samples were assigned as *D. melanogaster* as expected and there were no CO1-mapping reads in the three water negative control datasets.

2.4.4 The overall taxonomic composition of old and new samples is similar

To broadly assess the source of sequences in our datasets, we taxonomically binned sequences using or lab's metagenomic classification pipeline [38]. This first identified host-derived reads by mapping to an index composed of the *D. melanogaster*, *D. putrida*, and *S. pallida* genomes[39,40]. The pipeline then assembled remaining non-host reads and taxonomically assigned the resulting contigs and unassembled reads using a BLASTN search of the NCBI nucleotide database. We tabulated the fraction of reads in each dataset that were assigned to various taxa or that remained unassigned (**Fig. 3**).

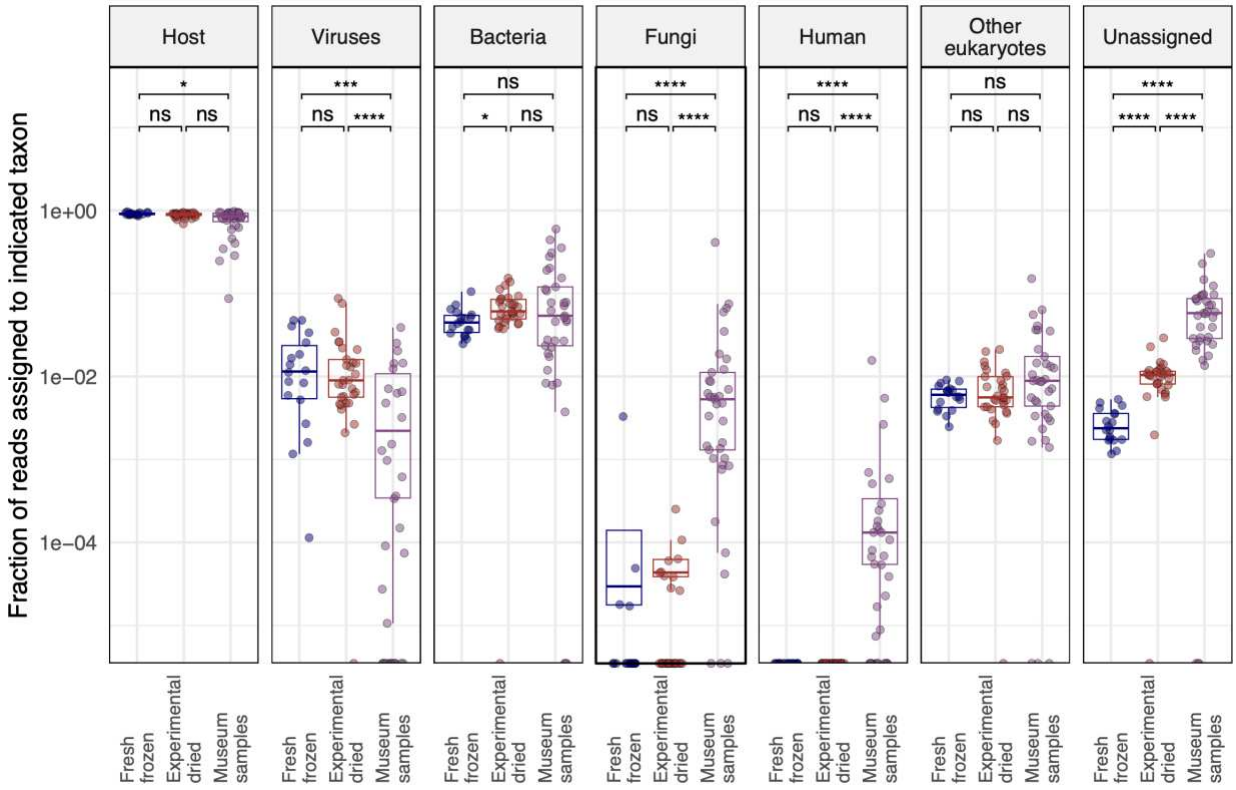


Figure 3: The overall taxonomic composition of old and new samples was similar. The fraction of reads assigned to the indicated taxa are plotted. Each point represents a dataset from an individual fly. Adjusted *p*-values significance levels from Wilcoxon test are indicated. Host-mapping reads were identified by mapping to combined *D. melanogaster*, *D. putrida*, and *S. pallida* genomes. Fractions of reads mapping to other taxa were determined by metagenomic classification of non-host-mapping reads. The "other eukaryotes" category accounts for all eukaryotic taxa apart from fungi and human. Unassigned read fractions account for reads that were not assigned by host-mapping or metagenomic classification.

Most reads mapped to the fly genomes in all datasets, although the median fraction of host-mapping reads fell from 90.5% in fresh samples to 85% in museum samples (**Fig. 3**). This decrease was accompanied by a corresponding increase in unassigned reads. Only 0.2% of reads were unassigned in fresh datasets, compared to 5.5% in museum datasets ($p=2.3x10^{-6}$). This increase in unassignable reads is likely attributed to the fact that short reads from fragmented RNA contain less information for

host mapping or taxonomic assignment. There were on average fewer virus-mapping reads in museum samples, which could reflect the fact that our FoCo-17 population was persistently infected by multiple viruses or could reflect decreased relative survival of viral RNA in old samples. There were increases in the proportion of fungi- and human-mapping reads in museum datasets. Fungal reads in old samples may have derived from saprophytic fungi. Human-mapping reads in museum datasets may have originated from handling of specimens by curators, accumulation of dust on samples, or an increased fraction of contamination-derived sequences in old datasets. The median fraction of human-mapping reads in old datasets was 5.4×10^{-5} and the maximum fraction was 1.6% of reads, in a dataset from Hawai'i. We concluded that some of the reads in our libraries, including human-assigned reads, may have derived from contamination. However, these accounted for a small fraction of datasets and the overall taxonomic composition of new and old datasets was similar (**Fig. 3**).

2.4.5 Entomological specimens harbor diverse viral sequences including previously unknown viruses

We recovered at least one virus sequence from 21 of the 36 sequenced museum samples and two or more virus sequences from 5 samples (**Fig. 4; Table 2**). All but three of the virus sequences corresponded to known *Drosophila*-infecting viruses[25,26]. Galbut virus was the most common virus, with six samples producing coding-complete genomes and six samples producing partial galbut virus genomes (**Fig. 4; Table 2**). The next most common viruses were *D. melanogaster* sigma virus and vera virus, which were detected in three samples each (**Fig. 4; Table 2**). The sample with the most viral

sequences was collected in 1908 in Illinois, USA: this fly yielded three coding-complete and one partial virus sequence.

We identified two coding-complete genomes from previously undescribed viruses. The first was a bunyavirus from a *D. melanogaster* collected in Ontario, Canada in 2010, which we named Puslinch virus. The Puslinch virus L protein shares only ~35% pairwise amino acid identity with its nearest known relatives, which were viruses in the genus Herbevirus (**Fig. 4; Fig. 5; Table 2**). The other new virus sequence was a sobemo-like virus from a *D. putrida* collected in Pennsylvania, USA in 2003 (**Fig. 4; Fig 5; Table 2**).

Additionally, we recovered a complete tobacco mosaic virus (TMV) genome with high coverage depth (850x) from the 1908 *D. melanogaster* specimen. We speculated that this sequence may have originated from contamination of the specimen, perhaps during handling by a smoker[41].

Positive control libraries, constructed from RNA from pooled FoCo-17 flies, contained the expected virus sequences (**Fig. 4**). To avoid potential read contamination via index hopping, we sequenced positive control libraries on a separate sequencing run from the museum specimens[42]. Negative control datasets were included in each run and contained fewer than 15k reads after trimming of adapter and low-quality bases. Negative control datasets contained zero reads mapping to any of the virus sequences present in any of the other datasets (**Fig. 4**)[39].

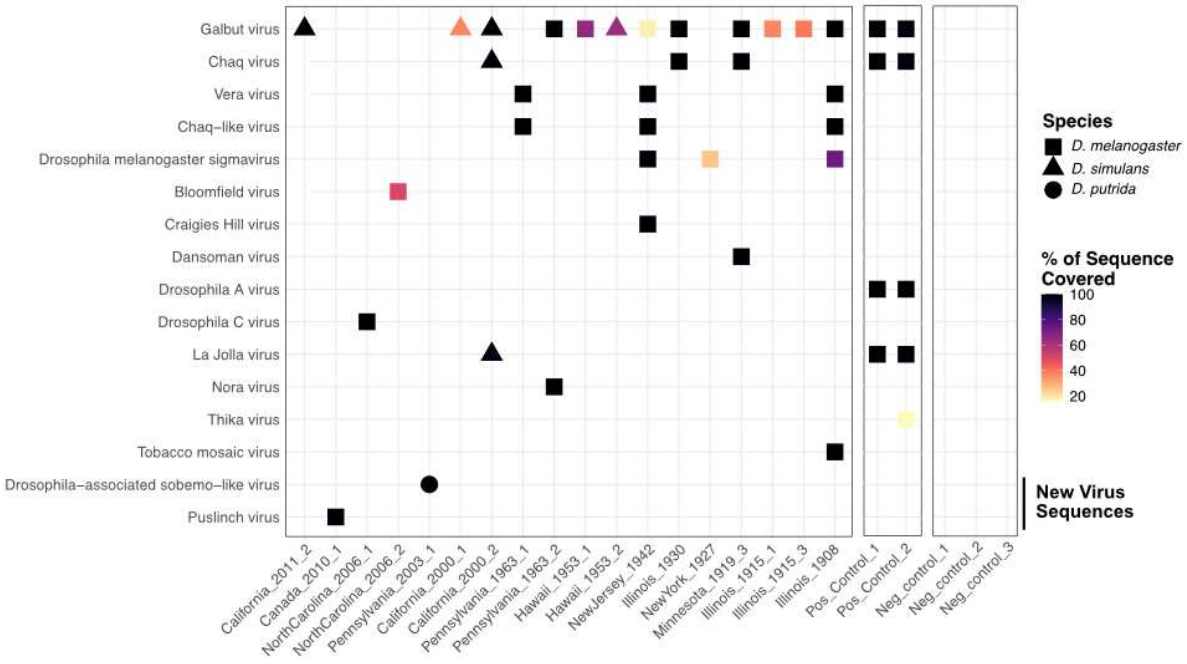


Figure 4. Entomological museum specimens harbor diverse viral sequences. Heatmap showing virus sequences recovered from individual museum samples. Shape corresponds to host species and color corresponds to the percent of reference sequence covered (**Table 2**). For segmented viruses, the average coverage across all segments is shown. Viruses detected in positive control datasets, made from pooled FoCo17 flies, are shown. No reads mapping to any of these viruses were present in any of the three negative control datasets.

Table 2: Virus sequences identified in *Drosophila* museum collection specimens.

Date Collected	Location	Sample ID	Virus	GenBank Sequence Accession	Nearest GenBank[1]	% Query Coverage[2,3]	% Identity[4,5]	Completeness	Average Coverage[6]	Estimated Evolutionary Rate[7]
Known										
1908	Illinois	Illinois_1908	Drosophila melanogaster sigmavirus	-	MH384277	73	99.4	Partial sequence	2	-
1908	Illinois	Illinois_1908	Tobacco mosaic virus	OR820564	KY810785	100	99.7	Coding complete sequence	850	7.13E-05
1908	Illinois	Illinois_1908	Vera virus	OR820566, OR820565	MT742171, MT742172	100	99.9 - 100	Coding complete sequence	9-34	1.09E-05 - 0.00E
1908	Illinois	Illinois_1908	Galbut virus	OR820562, OR820561, OR820563	OR729093, MT742165, OR729068	100	98.9 -99.3	Coding complete sequence	7-14	6.43E-05 - 1.00E-04
1908	Illinois	Illinois_1908	Chaq-like virus	OR820560	MT742173	100	99.9	Coding complete sequence	18	6.36E-06
1915	Illinois	Illinois_1915_1	Galbut virus	-	MT742164, MT742161, MT742162	37	96.1	Partial sequence	4	-
1915	Illinois	Illinois_1915_3	Galbut virus	-	MT742164, MT742161, MT742162	39	96.5	Partial sequence	1	-
1919	Minnesota	Minnesota_1919_3	Dansoman virus	OR820572, OR820573	MH384295, MH384270	100	97.4 - 98.3	Coding complete sequence	40-46	2.93E-04 - 3.93E-04
1919	Minnesota	Minnesota_1919_3	Chaq virus	OR820571	MH384367	100	96.9	Coding complete sequence	317	4.11E-04
1919	Minnesota	Minnesota_1919_3	Galbut virus	OR820575, OR820574, OR820576	MT742164, MT742165, OR729068	100	99.4 - 99.6	Coding complete sequence	247-523	4.55E-05 - 9.31E-05
1927	New York	NewYork_1927	Drosophila melanogaster sigmavirus	-	NC_038281	26	99.2	Partial sequence	2	-
1930	Illinois	Illinois_1930	Chaq virus	OR820567	MH384311	100	97.1	Coding complete sequence	32	4.88E-04
1930	Illinois	Illinois_1930	Galbut virus	OR820569, OR820568, OR820570	OR729094, MT742165, OR729070	100	89.3 - 99.7	Coding complete sequence	29-50	3.64E-05 - 1.19E-03
1942	New Jersey	NewJersey_1942	Craigies Hill virus	OR820579, OR820578	MH384377, MH384349	100	90.4 - 98.1	Coding complete sequence	22-40	2.95E-04 - 1.46E-03
1942	New Jersey	NewJersey_1942	Galbut virus	-	MT742164, MT742161, MT742162	18	97.1	Partial sequence	17	-
1942	New Jersey	NewJersey_1942	Chaq-like virus	OR820577	MT742173	100	99.9	Coding complete sequence	490	9.21E-06
1942	New Jersey	NewJersey_1942	Vera virus	OR820582, OR820581	MT742171, MT742172	100	99.9 - 100	Coding complete sequence	463-686	1.58E-05 - 0.00E
1942	New Jersey	NewJersey_1942	Drosophila melanogaster sigmavirus	OR820580	MH384306	100	99.2	Coding complete sequence	46	1.88E-04
1953	Hawai'i	Hawaii_1953_1	Galbut virus	-	MT742164, MT742161, MT742162	65	92.4	Partial sequence	3	-
1953	Hawai'i	Hawaii_1953_2	Galbut virus	-	MT742164, MT742161, MT742162	63	99.7	Partial sequence	3	-
1963	Pennsylvania	Pennsylvania_1963_1	Vera virus	OR820593, OR820592	MT742168, MT742172	100	99.6 - 100	Coding complete sequence	2512-2727	6.55E-05 - 0.00E
1963	Pennsylvania	Pennsylvania_1963_1	Chaq-like virus	OR820591	MT742173	100	99.8	Coding complete sequence	5323	3.64E-05
1963	Pennsylvania	Pennsylvania_1963_2	Galbut virus	OR820588, OR820587, OR820589	MH384303, MH384304, MH384276	100	96.4 - 99.2	Coding complete sequence	8-81	1.73E-04 - 7.98E-04
1963	Pennsylvania	Pennsylvania_1963_2	Nora virus	OR820590	JX220408	100	97.1	Coding complete sequence	16	5.98E-04
2000	California	California_2000_2	La Jolla virus	OR820598	MH384285	97	95.35	Partial sequence	11	5.81E-03
2000	California	California_2000_2	Chaq virus	OR820594	MT742163	100	82.8	Coding complete sequence	62	1.01E-02
2000	California	California_2000_2	Galbut virus	OR820596, OR820595, OR820597	MH384283, MH384336, MH384366	100	83.5 - 92.4	Coding complete sequence	150-271	6.00E-04 - 2.07E-02
2000	California	California_2000_1	Galbut virus	-	MT742164, MT742161, MT742162	37	95	Partial sequence	2	-
2006	North Carolina	NorthCarolina_2006_1	Drosophila C virus	OR820583	OK188767	100	98.3	Coding complete sequence	264	1.86E-03
2006	North Carolina	NorthCarolina_2006_2	Bloomfield virus	-	MF416371, KP714091, KP714093, KP714094	50	99.6 - 100	Partial sequence	0.4-2	-
2011	California	California_2011_2	Galbut virus	OR820600, OR820599, OR820601	MH384283, MH384336, MH384366	100	83.8 - 99.4	Coding complete sequence	16-18	2.00E-03 - 2.61E-02
Novel										
2003	Pennsylvania	Pennsylvania_2003_1	Drosophila-associated sobemo-like virus	OR820603, OR820602	UYL94340.1, QHA33877.1	100	38.1 - 70.6	Coding complete sequence	84-136	-
2010	Ontario, CAN	Canada_2010_1	Puslinch virus	OR820586, OR820585, OR820584	YP_009362026.1, YP_010840683.1, YP_009362024.1	100	32.8 - 39.1	Coding complete sequence	81-188	-

[1] Nearest GenBank sequence is provided for RdRp of Drosophila-associated sobemo-like virus and the L, M and S segment of Puslinch virus.

[2] % Query coverage from BLASTN alignment

[3] % Query coverage from BLASTN alignment for novel virus sequences was determined based on nearest GenBank sequence.

[4] nt, percentage nucleotide identity to closest GenBank sequence identified via BLASTn.

[5] For novel virus sequences, % nt identity to the reference sequence is shown.

[6] Average mapped read coverage across contigs as reported in Geneious.

[7] Calculated using the nearest GenBank sequence with date collected data.

2.4.6 Relatedness of old virus sequences to contemporary ones

Sequences from museum samples of previously described viruses ranged from 83.5% to 100% identical to existing sequences with 38% of the sequences sharing ≥ 99% pairwise identity to existing sequences (**Table 2**). Notably, all vera virus RNA 1 sequences, including one from 1908, were 100% identical to their closest available existing sequence, though the vera virus RNA 2 and chaq-like sequences from the same samples were not identical to existing sequences.

We initially considered a Bayesian approach to generate evolutionary rate estimates but found our data had insufficient temporal signal[43]. We instead generated initial estimates of evolutionary rates using differences in sequence identity and sampling times for the most closely related contemporary sequences (**Supp. Fig. 7; Table 2,**). This may underestimate rates due to saturation and might overestimate rates because the old sequences are likely not the direct ancestors of the contemporary sequences[44–46]. Sequences without a suitably similar contemporary sequence (e.g., USA 2000 and 2011 galbut virus *D. simulans* RNA 2 and 3) were not analyzed in this way. Estimated rates were generally higher in sequences from more recent samples (2000-2011) and lower in sequences from older samples, with sequences from 1908 showing the lowest rates (**Supp. Fig. 7**). This observation is consistent with the time-dependent rate phenomenon, in which evolutionary rate estimates decrease as the timescale of measurement increases[44,47].

The most common contemporary viral infection of *D. melanogaster* is galbut virus, the success of which is likely attributable to efficient biparental vertical transmission and minimal apparent fitness costs[26,30,48]. Detection of galbut virus in

museum samples by sequencing matched detection by RT-qPCR (**Table 1**). In maximum likelihood trees, three clades of galbut virus were evident: two clades (A and B) consisted of sequences from *D. melanogaster* while the third clade consisted of sequences from *D. simulans* (**Fig. 5A-C**). An endogenized galbut virus RNA 1 sequence was on its own branch most closely related to clade A (**Fig. 5A**)[49]. All the historic galbut virus sequences fell within existing galbut virus diversity and clustered with other sequences from the same host i.e., the two *D. simulans* sequences clustered with previously described *D. simulans* sequences and the *D. melanogaster* sequences with known *D. melanogaster* RNA1 sequences (**Fig. 5A**). Interestingly, the RNA1 and RNA2 sequence from Illinois 1930 clustered with clade A but this virus's RNA 3 was situated on a separate branch. The 1930 Illinois RNAs 1 and 2 were 99% identical to contemporary sequences in GenBank while RNA 3 was only 89% identical. This phylogenetic discordance is consistent with galbut virus reassortment[50].

The sigma virus sequence from 1942 clustered within the previously described diversity of *D. melanogaster* sigma viruses (**Fig. 5D**). Puslinch virus and *Drosophila*-associated sobemo-like virus sequences were situated on their own branches on trees of related sequences (**Fig. 5E-F**). These new virus sequences are highly divergent and may establish new genera or higher order taxa.

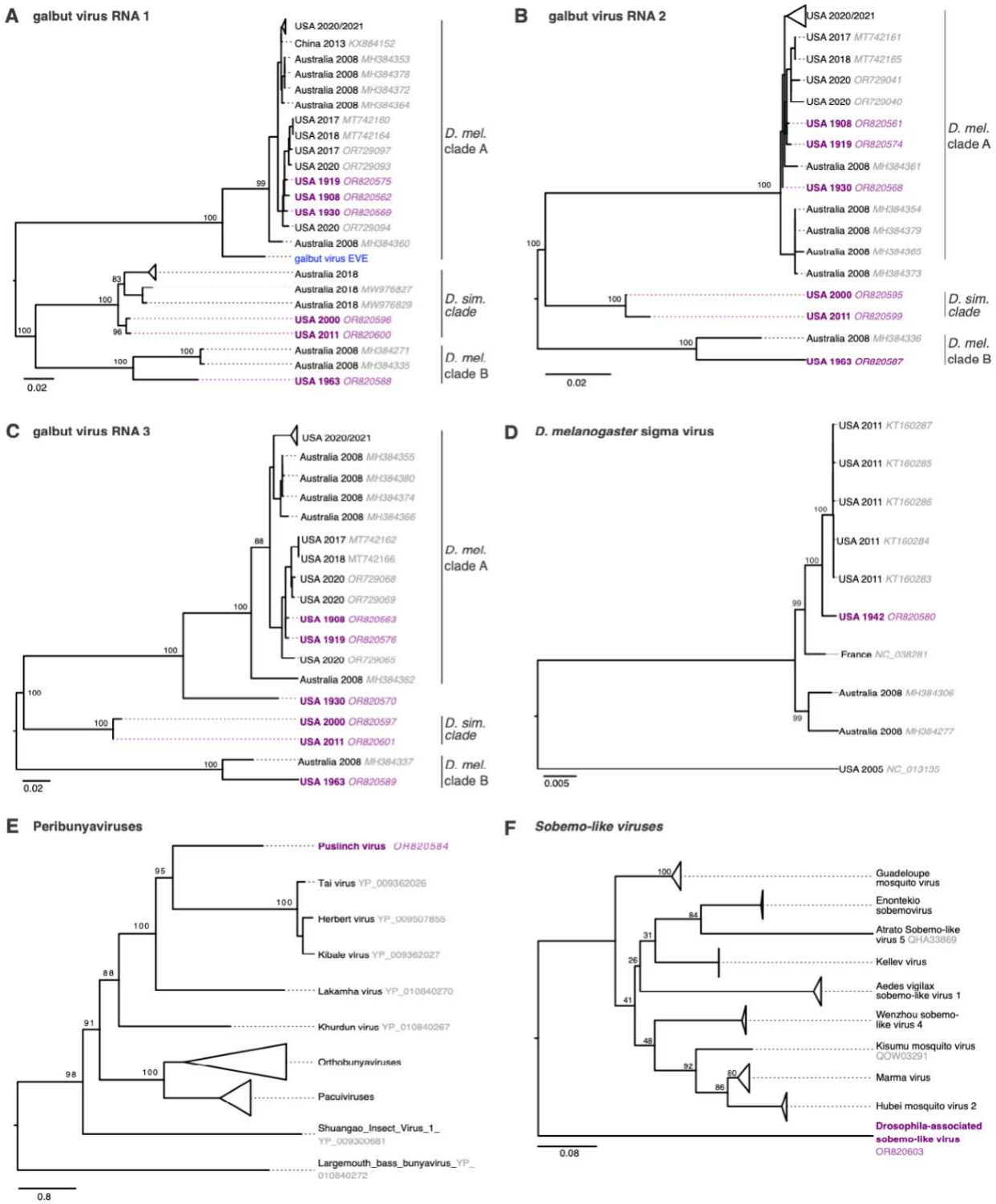


Figure 5: Trees showing museum sample virus sequences in the context of related contemporary sequences. (A) Maximum likelihood trees for galbut virus RNA 1 (RdRp) sequences, (B) galbut virus RNA 2 sequences, (C) galbut virus RNA3 sequences, (D) *D. melanogaster* sigma virus sequences, (E) L protein sequences for

viruses in the family *Peribunyaviridae*, and (F) putative RNA-dependent RNA polymerase protein sequences related to the new *Drosophila*-associated sobemo-like virus. Trees in (A)-(D) are based on nucleotide alignments and trees in (E) and (F) are based on protein sequence alignments. Purple color indicates museum sample-derived sequence generated in this study. Accession numbers are noted except when groups of closely related sequences are collapsed. All trees are midpoint rooted. Blue tip indicates the sequence of an endogenized galbut virus RNA 1 sequence.

2.4.7 Certain types of RNA survived better in old samples

We performed a variety of analyses to characterize the molecular and genomic properties of surviving RNA. An initial indication that not all types of RNA survived equally came from the strandedness of virus-mapping reads. We constructed libraries using a protocol that preserved information about the RNA strand from which reads derived[51]. Galbut virus is a partitivirus, which have dsRNA genomes[52]. But in infected flies, nearly all galbut virus RNA was positive (+) sense (98.2% in fresh frozen flies; **Fig. 6; Supp. Fig. 8**). This is presumably because most galbut virus RNA in infected cells is messenger RNA (mRNA). The fraction of +strand galbut virus RNA decreased in older samples. In experimentally dried samples an average of 89.1% of reads were from +strand RNA, and this value fell to 74.1% in museum specimens ($p = 1.4 \times 10^{-4}$; **Fig. 6**). The decreasing fraction of +strand galbut virus RNA could be explained by the preferential survival of double-stranded galbut virus RNA in older samples.

Coverage of galbut virus in museum datasets derived from all segments and was not concentrated in regions of RNA 1 corresponding to the PCR primers we use to detect galbut virus in our lab (**Supp. Fig. 8**). The lack of higher coverage in PCR product regions supports the conclusion that galbut virus sequences did not result from contamination by contemporary PCR products.

Other viruses had different fractions of +strand RNA that were consistent with the genome type of each virus (**Fig. 6**). For instance, a majority of reads from *D. melanogaster* sigma virus, a negative (-) strand RNA virus, were from -strand RNA[29]. RNA from Nora virus (a +strand virus) remained almost totally +strand in all samples[33]. The fraction of +strand vera virus RNA, another partitivirus, remained high in old samples (**Fig. 6**). This indicated that it was not simply dsRNA that survived in older samples.

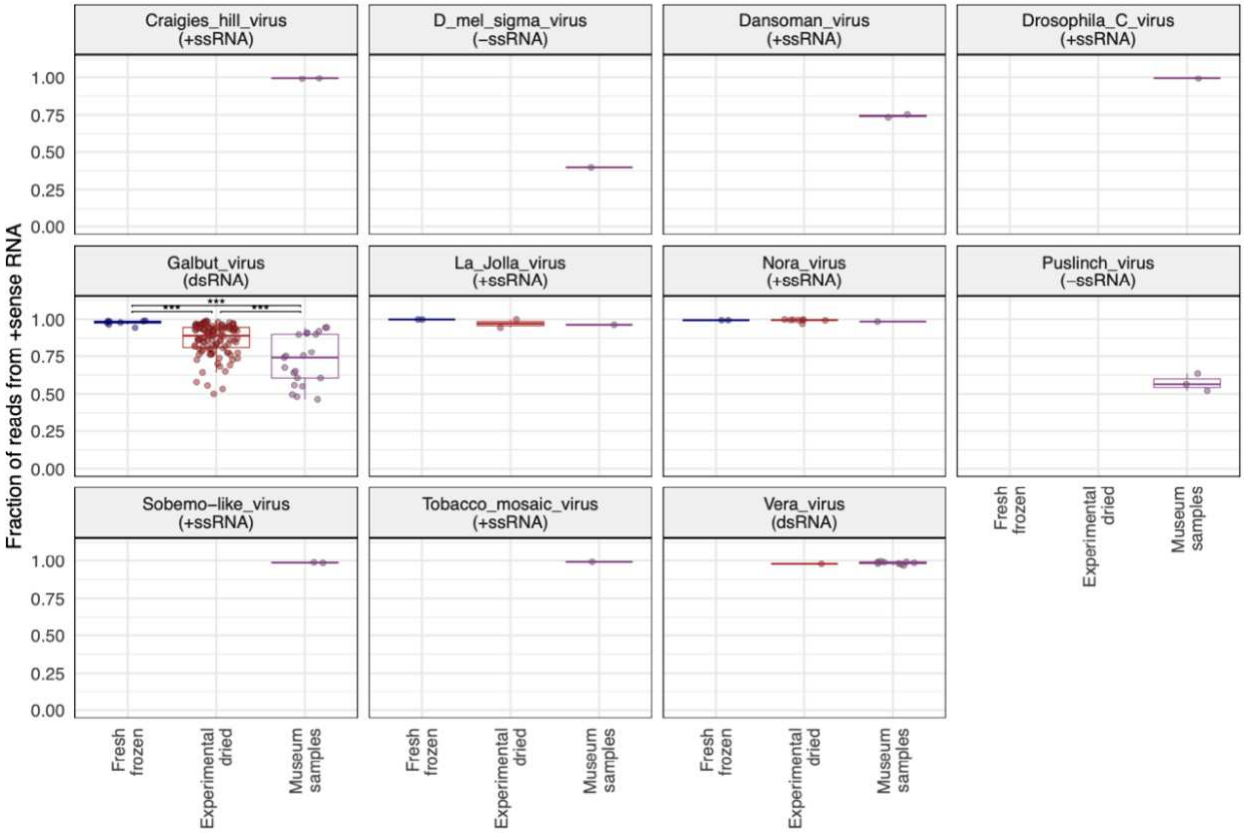


Figure 6: The strandedness of galbut virus RNA in old samples is consistent with preferential survival of dsRNA. The fraction of virus-mapping reads that originated from +strand RNA is plotted. Each point represents a dataset from an individual *D. melanogaster* fly. Significance levels of Wilcoxon test adjusted *p*-values are indicated: ns: p > 0.05; *: p <= 0.05; **: p <= 0.01; ***: p <= 0.001; ****: p <= 0.0001. Vera virus

and other viruses were not present in sufficient numbers in different datasets to statistically evaluate how their strand ratios changed over time.

Nevertheless, there was additional evidence from host-mapping reads that double-strandedness contributed to preferential RNA survival. Most host-mapping RNA derived from ribosomal RNA (rRNA), and most rRNA-mapping reads derived from the +strand (**Fig. 7**). In fresh samples, there was on average 698x more +strand rRNA than -strand rRNA. The small amount of -sense rRNA could have derived from anti-sense transcription of rRNA genes or pseudogenes[53]. In experimentally dried and museum samples, the +strand:-strand rRNA ratio dropped to 142:1 and 30:1, respectively ($p = 8.2x10^{-13}$ and $2.8x10^{-22}$; **Fig. 7**). This drop was driven by a relative decrease in +strand rRNA and a relative increase in -strand rRNA in older samples (**Supp**. **Fig. 9**). This pattern could be explained by the preferential survival of sense:antisense rRNA duplexes. However, even in old samples, there was 30x more +strand rRNA than –strand (**Fig. 7**), so sense:antisense duplexes could not account for all surviving rRNA.

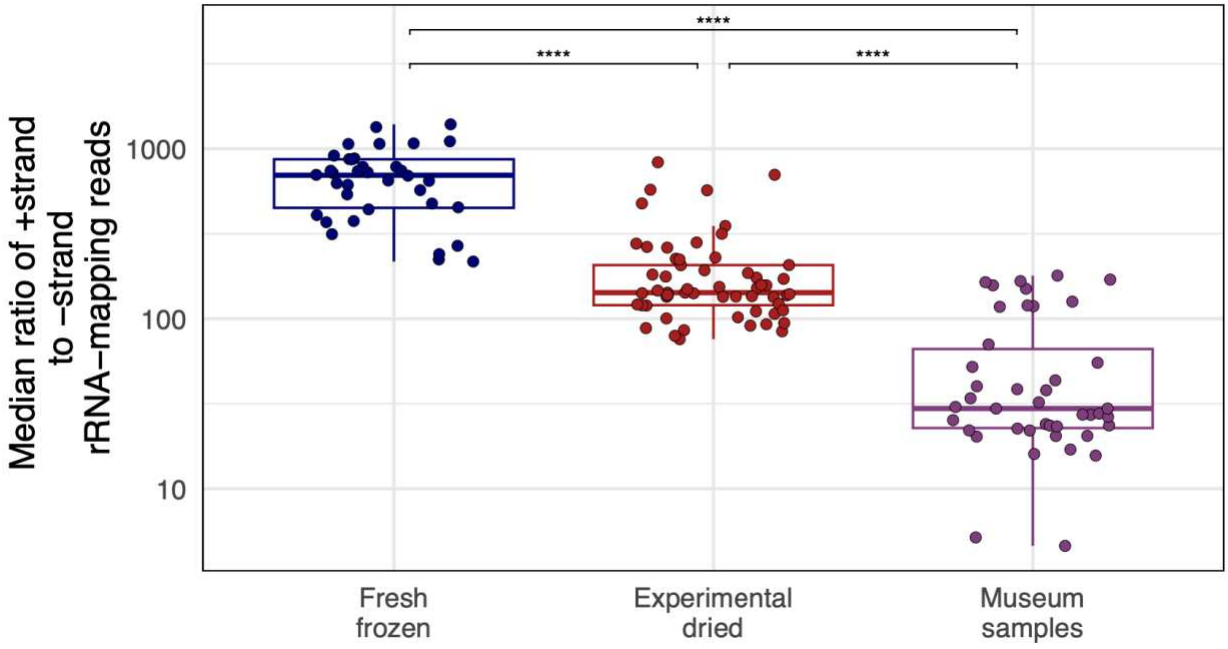


Figure 7: Strand ratios of surviving ribosomal RNA are consistent with preferential survival of dsRNA. The median ratio of +strand to –strand coverage of rRNA-mapping reads. Points represent individual-fly datasets. Adjusted *p*-values significance levels from Wilcoxon test are indicated.

Coverage levels across rRNAs varied more in old samples than they did in new samples, where coverage was relatively even (**Fig. 8**). For instance, coverage surrounding position 1500 of the 18S rRNA and position 3000 of the 28S rRNA had consistently lower coverage in old samples (**Fig. 8**). In contrast, in fresh samples coverage over these same regions was similar to average coverage.

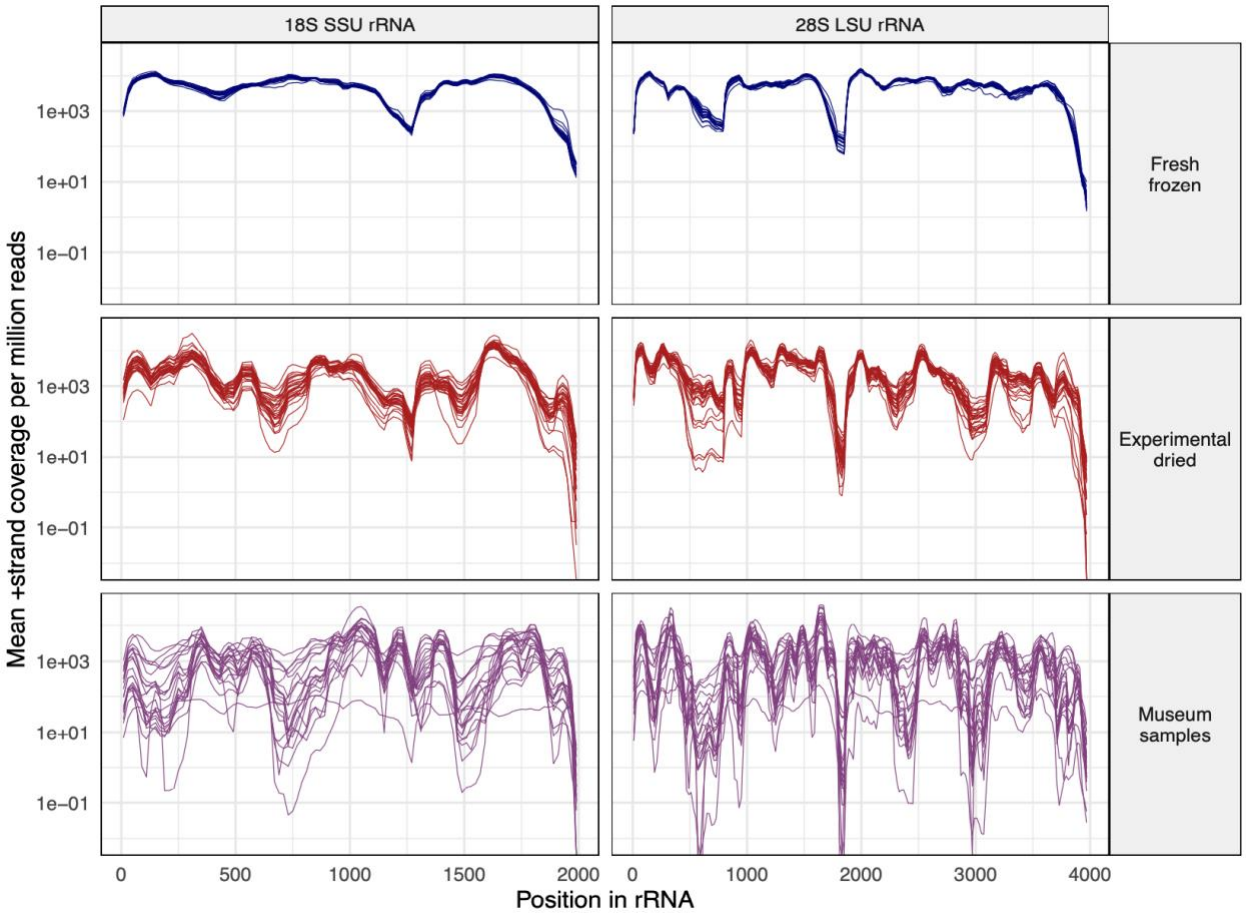


Figure 8: Certain regions of ribosomal RNA are underrepresented in old samples. The mean +strand rRNA coverage per million reads in 20 bp windows is plotted for *D. melanogaster* datasets. Each line represents a dataset from an individual fly. SSU: small subunit (18S) rRNA; LSU: large subunit (28S) rRNA.

To investigate why some rRNA regions might not survive as well as others, we took advantage of a high resolution cryo-electron-microscopy structure of the *D. melanogaster* 80S ribosome[54]. This structure includes ribosomal proteins and RNA (**Fig. 9A**). It is possible to resolve individual rRNA bases in the structure and to determine whether they are interacting with other bases, and whether they are present in the structure (**Fig. 9B**). We used this structure and RNApdbee software to individually assign the 5965 rRNA bases to one of four categories: paired with other rRNA bases (62.9%), unpaired (29.2%), present in higher-order secondary structures like pseudoknots (4.5%), or missing from the 3D structure altogether (3.4%; **Fig. 9B**) [54–56].

We calculated coverage of bases in each of these four categories relative to total median coverage (**Fig. 9C**). Paired bases had higher average coverage in older samples ($p = 1.9x10^{-19}$ in museum samples vs. fresh ones). Similarly, bases in higher-order secondary structures like pseudoknots had higher coverage in old samples ($p = 3.3x10^{-17}$ relative to fresh). In contrast, unpaired bases had lower average coverage in old samples ($p = 3.8x10^{-20}$ vs. fresh samples). Bases that were not captured in the 3D structure were the least well represented in older samples, with coverage levels 8x lower than overall average coverage in museum samples ($p = 1.0x10^{-19}$ relative to fresh frozen). Experimentally dried sample coverage levels were generally intermediate between fresh and museum samples, consistent with their intermediate age (**Fig. 9C**).

These patterns suggested that the molecular environment surrounding rRNA bases influenced the likelihood that they would survive. Being base-paired or in a higher-order RNA secondary structure protected bases, enabling them to survive over longer periods. In contrast, unpaired bases or bases not present in the 3D structure,

presumably because they were outside of the protective environment of the ribosome, were less likely to survive.

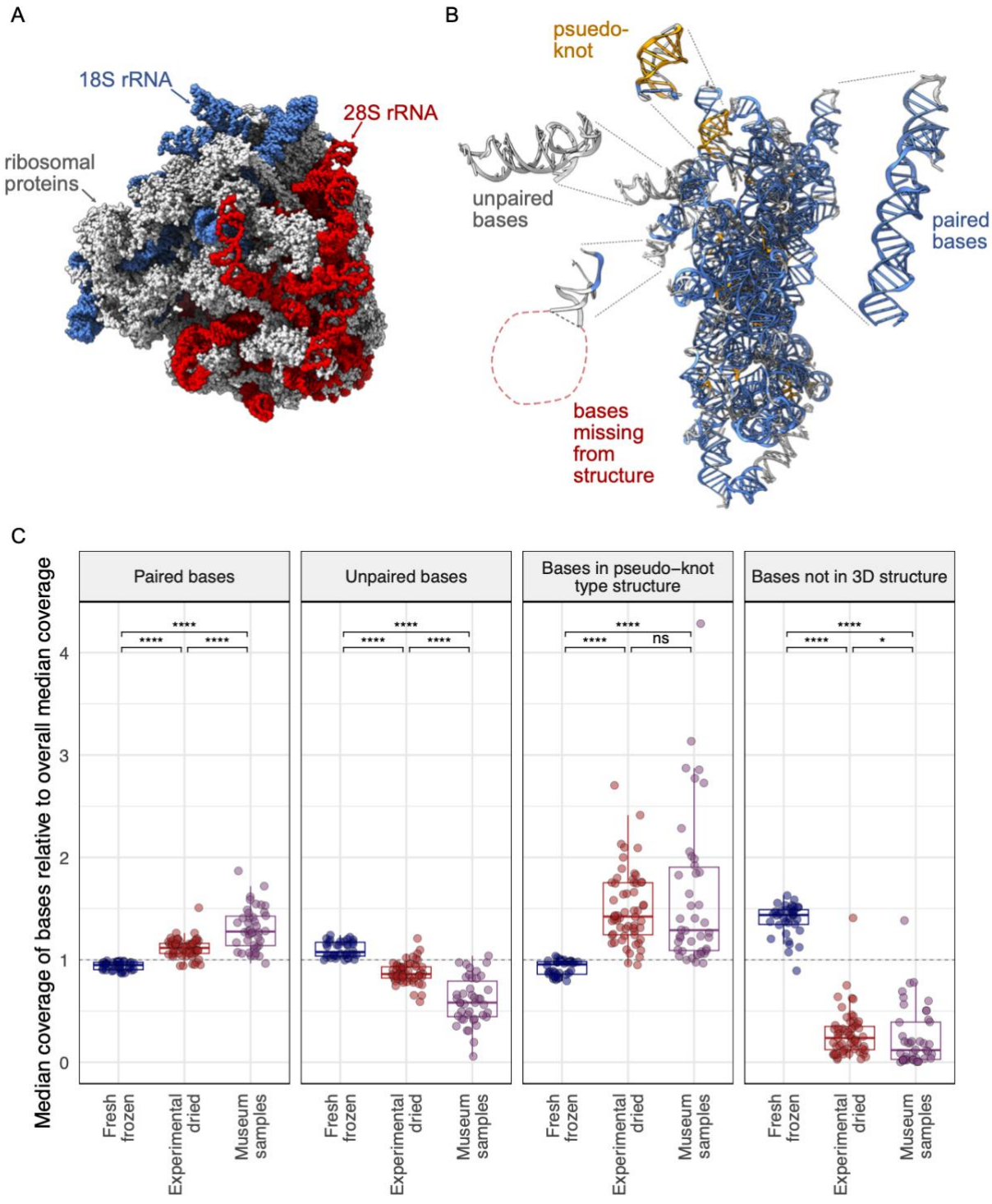


Figure 9: **Different types of ribosomal RNA exhibit differential survival in old samples.** (A) The structure of the *D. melanogaster* 80S ribosome from Anger at al, showing ribosomal proteins and the 18S and 28S ribosomal RNAs. (B) The 18S rRNA with individual bases color coded to indicate whether they are paired with other rRNA

bases (blue), unpaired (grey), or involved in higher order pseudo-knot type structures (gold). Some bases were not captured in the structure (red). Bases were binned into categories using RNApdbee software. The 18S rRNA is rotated relative to panel A. (C) The median coverage level of bases in the indicated categories relative to total median coverage is plotted. Each point represents a dataset from an individual *D. melanogaster* fly. Significance levels of Wilcoxon test adjusted *p*-values are indicated.

2.4.8 Different types of non-ribosomal RNA survive differentially in old samples

We also quantified the extent to which different types of non-ribosomal host RNA survived in old samples. We mapped reads from *D. melanogaster* samples to the *D. melanogaster* genome and used the ALFA software to quantify the different types of host RNA present in each dataset[57]. ALFA combines mapping information with genome annotation to assign mapped reads to one of a dozen RNA types (**Fig. 10**). As expected, most reads were categorized as rRNA, though there was relatively less rRNA in older samples (**Fig. 10A**; $p = 5.1 \times 10^{-6}$ for museum vs. fresh samples). There was also relatively less protein-coding mRNA in older samples (**Fig. 10B**; $p = 5.7 \times 10^{-4}$). Opposite strand RNA levels – that is, reads derived from the RNA strand opposite an annotated feature - were elevated in older samples, consistent with the preferential survival of sense:antisense transcript duplexes ($p = 1.8 \times 10^{-7}$). Highly structured transfer RNA (tRNA) levels were also elevated in older samples ($p = 5.9 \times 10^{-5}$)[58,59].

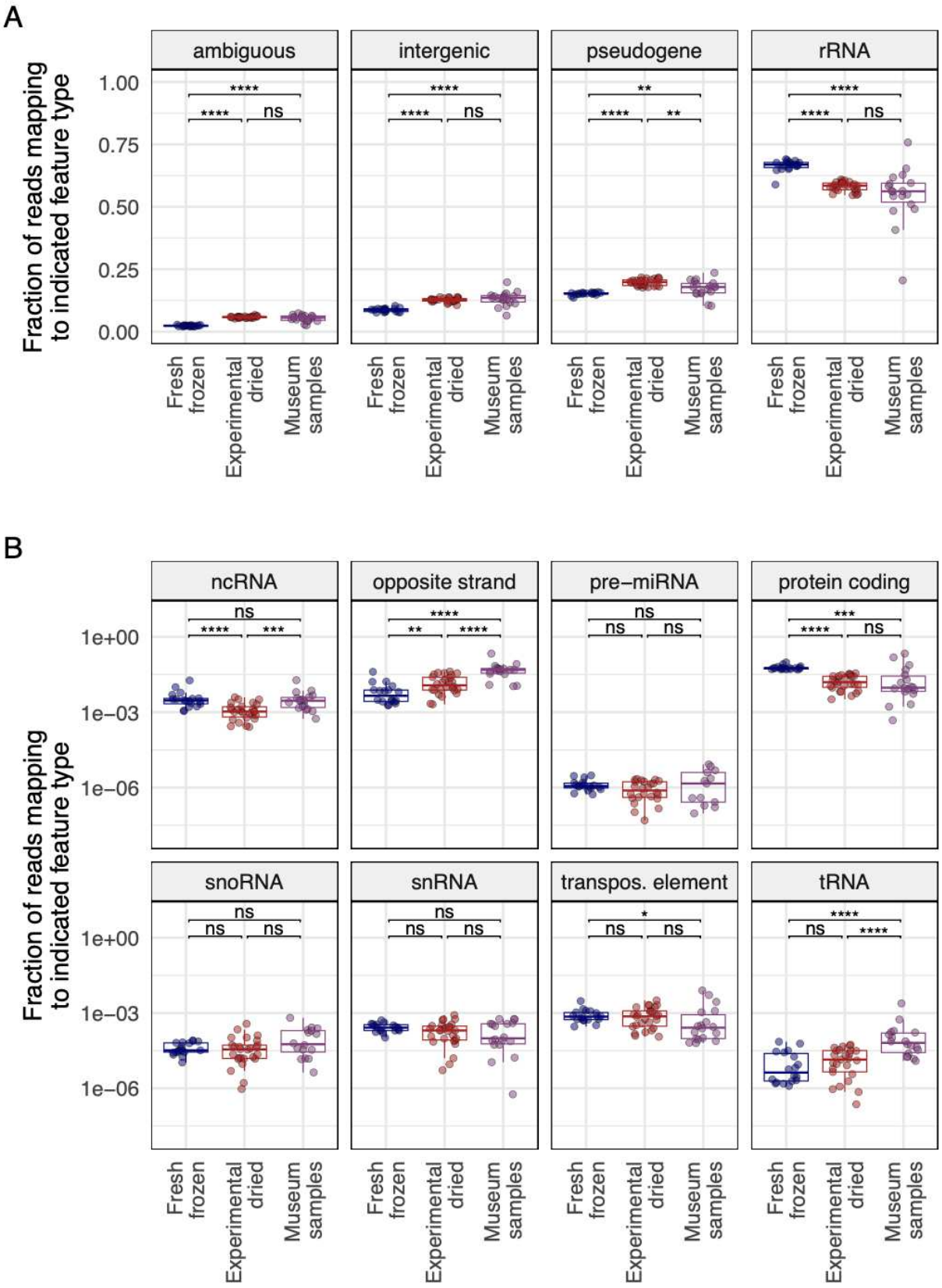

A
ambiguous
intergenic
pseudogene
rRNA
Fraction of reads mapping to indicated feature type
1.00
0.75
0.50
0.25
0.00
****
ns
**
Fresh frozen
Experimental dried
Museum samples
B
ncRNA
opposite strand
pre−miRNA
protein coding
snoRNA
snRNA
transpos. element
tRNA
1e+00
1e−03
1e−06
***
*

Figure 10: **Certain types of RNA survive better in old samples.** The fraction of reads in *D. melanogaster* datasets mapping to the indicated RNA types is plotted. Each point represents a dataset from an individual fly. Adjusted *p*-values significance levels from Wilcoxon test are indicated. (A) Types of RNA present at >5% median abundance. (B) Types of RNA present at <5% abundance. Abbreviations for different RNA types: ncRNA: non-coding RNA; miRNA: micro RNA; snoRNA: small nucleolar RNAs; snRNA: small nuclear RNA; tRNA: transfer RNA.

2.4.9 Old RNA is chemically damaged

Old DNA molecules are fragmented and chemically damaged[60]. Polymerases can misincorporate bases when copying damaged templates, causing damage to manifest as substitutions in library molecules[61–63]. For example, deaminated cytosines, common in ancient DNA, result in C-to-T substitutions in library molecules. In fact, signatures of chemical damage in sequence reads are taken as evidence that reads derive from old DNA and not from contaminating contemporary DNA.

We quantified mismatch patterns in mapped reads to look for signatures of chemical damage in old RNA (**Fig. 11**). Mismatch frequencies varied in different sample types. Some differences in mismatch frequencies could be attributed to samples being sequenced on different sequencing runs (we sequenced museum samples separately from other samples to avoid the possibility of read misassignment from index-hopping) [64]. The largest difference between old and new samples was an elevated frequency of C-to-T substitutions in reads from museum samples. The median level of C-to-T mismatches in datasets from museum samples was 40x higher than in fresh-frozen datasets (2.0×10^{-3} frequency of C-to-T mismatches vs. 5.0×10^{-5} (**Supp. Fig. 10**; $p = 1.0 \times 10^{-10}$). This increased frequency of C-to-T substitutions is consistent with deamination of cytosines in old RNA.

A-to-G substitutions were the next most elevated type of mismatch in old samples (**Fig. 11**). A-to-G substitutions occurred at a rate 4.4-fold higher in museum samples than in fresh datasets ($5.7x10^{-4}$ vs. $1.3x10^{-4}$; $p=1.6x10^{-10}$). Such substitutions can result from spontaneous deamination of adenine to hypoxanthine[60,65].

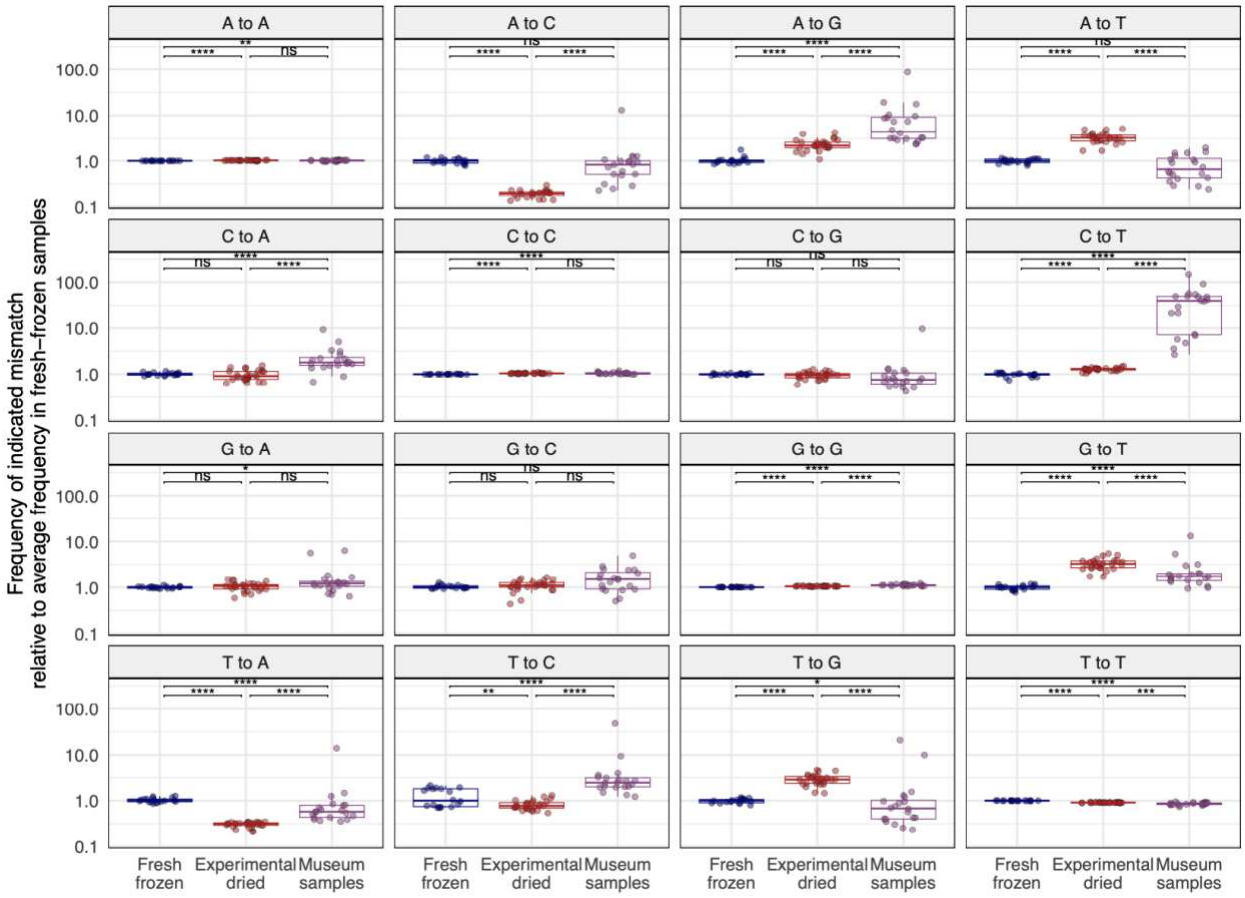


Figure 11. Old RNA is chemically damaged and exhibits mismatch patterns consistent with cytosine and adenine deamination. Mismatches in rRNA-mapping reads from *D. melanogaster* datasets were quantified and the frequency of each mismatch type relative to the median frequency in fresh-frozen datasets is plotted. Each point represents a dataset from an individual fly. Adjusted *p*-values significance levels from Wilcoxon test are indicated.

2.4.10 Host-mapping reads exhibit species specificity

We took advantage of the fact that our museum samples derived from 4 different species to investigate the possibility that host-mapping reads – specifically rRNA-mapping reads – might result from contemporary contamination. The likely source of fly-mapping contamination would be the *D. melanogaster* that we rear and study in our lab[48]. The 18S and 28S ribosomal RNA sequences of *D. melanogaster* and *D. simulans* share over 99% pairwise identity which makes them difficult to distinguish by competitive read mapping. We therefore collected and mapped reads to the relatively variable region of the ribosomal RNA between the 18S and 5.8S genes (the internal transcribed spacer 1, ITS) of the 4 fly species that we sampled (**Table S1**; **Fig. 12; Supp. Fig. 11**). All ITS-mapping reads in museum datasets mapped to the expected species ITS sequence and no reads mapped to an unexpected species (**Fig. 12**). Positive and negative control samples were as expected, with water negative control datasets containing no reads mapping to any of these 4 ITS sequences (**Supp. Fig. 11**). If fly rRNA-mapping reads were from contamination, we would have expected there to be *D. melanogaster*-mapping reads in the non-*melanogaster* and negative control datasets; instead, there were none. The perfect concordance of ITS mapping and sample species supported the conclusion that the host-mapping reads in our datasets were indeed from museum samples.

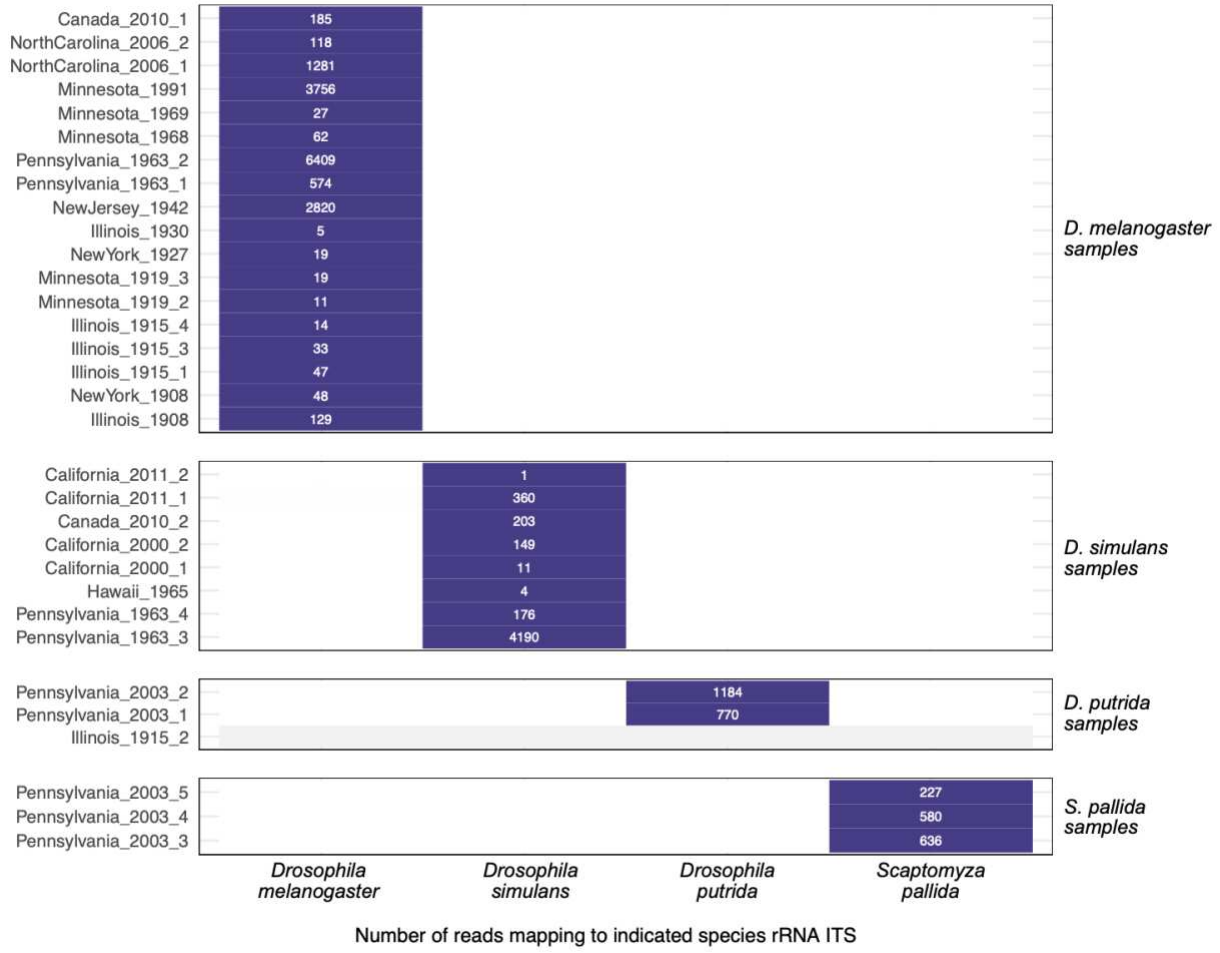


Figure 12: Host-mapping reads exhibit species specificity. The number of reads mapping to the indicated species rRNA internal transcribed spacer 1 (ITS) sequence for each museum sample dataset is shown. Samples are grouped according to their molecular species assignment (**Table S1**). Datasets with no ITS-mapping reads are represented with grey.

2.5 Discussion

In this study we took a two-pronged approach to investigate RNA stability in old biological samples. In flies and mosquitoes stored dry at room temperature for weeks or months, RNA grew increasingly fragmented, but RNA yields did not significantly decrease, and cellular and viral RNA remained detectable. Perhaps the most unexpected result from this experiment was the limited difference in RNA quality and

quantity between dried and frozen samples (**Fig. 1; Supp. Figs. 2, 4**). We recovered RNA from dried flies that were from as far back as the 1890s and used this RNA to make sequencing libraries. We recovered coding complete virus sequences from 11 old specimens and identified two novel virus sequences. We also showed that certain types of RNA tended to survive in old samples, including dsRNA and RNA in secondary structures or ribonucleoprotein complexes.

There were limitations to the experimental arm of our study. First, RNA recovery from individual flies and mosquitoes, even fresh ones, was variable. Second, we used outbred populations of flies and mosquitoes with different levels of virus infection between individuals. In future studies, optimization of our RNA isolation protocol to decrease yield variability between individual flies and mosquitoes would be useful. It would also be better to use pinned specimens from inbred populations with less variable infection phenotypes.

We found that not all RNA survived equally well and described molecular features that influenced how well different types of RNA survived. dsRNA survived better than ssRNA. Surviving RNA duplexes derived from intermolecular base pairing, such as sense-antisense duplexes (**Figs. 6, 7A, 10B**) and from intramolecular base pairing, such as paired rRNA bases (**Fig. 9C**). The preferential survival of dsRNA likely reflects decreased susceptibility to hydrolysis[66]. However, double strandedness could not explain all RNA survival, as RNA from certain viruses remained largely +strand (**Fig. 6**) and unpaired rRNA survived (albeit less than paired rRNA; **Figs. 7, 9**). It may be that ribonucleoprotein complexes like virus particles and ribosomes provide a protective environment in which RNA can survive longer.

Like old DNA, old RNA was fragmented and chemically damaged (**Figs. 2C, 11**). And, like DNA, deamination was the most evident type of damage (**Fig. 11**). Other types of damage may be common in old RNA but undetectable by sequencing. The extent to which ribonucleases, deaminases, and other enzymes involved in normal RNA processing continue to function in dead cells is unclear, but it is likely that most RNA fragmentation and damage in old samples results from spontaneous chemical rather than enzymatic degradation.

Environmental factors like temperature and humidity likely play a major role in the survival of RNA. It was notable that none of the 12 specimens from tropical Hawai'i yielded enough RNA to be detected fluorometrically and only two generated sequenceable libraries (**Fig. 2b, Table 1 & Supp. Fig. 5**). Humidity might be more important than temperature, as RNA was recovered from 1000-year-old corn in Arizona, where temperatures fluctuate to extremes, but humidity is very low[15]. If this were true, RNA might survive better in drier environments and in quickly desiccating samples like the small insects that are the focus of this work. Even for much larger samples like wooly mammoths, dehydration has been proposed to promote preservation of nuclear architecture and biological macromolecules[67].

The virus sequences we recovered revealed that the *D. melanogaster* virome has not changed much over the last 120 years. All the viruses that we detected in *D. melanogaster*, except for Puslinch virus, had already been described in contemporary *D. melanogaster*[25,26]. For the most part the old sequences were closely related to contemporary sequences (**Table 2**). For example, the viruses infecting the fly collected in 1908 were >98% identical to known *D. melanogaster* viruses (**Fig. 4; Table 2**).

The discrepant phylogenetic positions of the 1930 galbut virus segments is consistent with reassortment[68]. The 1930 galbut virus RNA 3 sequence that is positioned on its own branch may correspond to an extinct RNA 3 lineage, or simply one that has not been sampled before (**Fig. 5C**). Similarly, the endogenous galbut virus RNA 1 sequence in these phylogonies may represent an extinct lineage or an ancestor of contemporary galbut virus RNA 1 sequences (**Fig. 5A**)[49]. A potential benefit of large-scale sequencing of old samples is that it could provide estimates of how frequently virus lineages go extinct from a particular host.

Contamination from contemporary nucleic acid has been a long-standing concern for those sequencing ancient DNA[69]. There was evidence of low-level contamination in our datasets (**Fig. 3**), and it is difficult to completely avoid contamination in NGS datasets, even when sequencing new samples[70,71]. But, there were numerous lines of evidence to support the idea that our conclusions were based on sequences from actual old RNA and not from contemporary contamination. Old RNA was short and chemically damaged; properties that are shared with old DNA[60]. Characteristics of RNA from experimentally dried samples were generally intermediate between those of fresh RNA and museum RNA, consistent with their intermediate age (**Figs. 6-11**). We recovered different virus sequences from different old samples, and reads in individual datasets supported the same distinct virus sequences (**Table 2**). The virus sequences we recovered were for the most part known *Drosophila*-infecting viruses. Although many virus sequences from old samples were similar to existing contemporary sequences, most were not identical (**Table 2**). The exceptions were three vera virus RNA 1 sequences that were identical to existing sequences. But the RNA 2 and chaq-like

sequences from these samples were not identical to any contemporary sequences. The two new virus sequences we recovered were divergent, but their closest relatives were from other insects (**Table 2; Fig. 5**). The tobacco mosaic virus sequence from the 1908 dataset was an exception, which we can't explain except to suggest that this specimen was contaminated during handling by a smoker, as individual cigarettes were found to contain as many as 10^9 TMV RNA copies[41]. There was no peak in galbut virus coverage corresponding to our diagnostic RT-qPCR amplicons, which would be a prime candidate source of contaminating galbut virus sequences (**Supp. Fig 6**)[30,48]. All museum samples that were positive for galbut virus by RT-qPCR and generated a sequenceable library were also positive by sequencing (**Supp. Table 1**). Negative control datasets were uniformly free of virus and rRNA-mapping reads (**Fig. 4**; **Supp Fig. 11**). Galbut virus sequences from *D. simulans* museum samples clustered with contemporary *D. simulans* sequences, and the same for *D. melanogaster* sequences (**Fig. 5**).

Indeed, the species-specificity of host and virus sequences (**Figs. 5 & 12**) provides some of the strongest support for the idea that our analyses are based on nucleic acid from the actual old samples. If our conclusions were based on contaminating contemporary sequences, it would be necessary to invoke a complicated mechanism where contaminating host and virus nucleic acids sorted themselves by species into each sample (**Figs 5 & 12**). The overall apparent lack of contaminating virus and host sequences was consistent with the fact that all extractions were done in biosafety cabinets following a pre-PCR to post-PCR workflow, including DNase treatment before RT-qPCR and library preparation.

RNA can persist in biological samples over decades or centuries without freezing or fixation. The half-life of RNA in dead cells is clearly much longer than in living cells. The millions of dried specimens in museums represent a valuable source of RNA for reconstructing historic virus-host interactions[14,15]. It is tempting to speculate that the long-term survival of certain RNA might approach that of DNA[72]. It is time to move beyond the incorrect idea that old, dried samples are not a good source of useful RNA.

2.6 Materials & Methods

2.6.1 Sample Collection

Pinned Specimens: FoCo-17 *D. melanogaster* were collected and pinned using standard entomological collection techniques with storage at room temperature or stored frozen at -80°C[73]. Poza Rica *Aedes aegypti* mosquitoes were reared using standard lab techniques until adulthood[74]. Prior to blood feeding mosquitoes were collected and pinned or frozen.

Museum Specimens: Samples were obtained from museums and other institutions containing entomological specimens located in the United States of America and Canada. Altogether, 46 drosophilid samples, mainly *D. melanogaster* or *D. simulans*, were selected for the study. Museum species assignments were checked using molecular data as described below. One sample from California, California_2000_2, was initially identified as *D. melanogaster*, but our sequence data indicated that it was *D. simulans* (these species can be difficult to distinguish morphologically). Location, date of sample collection, sample storage type, extraction

method, RNA quality and RNA extraction concentration are summarized in **Figure 2**, **Table 1** and **Supplemental Table 1**

2.6.2 Sample Workflow

Deliberate protocols were used to minimize contamination during sample processing and library construction. Sample processing and library construction moved from dedicated pre-PCR rooms to a post-PCR room. RNA extractions were performed in a dedicated pre-PCR sample extraction room in a Class II type B2 biosafety cabinet which protects both personnel and samples. Initial library preparation steps, including reverse transcription, ligation, and setup of library amplification reactions were completed in a separate PCR setup room in an AirClean PCR workstation equipped with a HEPA filter (AirClean Systems, AC600 Series). Finally, library amplification and subsequent cleanup and pooling steps were performed in a third post-PCR lab space.

2.6.3 RNA Extraction

Non-Destructive Sampling: For samples for which destructive sampling was prohibited, we adapted the protocol described by Santos et. al., 2018[35]. Briefly, individual samples were placed in a PCR tube containing 200µL of solution containing 200mM Tris HCl, 250mM NaCl, 25mM EDTA, 0.5% SDS and 400 µg/mL proteinase K. Samples were then incubated at 56°C for 16 hours. After incubation the supernatant was moved to a new tube for RNA isolation and the specimen was transferred to a cryovial containing 80% ethanol. RNA was purified using the Zymo RNA Clean & Concentrator Kit as described below.

Ethanol Stored Sample Rehydration: Four samples arrived in 70% ethanol. To prepare these samples for extraction, each specimen was rehydrated using a gradient

of decreasing concentrations of ethanol. The specimens were suspended in 200µL 70%, 50%, 30% and 10% ethanol for 15 minutes on ice with agitation every 2-3 minutes prior to a final incubation in water for 15 minutes before RNA extraction following the destructive sampling method described below.

Destructive Sampling: A phenol/chloroform approach with mechanical disruption was used for RNA isolation of the experimentally dried and frozen samples as well as the museum specimens that could be destructively sampled. Briefly, each specimen was placed in a 2.0mL tube with a small BB (McMaster Carr, 1598K22) and 500µL of Trizol was added and incubated for 5 minutes. Tubes were then shaken in a TissueLyser (Qiagen, 9003240.) at 30 Hz for 3 minutes. 100µL of chloroform was added, thoroughly mixed and incubated for two minutes prior to being spun at 12,000xg for 10 minutes. The aqueous phase was retained for RNA purification using the Zymo RNA Clean & Concentrator-5 kit following manufacturer directions for the capture of small RNA fragments with slight modification (Zymo Research, R1013). Briefly, 225µL of sample was mixed with 225µL 100% ethanol and 225µL RNA binding buffer. Samples were transferred to spin columns with collection tubes and spun. To capture small RNAs, the flow through from the first spin was retained and mixed with an equal amount of 100% ethanol. This was transferred to the spin column and centrifuged. A wash using RNA wash buffer was done prior to a DNAse treatment following manufacturer recommendations (Zymo Research, E1010). After DNAse treatment, two more washes were complete prior to elution in 30µL of nuclease free water. All spins were performed at 9,000xg for 1 minute unless otherwise noted. For all extraction types,

FoCo-17 male and female flies were used as a positive control and an empty tube was used as a negative control.

2.6.4 RNA Quality and Quantification

RNA quality was assessed using a nanodrop spectrophotometer (Thermo Scientific ND2000USCAN) and RNA concentration was determined using the Qubit high-sensitivity RNA reagent (ThermoFisher, Q32852). RNA length was measured using the Agilent Tapestation High Sensitivity RNA screening tapes and reagents (Agilent, 5067-5579 & 5067-5580).

2.6.5 RT-qPCR

Pinned Samples: Several targets were assessed for presence/absence and evidence of degradation using RT-qPCR (primer sequences in **Supplemental Table 3**). cDNA was generated using 5.5µL RNA, 10µM random hexamer primers (IDT), 10µM dNTPs and water to 13µL and incubated at 65°C for 5 minutes. Next, 4µL 5X FS buffer, 1µL 0.1M DTT and 1µL reverse transcriptase was added and incubated at 50°C for 60 minutes followed by a heat inactivation step at 80°C for 10 minutes after which each sample was diluted in 90µL nuclease-free water. qPCR was conducted using NEB Luna Universal qPCR Mastermix following manufacturer recommendations and cycling conditions (NEB, M3003).

2.6.6 PCR controls

RNA from a pool of FoCo-17 flies was reverse transcribed to create a cDNA positive control for fly samples. For mosquito RT-qPCR, positive control cDNA was reverse transcribed from individual colony mosquitoes. We included these positive

controls and three negative controls in all qPCR runs. Negative controls consisted of an extraction blank (a no sample extraction), a no-sample RT control, and a no template qPCR control. The use of negative controls at each step would have allowed us to determine when cross-contamination had been introduced if it had been. Positive/negative status was determined using the melt curve (within 1 degree of the positive control) for each sample and target. Agarose gel electrophoresis of qPCR products was used to determine positive/negative status when Ct and melt curve were ambiguous.

Museum Collections: We used RT-qPCR to determine if any samples had detectable levels of galbut virus RNA and RpL32 mRNA. To increase our ability to detect highly fragmented RNA, we designed a second set of primer that target a small region of the galbut virus RNA 1 and *D. melanogaster* mRNA RpL32. A 2% ethidium bromide agarose gel was used to confirm positive/negative status. Galbut virus positive female and male FoCo-17 flies were used as positive controls and water was used as a negative control.

Supplemental Table 3: Primer sequences used in this study.

Primer pair name (#s in lab collection)	Forward	Reverse	Reference
Galbut virus (long) (#1600/1601)	CCGTGAAGCAAGGAATCAAT	TGCCGATTTTCTGCTCTTTT	[30]
Galbut virus-short (1948/1949)	AGAAGATGTGCTGTAGTGACAC	CGTGGAATTCCGAACGGCTA	This study

RpL32- long (1012/1013)	TGCTAAGCTGTCGCACAAATGG	TGCGCTTGTTCGATCCGTAAC	[75]
RpL32- short (1950/1951)	GCGCTTGTTCGATCCGTAAC	GCCGCTTCAAGGGACAGTAT	This study
Thika virus (1604/1605)	CAGCAGGTCCCTTGCTAAAG	TGGTCAGCATATGACCGAAA	[30]
Nora virus (1592/1593)	GCACCTGGTCGATTGAATCC	CGTTCAGGGCATAGTCAAGC	[30]
La Jolla virus (1692/1693)	ACCGTATGGCGTCGTACTTC	AAAGTATCAGCAGCGCGAAT	[30]
Actin (#1826/1827)	CGTTCGTGACATCAAGGAAA	GAACGATGGCTGGAAGAGAG	[76]
Verdadero virus (#1662/1663)	ATATGGGTCGTGTCGAAAGC	CACCCCGAAATTTTCTTCAA	[30]
Guadeloupe mosquito virus (#1688/1689)	TTGTAAATCCCCCTGTGCTC	CGCAAAAATTGGGTTTGAGT	This study

2.6.7 NGS Library Preparation

Museum collection samples were prepared for sequencing using the Kapa Biosystems Kapa RNA Hyper Prep kit (Roche) following manufacturer recommendations with modification of the fragmentation/priming step. Since the RNA was already highly fragmented, samples were minimally incubated at 65°C for 1 minute during the fragmentation/priming step. Due to low starting concentrations, 12 cycles

were used for library amplification. FoCo-17 fly RNA was used as a positive control and water was used as a negative control for library preparation and sequencing. Controls were sequenced on a separate run to avoid index hopping from the positive control sample[77]. Sample pooling was determined using High Sensitivity DNA Qubit reagents and final library quality control was done using an Agilent D1000 HS Tapestation and KapaQuant reagents following manufacturers recommendations (Roche, 07960140001). Museum samples were sequenced on three NextSeq 500 runs; one high output 75 cycle run and two mid output 150 cycle runs. All museum specimens were run on the first NextSeq run, runs two and three consisted of samples that we wanted to increase coverage of near complete virus sequences. Experimentally dried samples were sequenced on a high output 75 cycle run. The two positive control samples were sequenced on a 300 cycle MiSeq run. Additional fresh frozen samples were sequenced on a NextSeq mid output 150 cycle kit.

2.6.8 Data Analysis

Experimental Collections: All statistical analysis and data visualization, unless otherwise noted, was done using R/RStudio. Data cleaning and data visualization were done using the tidyverse package[78]. All scripts are available in the GitHub repository linked below. A p-value of $\alpha = 0.05$ or below was used as our significance threshold. RNA Concentration and Length: Four individual multiple linear regression models (MLRs) were constructed to test for statistical significance in RNA concentration and length between dried and frozen flies and mosquitoes over 72- and 52-weeks. The predictor time was treated as a continuous variable and the predictor storage type was converted to factor. A balanced one-way analysis of variance (ANOVA) was used to test

for significance using the car package[79]. The performance package was used to test model assumptions[80]. RT-qPCR: Raw qPCR data files were cleaned and compiled using the SangerTools R package and a custom R script[81]. Two-tailed Students T-tests from the rstatix package were used to determine statistical significance of dried and frozen Ct values[82]. P-values were adjusted for multiple hypothesis testing.

Identification of virus sequences in old samples: We used our lab's previously described metagenomic classification pipeline (https://github.com/stenglein-lab/taxonomy_pipeline) to identify and validate virus sequences in NGS datasets[30]. In brief, adapters and low quality reads were trimmed using cutadapt 3.5[83]. Fastqc was used to assess post-collapsed read quality 0.11.9[84]. Host reads were removed using bowtie2 2.4.5 and the remaining reads were assembled using spades 3.15.4[85,86]. Virus sequences were identified using BLASTn 2.12.0 and a BLASTx search using diamond 2.0.14 was used to identify novel sequences[87,88]. Draft sequences were validated by remapping of trimmed reads using bwa mem aligner version 0.7.17[89]. Final sequences were submitted to the NCBI nucleotide database and raw NGS data to the NCBI sequence read archive repository.

This metagenomic classification pipeline was also used to taxonomically assign non-host contigs and unassembled reads. Counts assigned to individual taxa were normalized to the number of quality- and adapter-trimmed reads.

Molecular species identification using cytochrome oxidase subunit 1 competitive mapping. The sequence of this mitochondrial-encoded gene is commonly used to distinguish related species[90]. We used the *D. melanogaster* CO1 coding sequence from NC_024511 as a BLASTN query to the NCBI nucleotide database to identify other

drosophilid CO1 sequences[91]. We restricted results to sequences from Drosophilidae and to alignments with >90% query coverage and E-value <1x10^{-50}. We used cd-hit-est to collapse sequences that shared >99% pairwise nucleotide identity to create a set of 545 representative CO1 sequences[92]. We mapped adapter and quality trimmed reads to these sequences using bowtie2 v.2.4.4 in end-to-end mode and retained alignments with a mapping quality ≥ 20.

Maximum Likelihood Trees: All available galbut virus and sigma virus sequences from individual flies (and not pools) with collection date metadata were downloaded from GenBank. Sequences in the Puslinch virus phylogeny include all available L protein sequences in the NCBI RefSeq protein database from the *Peribunyaviridae* family. Sequences in the sobemo-like virus tree include RdRp protein sequences from viral genomes identified by a BLASTN search of the NCBI nucleotide database. Alignments were generated using MAFTT v7.490 with default parameters. IqTree v1.0 (http://iqtree.cibiv.univie.ac.at) was used to generate maximum likelihood trees[93]. The following parameters were used on the web interface version of IqTree: substitution model- auto, Free-rate Heterogeneity- yes, bootstrap analysis- ultrafast with all other options set to default. FigTree was used for tree visualization (http://tree.bio.ed.ac.uk/software/figtree/). Trees were midpoint rooted.

Analysis of strandedness of virus-mapping reads. To tabulate the numbers of positive and negative strand reads mapping to virus sequences, we mapped quality- and adapter-trimmed reads to assembled virus sequences using the bwa mem aligner version 0.7.17[89]. The orientation of each mapped read was determined from the 0x10 bit flag of the resulting sam format output.

For all analyses of virus- and host -mapping reads, values were summarized and visualized using the tidyverse R packages[78]. Wilcoxon tests were used to statistically evaluate differences between groups using the rstatix R package. In all cases values were determined to be non-normally distributed by Shapiro-Wilk test[82]. *P*-values were adjusted for multiple testing using the Holm–Bonferroni method and were indicated on plots using the ggpubr R package using the following significance levels: ns: p > 0.05; *: p <= 0.05; **: p <= 0.01; ***: p <= 0.001; ****: p <= 0.0001[94]. Analyes were implemented as nextflow workflows available at: https://github.com/LKeene/Old_Collections_Figures/[95].

Analysis of ribosomal rRNA-mapping reads. Adapter- and quality-trimmed reads from *D. melanogaster* datasets were mapped to the 18S and 28S rRNA sequences from Anger et al (chains B5 and A2 from[54]) using bwa mem aligner version 0.7.17[89]. Per-base, per-strand coverage was calculated using the BamToCov tool v2.7.0[96]. The *D. melanogaster* 80S ribosome structure was visualized using ChimeraX v.1.7.1[97]. Bases were binned into one of four categories (paired, non-paired, higher-order structure, absent from structure) by feeding this structure (protein data bank accession 4V6W) into RNApdbee 2.0[55].

*Analysis of non-ribosomal RNA types. A*dapter- and quality-trimmed reads from *D. melanogaster* datasets were mapped to the *D. melanogaster* genome version BDGP6.32[98] using bwa mem aligner version 0.7.17[89]. Mapping output and genome annotation were input into ALFA v1.1.1 to tabulate the fractions of reads mapping to each of the annotated genomic biotypes[57].

Analysis of mismatch signatures of chemical damage. Trimmed reads from *D. melanogaster* datasets were mapped to rRNA as above. Mismatches of bases with a basecall quality score >30 were tabulated.

Competitive mapping to ribosomal internal transcribed spacer sequences. We collected sequences between the 18S and 5.8S rRNA genes for the four museum species. Sequences correspond to Genbank accessions: *D. melanogaster*: NR_133558.1:2861-722; *D. simulans*: NGVV02000020.1:5755-5071; *D. putrida*: AF184042.1:153-611; *S. pallida*: JAECXP010000218.1:7393-7904. We mapped adapter and quality trimmed reads to these sequences using bowtie2 v.2.4.4 in end-to-end mode and retained alignments with a mapping quality ≥ 30.

2.7 Acknowledgements

The authors would like to acknowledge the museum institutions and staff for their willingness to share specimens for this work: Robin Thomson, University of Minnesota Insect Collection; Timothy McCabe, New York State Museum; Chris Paradise, The Davidson College Entomology Collection; Matt Gimmel, Santa Barbara Museum of Natural History; Camiel Doorenweerd, University of Hawaii Insect Museum; Jayme Sones, The Centre for Biodiversity and Genomics at the University of Guelph; Laura Porturas, The Frost Entomology Center at Pennsylvania State University; Tommy McElrath, Illinois Natural History Survey. The authors would also like to thank Dr. Brian Foy for providing *Ae. aegypti* mosquitoes, and Dr. Rebekah Kading for the supplies to pin and store dried insects. We thank Darren Obbard, Dan Sloan, and Jeff Wilusz for helpful discussions.

2.8 Funding

NSF IOS 2048214 (MDS), NIH T32GM132057 (AHK), Computational resources supported by NIH/NCATS Colorado CTSA Grant Number UL1 TR002535

2.9 Data availability

All metadata, RT-qPCR raw data, plots, and analysis and visualization code can be found at: https://github.com/LKeene/Old_Collections_Figures. Raw sequence data has been uploaded to the NCBI SRA database; assembled viral sequences have been deposited in GenBank; both can be found under NCBI BioProject accession PRJNA1034757.

CHAPTER 2: REFERENCES

1. Leonardi, M. *et al.* Evolutionary Patterns and Processes: Lessons from Ancient DNA. *Syst Biol* **66**, syw059 (2016).
2. Orlando, L. & Cooper, A. Using ancient DNA to understand evolutionary and ecological processes. *Annu Rev Ecol Evol Syst* **45**, 573–598 (2014).
3. Green, R. E. *et al.* A draft sequence of the neandertal genome. *Science (1979)* **328**, 710–722 (2010).
4. Duchêne, S., Ho, S. Y. W., Carmichael, A. G., Holmes, E. C. & Poinar, H. The Recovery, Interpretation and Use of Ancient Pathogen Genomes. *Current Biology* **30**, R1215–R1231 (2020).
5. Ng, T. F. F. *et al.* Preservation of viral genomes in 700-y-old caribou feces from a subarctic ice patch. *Proc Natl Acad Sci U S A* **111**, 16842–7 (2014).
6. Fromm, B. *et al.* Ancient microRNA profiles of 14,300-yr-old canid samples confirm taxonomic origin and provide glimpses into tissue-specific gene regulation from the Pleistocene. *RNA* **27**, 324–334 (2021).
7. Faria, N. R. *et al.* The early spread and epidemic ignition of HIV-1 in human populations. *Science (1979)* **346**, 56–61 (2014).
8. Taubenberger, J. K., Reid, A. H., Krafft, A. E., Bijwaard, K. E. & Fanning, T. G. Initial Genetic Characterization of the 1918 “Spanish” Influenza Virus. *Science (1979)* **275**, 1793–1796 (1997).
9. Belyi, V. A., Levine, A. J. & Skalka, A. M. Sequences from Ancestral Single-Stranded DNA Viruses in Vertebrate Genomes: the Parvoviridae and Circoviridae Are More than 40 to 50 Million Years Old . *J Virol* **84**, 12458–12462 (2010).
10. Ávila-Arcos, M. C. *et al.* One hundred twenty years of koala retrovirus evolution determined from museum Skins. *Mol Biol Evol* **30**, 299–304 (2013).
11. Enard, D. & Petrov, D. A. Ancient RNA virus epidemics through the lens of recent adaptation in human genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences* **375**, 20190575 (2020).
12. Yang, E. *et al.* Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res* **13**, 1863–72 (2003).
13. Fordyce, S. L., Kampmann, M.-L., van Doorn, N. L. & Gilbert, M. T. P. Long-term RNA persistence in postmortem contexts. *Investig Genet* **4**, 7 (2013).
14. Smith, O. *et al.* A complete ancient RNA genome: Identification, reconstruction and evolutionary history of archaeological Barley Stripe Mosaic Virus. *Sci Rep* **4**, 4003 (2014).
15. Peyambari, M., Warner, S., Stoler, N., Rainer, D. & Roossinck, M. J. A 1,000-Year-Old RNA Virus. *J Virol* **93**, e01188-18 (2019).

16. Mármol-Sánchez, E. *et al.* Historical RNA expression profiles from the extinct Tasmanian tiger. *Genome Res* **33**, 1299–1316 (2023).
17. Cobb, N. S. *et al.* Assessment of North American arthropod collections: prospects and challenges for addressing biodiversity research. *PeerJ* **7**, e8086 (2019).
18. Watts, P. C., Thompson, D. J., Allen, K. A. & Kemp, S. J. How useful is DNA extracted from the legs of archived insects for microsatellite-based population genetic analyses? *J Insect Conserv* **11**, 195–198 (2007).
19. Heintzman, P. D., Elias, S. A., Moore, K., Paszkiewicz, K. & Barnes, I. Characterizing DNA preservation in degraded specimens of Amara alpina (Carabidae: Coleoptera). *Mol Ecol Resour* **14**, 606–615 (2014).
20. Goldstein, P. Z. & Desalle, R. Calibrating phylogenetic species formation in a threatened insect using DNA from historical specimens. *Mol Ecol* **12**, 1993–1998 (2003).
21. Tin, M. M.-Y., Economo, E. P. & Mikheyev, A. S. Sequencing Degraded DNA from Non-Destructively Sampled Museum Specimens for RAD-Tagging and Low-Coverage Shotgun Phylogenetics. *PLoS One* **9**, e96793 (2014).
22. Lalonde, M. M. L. & Marcus, J. M. How old can we go? Evaluating the age limit for effective DNA recovery from historical insect specimens. *Syst Entomol* **45**, 505–515 (2020).
23. Gilbert, M. T. P., Moore, W., Melchior, L. & Worobey, M. DNA Extraction from Dry Museum Beetles without Conferring External Morphological Damage. *PLoS One* **2**, e272 (2007).
24. Shpak, M., Ghanavi, H. R., Lange, J. D., Pool, J. E. & Stensmyr, M. C. Genomes from historical Drosophila melanogaster specimens illuminate adaptive and demographic changes across more than 200 years of evolution. *PLoS Biol* **21**, e3002333 (2023).
25. Webster, C. L., Longdon, B., Lewis, S. H. & Obbard, D. J. Twenty-Five New Viruses Associated with the Drosophilidae (Diptera). *Evolutionary Bioinformatics* **12s2**, EBO.S39454 (2016).
26. Webster, C. L. *et al.* The Discovery, Distribution, and Evolution of Viruses Associated with Drosophila melanogaster. *PLoS Biol* **13**, e1002210 (2015).
27. Yamaguchi, M. & Yoshida, H. Drosophila as a Model Organism. in *Advances in Experimental Medicine and Biology* vol. 1076 1–10 (Springer New York LLC, 2018).
28. Beckingham, K. M., Armstrong, J. D., Texada, M. J., Munjaal, R. & Baker, D. A. Drosophila melanogaster--the model organism of choice for the complex biology of multi-cellular organisms. *Gravit Space Biol Bull* **18**, 17–29 (2005).
29. Longdon, B., Wilfert, L. & Jiggins, F. M. *The Sigma Viruses of Drosophila from: Rhabdoviruses: Molecular Taxonomy, Evolution, Genomics, Ecology, Host-Vector Interactions, Cytopathology and Control*. (Caister Academic Press, U.K., 2012).
30. Cross, S. T. *et al.* Partitiviruses Infecting Drosophila melanogaster and Aedes aegypti Exhibit Efficient Biparental Vertical Transmission. *J Virol* **94**, (2020).

31. Vera-Maloof, F. Z., Saavedra-Rodriguez, K., Elizondo-Quiroga, A. E., Lozano-Fuentes, S. & Black Iv, W. C. Coevolution of the Ile1,016 and Cys1,534 Mutations in the Voltage Gated Sodium Channel Gene of Aedes aegypti in Mexico. *PLoS Negl Trop Dis* **9**, e0004263 (2015).
32. Ortiz-Baez, A. S., Shi, M., Hoffmann, A. A. & Holmes, E. C. RNA virome diversity and Wolbachia infection in individual Drosophila simulans flies. *J Gen Virol* **102**, 001639 (2021).
33. Habayeb, M. S., Ekengren, S. K. & Hultmark, D. Nora virus, a persistent vitus in Drosophila, defines a new picorna-like virus family. *Journal of General Virology* **87**, 3045–3051 (2006).
34. Shi, C. *et al.* Stable distinct core eukaryotic viromes in different mosquito species from Guadeloupe, using single mosquito viral metagenomics. *Microbiome* **7**, 121 (2019).
35. Santos, D., Ribeiro, G. C., Cabral, A. D. & Sperança, M. A. A non-destructive enzymatic method to extract DNA from arthropod specimens: Implications for morphological and molecular studies. *PLoS One* **13**, e0192200 (2018).
36. Sturtevant, A. H. A New Species Closely Resembling *Drosophila Melanogaster*. *Psyche (Camb Mass)* **26**, 153–155 (1919).
37. Sturtevant, A. H. Notes on North American Drosophilidae with Descriptions of Twenty-Three New Species. *Ann Entomol Soc Am* **9**, 323–343 (1916).
38. Cross, S. T. *et al.* Co-Infection Patterns in Individual Ixodes scapularis Ticks Reveal Associations between Viral, Eukaryotic and Bacterial Microorganisms. *Viruses* **10**, 388 (2018).
39. Hoskins, R. A. *et al.* The Release 6 reference sequence of the Drosophila melanogaster genome. *Genome Res* **25**, 445–58 (2015).
40. Kim, B. Y. *et al.* Single-fly genome assemblies fill major phylogenomic gaps across the Drosophilidae Tree of Life. *PLoS Biol* **22**, e3002697 (2024).
41. Balique, F., Colson, P. & Raoult, D. Tobacco mosaic virus in cigarettes and saliva of smokers. *Journal of Clinical Virology* **55**, 374–376 (2012).
42. Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* **40**, e3 (2012).
43. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* **2**, vew007 (2016).
44. Aiewsakun, P. & Katzourakis, A. Time-Dependent Rate Phenomenon in Viruses. *J Virol* **90**, 7184–7195 (2016).
45. Ghafari, M., Simmonds, P., Pybus, O. G. & Katzourakis, A. A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses. *Current Biology* **31**, 4689-4696.e5 (2021).
46. Drummond, A., Pybus, O. G. & Rambaut, A. Inference of viral evolutionary rates from molecular sequences. *Adv Parasitol* **54**, 331–58 (2003).

47. Duchêne, S., Holmes, E. C. & Ho, S. Y. W. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proceedings of the Royal Society B: Biological Sciences* **281**, 20140732 (2014).
48. Cross, S. T. *et al.* Galbut Virus Infection Minimally Influences Drosophila melanogaster Fitness Traits in a Strain and Sex-Dependent Manner. *Viruses* **15**, 539 (2023).
49. Wallace, M. A. *et al.* The discovery, distribution, and diversity of DNA viruses associated with *Drosophila melanogaster* in Europe. *Virus Evol* **7**, veab031 (2021).
50. Lowen, A. C. It's in the mix: Reassortment of segmented viral genomes. *PLoS Pathog* **14**, e1007200 (2018).
51. Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* **37**, e123 (2009).
52. Nibert, M. L. *et al.* Taxonomic reorganization of family Partitiviridae and other recent progress in partitivirus research. *Virus Res* **188**, 128–41 (2014).
53. Robicheau, B. M., Susko, E., Harrigan, A. M. & Snyder, M. Ribosomal RNA genes contribute to the formation of pseudogenes and junk DNA in the human genome. *Genome Biol Evol* **9**, 380–397 (2017).
54. Anger, A. M. *et al.* Structures of the human and Drosophila 80S ribosome. *Nature* **497**, 80–5 (2013).
55. Zok, T. *et al.* RNApdbee 2.0: multifunctional tool for RNA structure annotation. *Nucleic Acids Res* **46**, W30–W35 (2018).
56. Staple, D. W. & Butcher, S. E. Pseudoknots: RNA structures with diverse functions. *PLoS Biol* **3**, e213 (2005).
57. Bahin, M. *et al.* ALFA: annotation landscape for aligned reads. *BMC Genomics* **20**, 250 (2019).
58. Sharp, S. J., Schaack, J., Cooley, L., Burke, D. J. & Söll, D. Structure and transcription of eukaryotic tRNA genes. *CRC Crit Rev Biochem* **19**, 107–44 (1985).
59. Zhang, J. Recognition of the tRNA structure: Everything everywhere but not all at once. *Cell Chem Biol* **31**, 36–52 (2024).
60. Dabney, J., Meyer, M. & Pääbo, S. Ancient DNA damage. *Cold Spring Harb Perspect Biol* **5**, a012567 (2013).
61. Shibutani, S., Takeshita, M. & Grollman, A. P. Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. *Nature* **349**, 431–4 (1991).
62. Sloan, D. B., Broz, A. K., Sharbrough, J. & Wu, Z. Detecting Rare Mutations and DNA Damage with Sequencing-Based Methods. *Trends Biotechnol* **36**, 729–740 (2018).
63. Rhee, Y., Valentine, M. R. & Termini, J. Oxidative base damage in RNA detected by reverse transcriptase. *Nucleic Acids Res* **23**, 3275–82 (1995).
64. Sinha, R. *et al.* Index switching causes 'spreading-of-signal' among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. doi:10.1101/125724.

65. Valentine, M. R. & Termini, J. Kinetics of formation of hypoxanthine containing base pairs by HIV-RT: RNA template effects on the base substitution frequencies. *Nucleic Acids Res* **29**, 1191–9 (2001).
66. Zhang, K., Hodge, J., Chatterjee, A., Moon, T. S. & Parker, K. M. Duplex structure of double-stranded RNA provides stability against hydrolysis relative to single-stranded RNA. *Environ Sci Technol* **55**, 8045–8053 (2021).
67. Sandoval-Velasco, M. *et al.* Three-dimensional genome architecture persists in a 52,000-year-old woolly mammoth skin sample. *Cell* **187**, 3541-3562.e51 (2024).
68. Petrzik, K. Evolutionary forces at work in partitiviruses. *Virus Genes* **55**, 563–573 (2019).
69. Skoglund, P. *et al.* Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc Natl Acad Sci U S A* **111**, 2229–2234 (2014).
70. Laurence, M., Hatzis, C. & Brash, D. E. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS One* **9**, e97876 (2014).
71. Naccache, S. N. *et al.* The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *J Virol* **87**, 11966–77 (2013).
72. Kjær, K. H. *et al.* A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA. *Nature* **612**, 283–291 (2022).
73. Krogmann, L. & Holstein, J. *Chapter 18 Preserving and Specimen Handling: Insects and Other Invertebrates*. vol. 8 (ABC TAXA, 2010).
74. Foster, W. A. Colonization and Maintenance of Mosquitoes in the Laboratory. in *Pathology, Vector Studies, and Culture* 103–151 (Elsevier, 1980). doi:10.1016/B978-0-12-426102-0.50009-9.
75. Cao, C., Magwire, M. M., Bayer, F. & Jiggins, F. M. A Polymorphism in the Processing Body Component Ge-1 Controls Resistance to a Naturally Occurring Rhabdovirus in Drosophila. *PLoS Pathog* **12**, e1005387 (2016).
76. Dzaki, N., Ramli, K. N., Azlan, A., Ishak, I. H. & Azzam, G. Evaluation of reference genes at different developmental stages for quantitative real-time PCR in Aedes aegypti. *Sci Rep* **7**, 43618 (2017).
77. van der Valk, T., Vezzi, F., Ormestad, M., Dalén, L. & Guschanski, K. Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. *Mol Ecol Resour* **20**, 1171–1181 (2020).
78. Wickham, H. *et al.* Welcome to the Tidyverse. *J Open Source Softw* **4**, 1686 (2019).
79. Fox, J. & Weisberg, S. *An R Companion to Applied Regression*. (Sage, 2019).
80. Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P. & Makowski, D. performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *J Open Source Softw* **6**, 3139 (2021).

81. Laldin, A. & Hutson, G. SangerTools: Tools for Population Health Management Analytics. (2022).
82. Kassambara, A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. *https://rpkgs.datanovia.com/rstatix/* (2023).
83. Martin, M. Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads. *EMBnet J* **17**, 10–12 (2011).
84. Andrews, S. *et al.* FastQC: a quality control tool for high throughput sequence data. https://qubeshub.org/resources/fastqc (2012).
85. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
86. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. *Curr Protoc Bioinformatics* **70**, e102 (2020).
87. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**, 366–368 (2021).
88. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
89. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
90. Hebert, P. D. N., Ratnasingham, S. & deWaard, J. R. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc Biol Sci* **270 Suppl 1**, S96-9 (2003).
91. Morgulis, A. *et al.* Database indexing for production MegaBLAST searches. *Bioinformatics* **24**, 1757–64 (2008).
92. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–2 (2012).
93. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268–274 (2015).
94. Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.6.0. *https://rpkgs.datanovia.com/ggpubr/* (2023).
95. Zhang, J. *et al.* International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. *Database (Oxford)* **2011**, bar026 (2011).
96. Birolo, G. & Telatin, A. BamToCov: An efficient toolkit for sequence coverage calculations. *Bioinformatics* **38**, 2617–2618 (2022).
97. Meng, E. C. *et al.* UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci* **32**, e4792 (2023).
98. Hoskins, R. A. *et al.* The Release 6 reference sequence of the Drosophila melanogaster genome. *Genome Res* **25**, 445–458 (2015).

CHAPTER 3: EXPERIMENTAL ASSESSMENT OF 3D-PRIINTED TRAPS AND CHEMICAL ATTRACTANTS FOR THE COLLECTION OF WILD *DROSOPHILA MELANOGASTER*[2]

3.1 Summary

Drosophila melanogaster, the common fruit fly, has been instrumental to our understanding of evolution, genetics and disease. There are benefits to studying these flies in the wild, including assessment of their naturally occurring microbiota. To facilitate efforts to catch wild *D. melanogaster*, we designed two fly traps and evaluated several candidate attractants. The first trap utilized a stable food substrate that can be used to catch live flies to establish new lab colonies. The second trap was designed to be reusable and easy to ship to enable the collection of flies over time from diverse locations. We evaluated several chemical attractants derived from banana and from marula fruit, which is the proposed ancestral food host of *D. melanogaster*. We found that wild flies were preferentially attracted to banana-based odorants over marula-derived ones. Overall, these traps and attractants represent an inexpensive and simple option for the collection of wild *D. melanogaster* and related species for sampling or colony establishment.

3.2 Keywords

Drosophila melanogaster, wild *Drosophila*, microbiota, fly collection, citizen science, field application, marula fruit, crowdsourced sample collection

[2] This work has been submitted for publication: Keene-Snickers, A.H., Dunham, T.D. and Stenglein, M.D. (2025) Experimental assessment of 3D-printed traps and chemical attractants for the collection of wild *Drosophila melanogaster*. Submitted to Fly. https://doi.org/10.1101/2025.01.28.635319

3.3 Introduction

Research involving *Drosophila melanogaster* has led to major advances in our understanding of genetics, molecular and cellular biology, and disease[1]. Although laboratory fly lines are useful, neither their genetics nor the laboratory environment necessarily recapitulate the situation in the wild. Studies of *Drosophila* evolution, microbiome, and virus infection have benefited from the use of wild-caught flies[2–5]. The study of the wild fly microbiota reflects an increasing general interest in the natural microbiota of model organisms. For instance, viruses isolated from wild-caught *Caenorhabditis* nematodes have enabled studies of small interfering RNA responses to natural virus infection[6]. Wilding mice, created by implanting laboratory mouse embryos into wild mothers, exhibit more natural immune response[7,8]. Similarly, wild *D. melanogaster* harbor diverse microbes not necessarily present in laboratory strains. These interact with the host and other microbiota, altering host biology and evolution[2,9–12]. The capture and maintenance of wild derived flies in the lab has been instrumental to understanding the impact of common *D. melanogaster* infecting viruses[2,3,13–16].

To capture wild *D. melanogaster*, researchers typically use manual or electronic aspirators, net swipes, or baited traps consisting of fruit (commonly banana) and yeast[17–20]. Although these traps are effective at capturing wild flies, they involve specialized equipment or could be technically challenging for non-experts[17,20]. Other trap types aim to attract and kill pest *Drosophila* species. *Drosophila suzukii*, a major agricultural pest of soft fruits can be efficiently captured and killed using vinegar-based traps[21–23]. The attractiveness of a trap is important to how successful it will be, particularly in natural environments where there is competition from other food sources.

In this study, we experimentally assessed the stability and attractiveness of different bait substrates, trap designs, and banana- and marula-derived attractants. The marula fruit is proposed to have been an important host fruit for ancestral *D. melanogaster* [24]. Populations of *D. melanogaster* in Sub-Saharan Africa have polymorphisms in the Or22a-expressing olfactory sensory neurons that increase their sensitivity to ethyl isovalerate, a compound in marula fruit[24]. Other compounds in marula fruit include isoamyl acetate, which is also a main component of synthetic banana flavoring.

Our goal was to design traps that gave us the ability to collect and retain live flies if we wanted and to be easily used by participatory scientists for crowdsourced fly collection. To this end we designed two traps. The first trap focused on the capture of live flies for the generation of new lab colonies. This trap relied on an environmentally stable cornmeal-based food and the addition of isoamyl acetate to increase attractiveness. Our second trap was designed to be easily used by citizen scientists. This trap used banana and yeast, a maximally attractive and readily available bait. These traps could also be reused making them useful for longitudinal collections.

3.4 Results

3.4.1 Cornmeal-based fly food was most stable

We first set out to assess the stability of traps baited with different food substrates over two weeks. We tested cornmeal-, potato-, and banana-based foods, all of which are commonly used to rear *D. melanogaster*. Each type was made plain or with synthetic banana flavoring (isoamyl acetate). Food bottles were placed in a controlled

indoor environment or outside for two weeks. Bottles were visually assessed daily to identify signs of decay and microbial growth.

Banana-based food was the least stable with visible mold and bacterial growth occurring in a matter of days (**Supplemental Fig. 1**). Potato-based food was more stable: after two weeks there was visible mold and bacterial growth but less than the banana-based food (**Supplemental Fig. 1**). Cornmeal-based food was the most stable in indoor and outdoor environments (**Supplemental Fig 1**). The indoor cornmeal-based food looked nearly the same at the end of the two weeks as it did at the start. The presence of synthetic banana flavoring did not impact the stability of any of the foods. For all food types there was considerable evaporation in the bottles that caused moisture to accumulate, particularly outdoors (**Supplemental Fig. 1**).

3.4.2 Banana-based compounds increase the attractiveness of plain food to *D. melanogaster*

We next sought to augment the attractiveness of cornmeal food baited traps. Although flies can be maintained on such food in the lab, it was unclear whether these foods would be sufficiently attractive. We expected that there would be competition from other food sources, particularly in outdoor environments, where organic volatiles known to attract *D. melanogaster* are emitted from rotting plant debris[30,31]. *D. melanogaster* are attracted to fermenting banana catalyzed by the addition of yeast[32]. However, this banana and yeast mixture decays over the course of several days creating a sticky substance that makes it difficult to retrieve flies. We therefore employed a choice assay with our wild derived laboratory line of *D. melanogaster* (FoCo-17) where the flies were

exposed to plain cornmeal-based food with or without homemade banana extract, or to banana/yeast bait[26].

We released flies into mesh enclosures in controlled indoor or outdoor environments containing traps baited with different food types. We left flies for 24-48 hours and recorded the number of flies in each bottle. Flies preferred food with a banana component (**Fig. 1**). Banana and yeast food was more attractive than either plain cornmeal ($p = 1.7x10^{-10}$) or cornmeal food supplemented with homemade banana extract ($p = 5.4x10^{-3}$). This preference was less pronounced in outdoor conditions ($p = 8.8x10^{-3}$ & $p = 0.83$, **Fig. 1**). This shows that banana and yeast was an effective fly attractant, as expected. However, a banana-based additive also increased attractiveness.

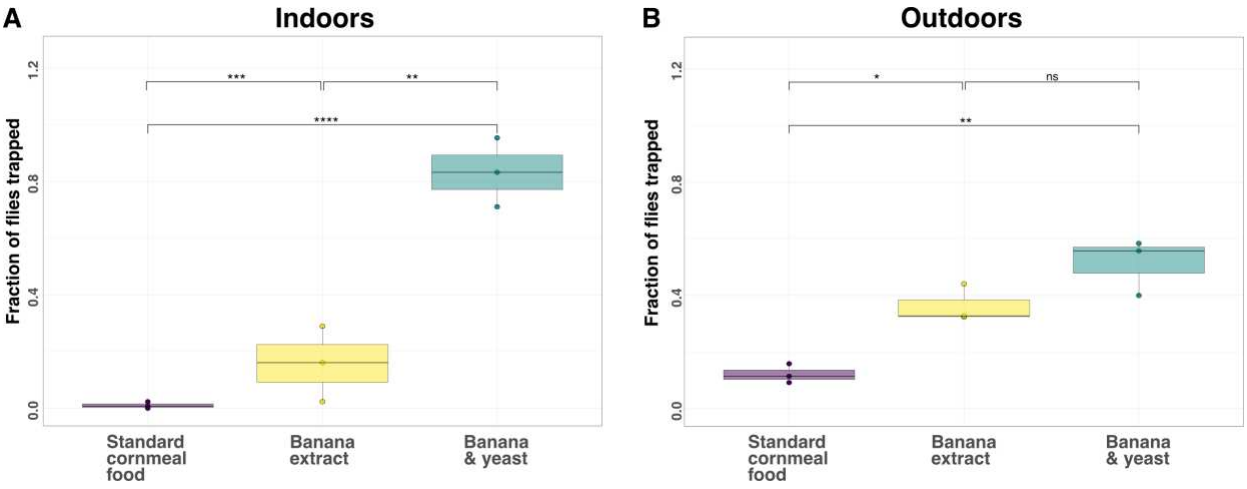


Figure 1 *D. melanogaster* were more attracted to banana-based baits. Fly bottles were either filled with plain cornmeal-based food, cornmeal-based food with homemade banana extract or banana and yeast. A 3D printed fly trap lid was used to seal the bottles. FoCo-17 flies were released in controlled environments either inside (n = 143, n = 197, n = 173) or outside (n=173, n=61, n=220) and left for 24-48 hours. Bottles were collected and the proportion of flies in each bottle was tabulated. Significance values from a negative binomial model are: $p < 0.0001$ '****', $0.0001 < p > 0.001$ '***', $0.001 < p > 0.01$ '**', $0.01 < p > 0.05$ '*' and $p > 0.05$ 'ns'.

3.4.3 Banana extract was more attractive to *D. melanogaster* than marula tree derivatives

The fruit of the marula tree has been proposed to be the ancestral host of *D. melanogaster*[24]. We therefore tested whether ethyl isovalerate, a volatile ester emitted from marula fruit, would be better at attracting flies than banana-based compounds. Southern African *D. melanogaster* populations have polymorphisms in the Or22a odorant receptor that increase sensitivity to this compound, but this has not yet been tested using North American populations[24]. Since flies were highly attracted to banana, we also tested the use of synthetic banana flavoring (isoamyl acetate) as an inexpensive and readily available chemical attractant. Isoamyl acetate is also a prominent ester in marula fruit[24,33].

We also wanted to test whether the effectiveness of these additives was impacted by the food base. Fly bottles with cornmeal or potato food and the various attractants were placed in indoor or outdoor enclosures. Flies from the wild-derived FoCo-17 population were released for 24-48 hours and the number of flies in each trap was recorded.

Across all replicates in both environments regardless of food type, banana-based odorants were the most attractive (**Fig. 2**). Surprisingly, there were very few flies in the ethyl isovalerate or marula fruit oil bottles (**Fig. 2**). There was a decrease in the relative attractiveness of banana extracts outdoors (**Fig. 2B & D**). Standard cornmeal-based food with either homemade or store-bought banana extract produced statistically indistinguishable results (**Fig. 2**). Due to the ease of use and chemical consistency, we opted to use synthetic banana flavoring (isoamyl acetate) for subsequent experiments.

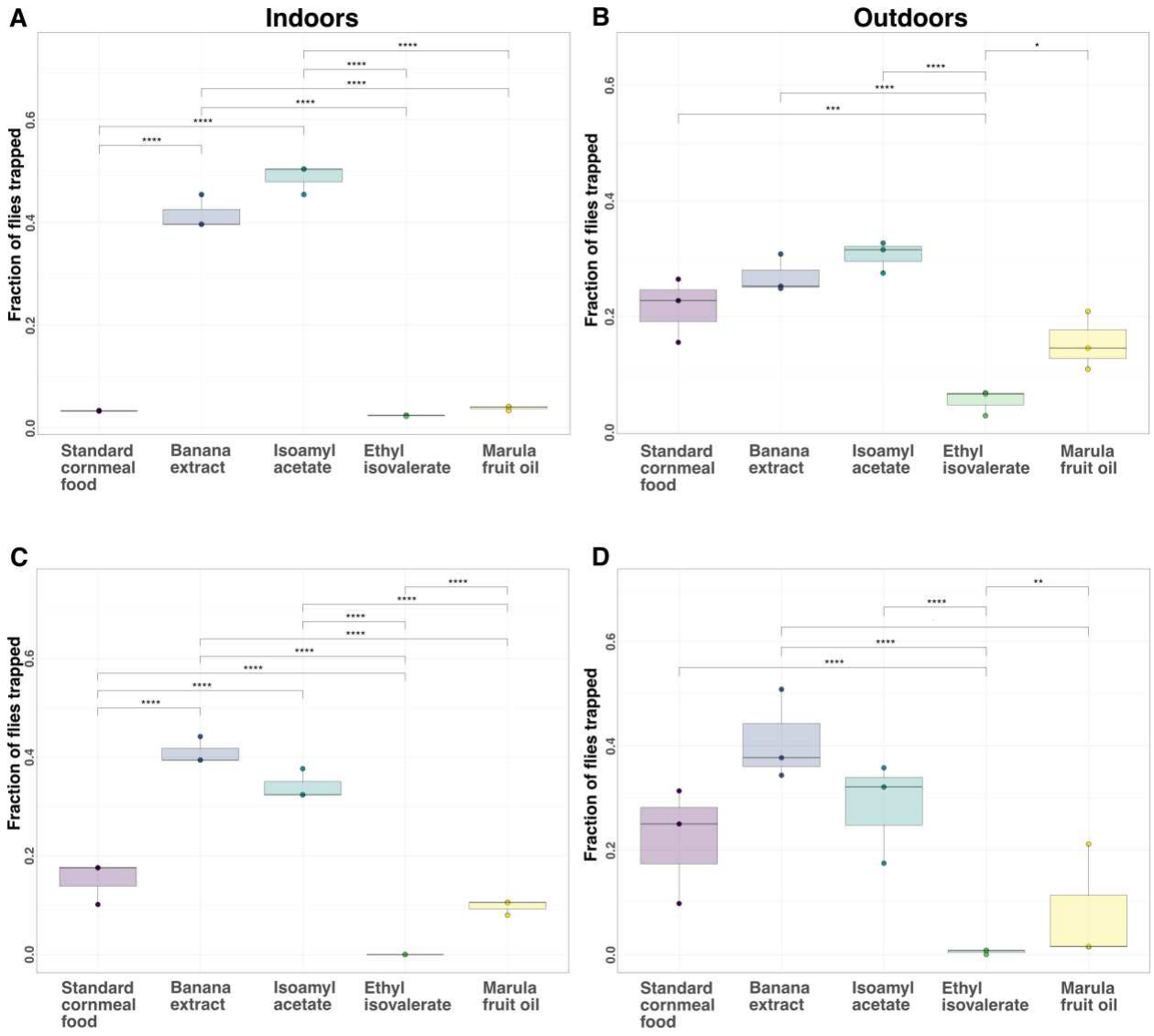


Figure 2: *D. melanogaster* prefer food with a banana-based attractant over food with derivatives of the marula fruit. The relative attractiveness of cornmeal-based (A & B) or potato-based (C & D) fly food with various baits were assessed in competition experiments. Experiments were conducted indoors (A,C) and outdoors (B,D). FoCo-17 flies were released for 24-48 hours in enclosures containing baited fly traps. The number of flies in each replicate was: (A) indoor n= 121, 121, 88; (B) outdoor n = 211, 206, 378 and (C) indoor n = 142, 142, 138; (D) outdoor n = 134, 246, 260. Significance values determined from a negative binomial model: $p < 0.0001$ '****', $0.0001 < p > 0.001$ '***', $0.001 < p > 0.01$ '**', $0.01 < p > 0.05$ '*' and $p > 0.05$ 'ns'.

The lack of attractiveness of ethyl isovalerate led us to test whether we were using an appropriate concentration. We performed an experiment using five ethyl

isovalerate concentrations: 0.45%, 0.9%, 1.8% (~the natural concentration in marula fruit[33]), 3.6% and 7.2%. Plain cornmeal bottles and bottles with cornmeal plus isoamyl acetate at 1.5%, the natural concentration found in marula fruit[33], served as controls. As with the previous experiments, FoCo-17 flies were released in our indoor mesocosm for 24-48 hours and then counted to determine the proportion of flies in each bottle.

Flies were most attracted to plain food or food containing synthetic banana extract (**Fig. 3**). The ethyl isovalerate bottles in aggregate attracted on average 31.8% of the flies with the 1.8% ethyl isovalerate bottle attracting 10.5% of the flies. The 3.6% and 7.2% bottles attracted the fewest flies. In contrast, synthetic banana flavoring attracted on average 37.4% of the flies and plain food 30.6% of the flies (**Fig. 3**). Food with banana extract was not significantly more attractive than plain food. It is possible that the plain and banana extract food bottles were not placed far enough apart due to the limited space in the mesocosm. Overall, the results of these experiments confirmed our previous findings that *D. melanogaster* were more attracted to isoamyl acetate than ethyl isovalerate.

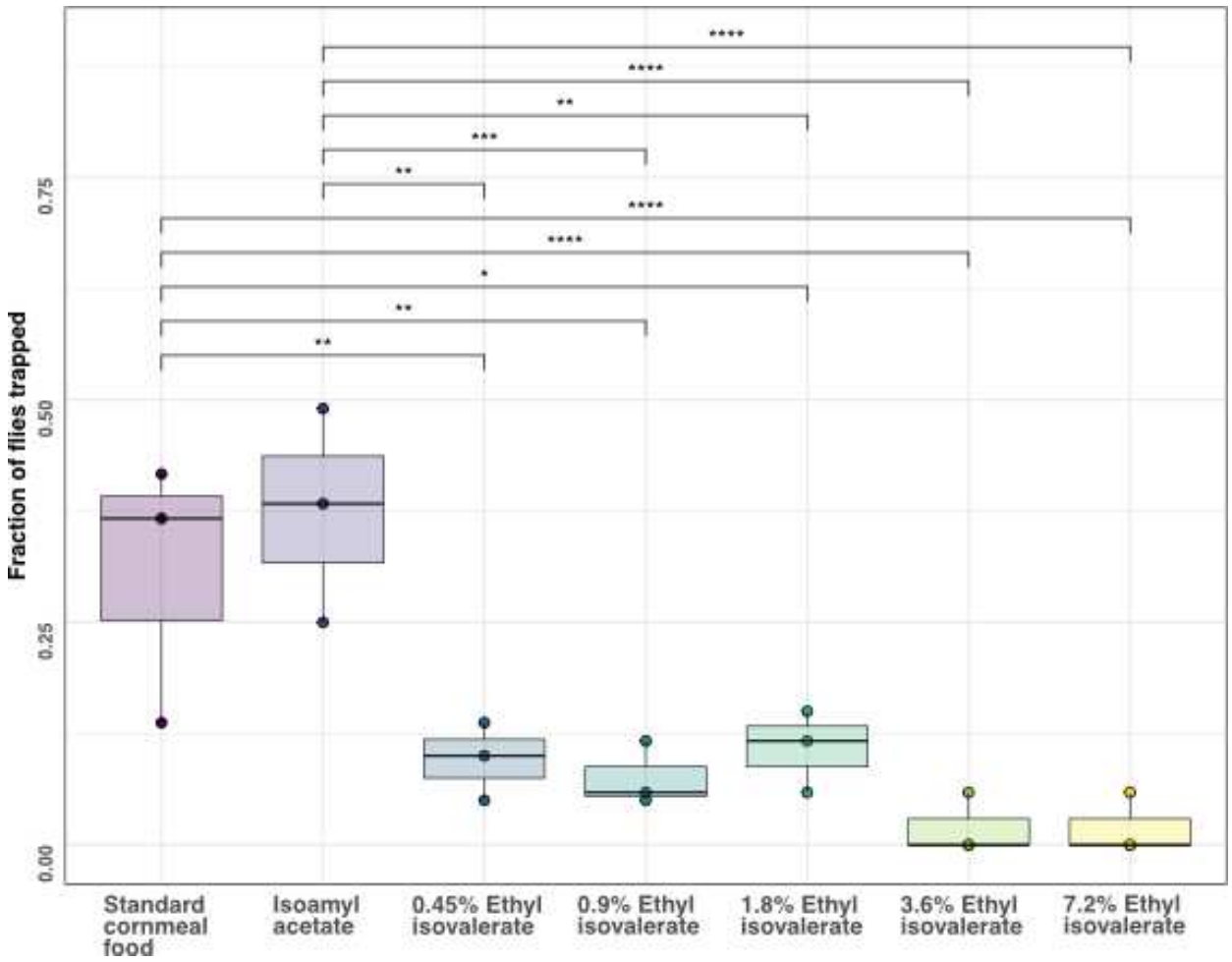


Figure 3: An outbred population of *D. melanogaster* from Colorado was not attracted to ethyl isovalerate. Flies from the FoCo-17 wild-derived outbred population of *D. melanogaster* were released (n=51, 60, 60) in an indoor microcosm cage for 24-48 hours. Cornmeal-based food with the following additives were placed in each cage: plain food, 1.5% synthetic banana extract and food supplemented with 0.45%, 0.9%, 1.8%, 3.6% or 7.2% ethyl isovalerate. After 24-48 hours the number of flies were counted. Significance values determined from pairwise comparisons of a negative binomial model are: $p < 0.0001$ '****', $0.0001 < p > 0.001$ '***', $0.001 < p > 0.01$ '**', $0.01 < p > 0.05$ '*' and $p > 0.05$ 'ns'.

3.4.4 A kit for capture of wild *D. melanogaster*

As a pilot test of the ability of these traps to be used by non-specialists, we sent traps with cornmeal-based food with isoamyl acetate to three volunteers across the

United States (**Fig. 4A**). Volunteers were instructed to place the traps in an outdoor or indoor location with fruit fly activity for roughly a week. They were told to watch for trapped flies and signs of larval development. Once first instar larva could be seen, the volunteers were instructed to seal the bottle with parafilm and ship it back to our lab. We received bottles from two locations and were able to generate colonies that persist three years later.

We next tested a modified trap that simplified fly capture and trap reuse. This trap included a 3D printed “table” covered in fine mesh to separate flies from the bait (**Fig. 4B**). This trap design takes advantage of the attractiveness of banana/yeast bait while preventing flies from getting entrapped.

These traps proved to be an effective means to capture flies. Volunteers were instructed to add one to two slices of banana and a small pinch of yeast to the bottom of the jar, place the “table” over the banana and close the jar with the trap lid. They then placed the trap in an area with fruit fly activity and monitored the trap for flies and rotting of the bait. Volunteers periodically froze the trap to kill the flies and moved them to the 2.0 mL tube. The tube was stored in the freezer and another round of collection began with fresh bait. At the end of the collection period (1-3 months) they shipped the 2.0 mL tubes back to our lab. Once at our lab, fly species was determined using morphological characteristics and stored in a -80°C freezer until further analysis.

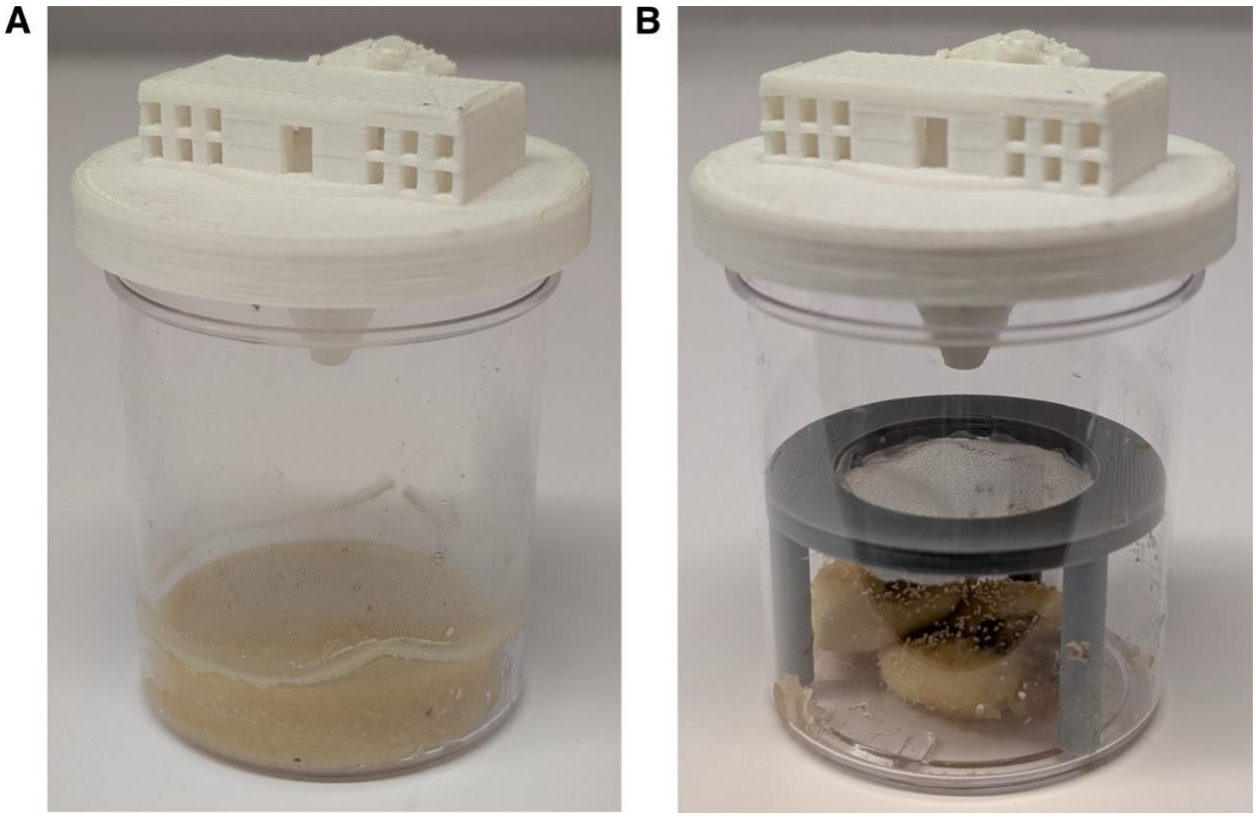


Figure 4: A 3D printed fly trap kit to collect wild *D. melanogaster*. Fly trap designs for the collection of live flies (A) or for repeated fly collection over time (B). Trap components are 3D printed. Cornmeal-based food with isoamyl acetate can be used as a stable food that facilitates egg laying (A) or banana and yeast can be used as bait without the risk of flies becoming trapped in the decomposing fruit (B).

3.5 Discussion

Here we describe the design and effectiveness of two wild *D. melanogaster* fly traps. The first trap is characterized by a stable cornmeal and agar-based food supplemented with propionic acid to inhibit bacterial growth. Eggs laid by adult flies remained viable in this food even after shipment. This made it possible to generate new laboratory colonies from various locations after collection by non-specialists. We have maintained two stocks generated from flies collected in this manner in 2021.

The second trap was designed to facilitate sample collection and trap reuse. This makes it ideal for crowdsourced sample collection and longitudinal collection. One feature of this trap is that it can be combined with other baits for the collection of specific *Drosophila* species[20]. In a small field trial, kits were sent to nine volunteers with no previous *Drosophila* collecting experience. Volunteers were able to setup the kits and three locations successfully collected flies that were received by our laboratory.

We tested the attractiveness of marula fruit oil and ethyl isovalerate as a *D. melanogaster* attractants because it had been proposed that ancestral populations of these flies evolved to specialize on marula fruit[24]. However, we found that banana-based additives were more attractive, at least to our wild-derived flies from Colorado (**Fig. 2 & Fig. 3**). Although *D. melanogaster* are considered to be generalists, many species within the *melanogaster* subgroup specialize on a single fruit with populations of the same species showing different specializations[24,34,35]. It is possible that the population of *D. melanogaster* we tested have either lost sensitivity to ethyl isovalerate or gained sensitivity to volatiles released by different rotting fruits[24,31]. This could explain the observed unattractiveness of this compound. There is considerable variation in odorant receptors that detect esters emitted by yeast and fermenting fruit [36].

The use of community engagement to further scientific research has become increasingly popular[37–39]. The traps are designed for use by untrained volunteers and to be baited with readily available materials. *D. melanogaster* are attracted to a variety of other fruits if banana is unavailable[40]. Overall, our flexible designs enable researchers and non-specialists to obtain flies from diverse locations for minimal cost and effort.

3.6 Methods & Materials

3.6.1 Fly Food

Cornmeal-based food: We used the standard cornmeal food based recipe from the Bloomington Drosophila Stock Center (https://bdsc.indiana.edu/information/recipes/bloomfood.html). Twenty-five to 50 mL of food was placed in fly bottes (VWR, 75813-140). For experiments involving ethyl isovalerate, ethyl isovalerate (Sigma-Aldrich 112283) was added to final percent volume concentrations of 0.45%, 0.9%, 1.8%, 3.6% or 7.2%. For experiments involving isoamyl acetate, synthetic banana flavoring (Kroger Brand Imitation Banana, 87-9110) was added to a final concentration of 1.5%.

Potato-based food: was prepared as described by Waddle, 1996 (www.anapsid.org/fruitfly.html), except baker's yeast was substituted for brewer's yeast[25]. Briefly, 120g of instant potato was mixed with 12g yeast. Mold inhibitor was made by adding 1g of Tegosept to 1 L of boiling water. Mold inhibitor was added to the potato/yeast mixture to rehydrate the potatoes. The volume of mold inhibitor varied from batch to batch based on how much liquid was needed to fully rehydrate the potato/yeast mixture. Twenty-five to 50 mL of food was then placed in fly bottles and sprinkled with baker's yeast. For attractiveness experiments, 4g of agar dissolved in 200mL water was added to increase shelf life.

Banana-based food: was prepared as described by Hundt, 1996 (www.anapsid.org/fruitfly.html)[25]. Briefly, 4 bananas were mashed together with 1/8 cup cane sugar until liquified. Rolled oats were added until the mixture became firm. Twenty-five to 50 mL of food was transferred to fly bottles and sprinkled with baker's yeast.

3.6.2 Fly food attractants

All potential attractants except the homemade banana extract were obtained through commercial vendors: isoamyl acetate (Kroger Brand Imitation Banana, 87-9110), ethyl isovalerate (Sigma-Aldrich, 112283) and marula oil (Rocky Mountain Oil Company, All Organic marula Oil). Homemade banana extract was made by manually mashing bananas until mostly liquid. The pulp and liquid was then separated using a mesh strainer. The liquid was retained and stored in a -20°C freezer until use.

3.6.3 Outdoor experiments

For attractiveness experiments, wild-derived flies from our lab's FoCo17 colony were used[26]. Microcosms were constructed using bug dorms (dimensions: W32.5 x D32.5 x H77.0 cm; similar to: bugdorm, BD4F3074). Microcosms were placed in an outdoor environment that experienced mixed sun and shade with temperatures ranging from 13°C to 38°C. Microcosms were protected from rain by a tarp. Flies released into microcosms were given 24-48 hours to select one of the foods, either plain or supplemented with one of the marula fruit or banana additives. To trap the flies, we used food bottles capped with 3D printed trap lids. These lids allowed flies to easily enter the bottle but not as easily get out. Bottles were collected and frozen before the number of flies in each bottle were counted.

3.6.4 Indoor experiments

Population cages in our insectary were used for these experiments. These cages were approximately D1.8 x W1.2 x H2.4 m in size. Procedures were the same as for the outdoor experiments: food was placed in bottles with the trap lid, flies were released for

24-48 hours before food bottles were collected, frozen and the number of flies in each bottle was determined. Indoor microcosms were on a 12-hour light-dark cycle and temperature was kept at ~21°C.

3.6.5 Live collection trap

3D-printed trap components were designed in the OpenSCAD scripting design language and printed in polylactic acid on a LulzBot Taz6 printer (Fargo Additive Manufacturing Equipment). Designs are available at https://github.com/stenglein-lab/3D_parts/tree/master/fly_trap. For our initial trap testing, we used threadless fly bottles (VWR, 75813-140) and attached the trap lid to the bottle using parafilm. Subsequent experiments involving volunteer collectors used 4 oz jars (Uline, S-9934) and the 3D printed trap containing cornmeal-based food as described above. Before pouring the food into bottles, we added isoamyl acetate at a 1:100 ratio to food. Fifty mL of food was poured into bottles and left to cool overnight. Bottles were stored at 4°C until use. Each bottle was sealed with a cotton plug and parafilm. Traps were shipped to volunteers along with the 3D printed trap lid and parafilm. Volunteers were instructed to place the trap lid on the bottle and place the trap in a location with fruit fly activity (one outdoor location and one indoor location). After ~1 week volunteers were instructed to seal the bottle with the provided parafilm and to ship it back to our lab (**Fig. 4A**).

3.6.6 Reusable trap

We created kits containing materials for trapping and returning flies. Kits included a clear plastic jar (Uline, S-9934), a 3D printed trap lid, a 3D printed barrier with organdy mesh (Oriole Textile, 2060), Fleischmann's Instant Dry Yeast, a screw top 2.0 mL tube

and a pre-paid shipping return envelope. Volunteers were instructed to place 1-2 slices of banana in the bottom of the jar with a small pinch of yeast. The barrier was placed over the banana/yeast ensuring that flies remained separated from bait. Volunteers were instructed to replace the banana/yeast every 3-5 days depending on the environmental conditions (hot/humid environments were told to replace the banana/yeast more often) (**Fig. 4B**). Flies were collected from traps that had been placed in a -20˚C freezer for ≥ 30 minutes. Once dead, flies could be removed by tapping the jar upside down or with forceps and stored frozen at -20˚C in 2.0 mL tubes. Tubes were returned in the provided prepaid shipping envelope.

3.6.7 Data analyses

Data visualization and statistical analyses were performed in RStudio (v 2024.09.0+375) using the tidyverse package[27]. Analysis code is available at https://github.com/LKeene/fly_trap_design . A p-value of $\alpha = 0.05$ or lower was used as our significance threshold. For attractiveness experiment, negative binomial models were generated using the MASS package and comparisons were made using the emmeans package[28,29]. Pairwise comparisons were adjusted for multiple hypothesis testing using Tukey's method. For Figure 3C, a pseudocount of one was added to all values to accommodate zero counts.

3.7 Acknowledgements

We thank volunteers and their willingness to participate in trap testing: Shaun Cross, Brendan, Arabella, and Malila Kavaney, Amanda, Mike, Clara, and Zoe Furlotti,

Lisa Harrington, Paul Keene, Khwannarin Khemsom, and Anastazia Jablunovsky. We would also like to thank Keller Gonzales for technical assistance.

3.8 Funding

NSF IOS 2048214 (AHK, TJD and MDS) and NIH T32GM132057 (AHK).

3.9 Declaration of interest:

The authors declare that they have no known financial interest or personal relationship that could have impacted the work described in this manuscript.

3.10 Data availability statement

The raw data that support the findings of this study and the scripts that were used to generate the figures are available on GitHub at https://github.com/LKeene/fly_trap_design.

CHAPTER 3: REFERENCES

1. Yamaguchi, M. & Yoshida, H. Drosophila as a Model Organism. in *Advances in Experimental Medicine and Biology* vol. 1076 1–10 (Springer New York LLC, 2018).
2. Martinez, J. *et al.* Virus evolution in wolbachia-infected drosophila. *Proceedings of the Royal Society B: Biological Sciences* **286**, (2019).
3. Ortiz-Baez, A. S., Shi, M., Hoffmann, A. A. & Holmes, E. C. RNA virome diversity and Wolbachia infection in individual Drosophila simulans flies. *J Gen Virol* **102**, 001639 (2021).
4. Webster, C. L. *et al.* The Discovery, Distribution, and Evolution of Viruses Associated with Drosophila melanogaster. *PLoS Biol* **13**, e1002210 (2015).
5. Webster, C. L., Longdon, B., Lewis, S. H. & Obbard, D. J. Twenty-Five New Viruses Associated with the Drosophilidae (Diptera). *Evolutionary Bioinformatics* **12s2**, EBO.S39454 (2016).
6. Félix, M.-A. & Wang, D. Natural Viruses of *Caenorhabditis* Nematodes. *Annu Rev Genet* **53**, 313–326 (2019).
7. Ma, J. *et al.* Laboratory mice with a wild microbiota generate strong allergic immune responses. *Sci Immunol* **8**, (2023).
8. Rosshart, S. P. *et al.* Laboratory mice born to wild mice have natural microbiota and model human immune responses. *Science (1979)* **365**, (2019).
9. Shi, M. *et al.* No detectable effect of Wolbachia wMel on the prevalence and abundance of the RNA virome of Drosophila melanogaster. *Proceedings of the Royal Society B: Biological Sciences* **285**, (2018).
10. Palmer, W. H., Varghese, F. S. & Van Rij, R. P. Natural variation in resistance to virus infection in dipteran insects. *Viruses* vol. 10 Preprint at https://doi.org/10.3390/v10030118 (2018).
11. Van Rij, R. P. *et al.* The RNA silencing endonuclease Argonaute 2 mediates specific antiviral immunity in Drosophila melanogaster. *Genes Dev* **20**, 2985–2995 (2006).
12. Nayak, A. *et al.* Cricket paralysis virus antagonizes Argonaute 2 to modulate antiviral defense in Drosophila. *Nat Struct Mol Biol* **17**, 547–554 (2010).
13. Cogni, R., Ding, S. D., Pimentel, A. C., Day, J. P. & Jiggins, F. M. Wolbachia reduces virus infection in a natural population of Drosophila. *Commun Biol* **4**, 1327 (2021).
14. Wallace, M. A. & Obbard, D. J. Naturally occurring viruses of *Drosophila* reduce offspring number and lifespan. *Proceedings of the Royal Society B: Biological Sciences* **291**, (2024).
15. Shi, M. *et al.* No detectable effect of Wolbachia wMel on the prevalence and abundance of the RNA virome of Drosophila melanogaster. *Proceedings of the Royal Society B: Biological Sciences* **285**, (2018).

16. Mackay, T. F. C. *et al.* The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**, 173–8 (2012).
17. Markow, T. A. & O'Grady, P. M. Collecting Drosophila in the wild. in *Drosophila* 145–153 (Elsevier, 2006). doi:10.1016/B978-012473052-6/50004-4.
18. Topal, P., Garg, D. & S. Fartyal, R. Fruit Flies (*Drosophila spp.*) Collection, Handling, and Maintenance: Field to Laboratory. in *The Wonders of Diptera - Characteristics, Diversity, and Significance for the World's Ecosystems* (IntechOpen, 2021). doi:10.5772/intechopen.97014.
19. Faria, V. G. & Sucena, É. From Nature to the Lab: Establishing Drosophila Resources for Evolutionary Genetics. *Front Ecol Evol* **5**, (2017).
20. Carson, H. L. & Heed, W. B. Methods of Collecting Drosophila. in *The Genetics and Biology of Drosophila* (eds. Ashburner, M., Thompson, J. N. & Carson H.L.) vol. 3d 1–28 (Academic Press, London, 1983).
21. Landolt, P. J., Adams, T. & Rogg, H. Trapping spotted wing drosophila, Drosophila suzukii (Matsumura) (Diptera: Drosophilidae), with combinations of vinegar and wine, and acetic acid and ethanol. *Journal of Applied Entomology* **136**, 148–154 (2012).
22. Iglesias, L. E., Nyoike, T. W. & Liburd, O. E. Effect of trap design, bait type, and age on captures of Drosophila suzukii (Diptera: Drosophilidae) in berry crops. *J Econ Entomol* **107**, 1508–1512 (2014).
23. Dadour, I. R. & Cook, D. F. THE EFFECTIVENESS OF FOUR COMMERCIAL FLY TRAPS AT CATCHING INSECTS. *Aust J Entomol* **31**, 205–208 (1992).
24. Mansourian, S. *et al.* Wild African Drosophila melanogaster Are Seasonal Specialists on Marula Fruit. *Current Biology* **28**, 3960-3968.e3 (2018).
25. Kaplan, M., Hundt, A., Waddle, F. & Nehring, N. Homemade fruit Fly Culture media. *Melissa Kaplan's Herp Care Collection* (1996).
26. Cross, S. T. *et al.* Partitiviruses Infecting Drosophila melanogaster and Aedes aegypti Exhibit Efficient Biparental Vertical Transmission. *J Virol* **94**, (2020).
27. Wickham, H. *et al.* Welcome to the Tidyverse. *J Open Source Softw* **4**, 1686 (2019).
28. Venables, W. & Ripley, B. *Modern Applied Statistics with S*. (Springer, New York, 2002).
29. Lenth, R. V. *et al.* emmeans: Estimated Marginal Means, aka Least-Squares Means. *CRAN: Contributed Packages* Preprint at https://doi.org/10.32614/CRAN.package.emmeans (2017).
30. Ruebenbauer, A., Schlyter, F., Hansson, B. S., Löfstedt, C. & Larsson, M. C. Genetic Variability and Robustness of Host Odor Preference in Drosophila melanogaster. *Current Biology* **18**, 1438–1443 (2008).
31. Soto-Yéber, L., Soto-Ortiz, J., Godoy, P. & Godoy-Herrera, R. The behavior of adult Drosophila in the wild. *PLoS One* **13**, e0209917 (2018).
32. Markow, T. A. & O'Grady, P. M. Chapter 4 - Collecting Drosophila in the wild. in *Drosophila* (eds. Markow, T. A. & O'Grady, P. M.) 145–153 (Academic Press, San Diego, 2006). doi:https://doi.org/10.1016/B978-012473052-6/50004-4.

33. Viljoen, A. M., Kamatou, G. P. P. & Başer, K. H. C. Head-space volatiles of marula (Sclerocarya birrea subsp. caffra). *South African Journal of Botany* **74**, 325–326 (2008).
34. Auer, T. O. *et al.* Olfactory receptor and circuit evolution promote host specialization. *Nature* **579**, 402–408 (2020).
35. Comeault, A. A. *et al.* A nonrandom subset of olfactory genes is associated with host preference in the fruit fly Drosophila orena. *Evolution Letters* vol. 1 73–85 Preprint at https://doi.org/10.1002/evl3.7 (2017).
36. de Bruyne, M., Smart, R., Zammit, E. & Warr, C. G. Functional and molecular evolution of olfactory neurons and receptors for aliphatic esters across the Drosophila genus. *J Comp Physiol A Neuroethol Sens Neural Behav Physiol* **196**, 97–109 (2010).
37. Lukyanenko, R., Wiggins, A. & Rosser, H. K. Citizen Science: An Information Quality Research Frontier. *Information Systems Frontiers* **22**, 961–983 (2020).
38. Fraisl, D. *et al.* Citizen science in environmental and ecological sciences. *Nature Reviews Methods Primers* **2**, 64 (2022).
39. Encarnação, J., Teodósio, M. A. & Morais, P. Citizen Science and Biological Invasions: A Review. *Front Environ Sci* **8**, (2021).
40. Dweck, H. K. M. *et al.* The Olfactory Logic behind Fruit Odor Preferences in Larval and Adult Drosophila. *Cell Rep* **23**, 2524–2531 (2018).

CHAPTER 4: DIVERSE GEOGRAPHIC COLLECTION OF WILD *DROSOPHILA MELANOGASTER* CHARACTERIZES THE GENOTYPE-PHENOTYPE RELATIONSHIP OF A PERSISTENT VIRAL INFECTION

4.1 Summary

Galbut virus is a remarkably successful virus of the model organism *Drosophila melanogaster*. To better understand the diversity and evolution of this virus, we screened 974 flies collected from 13 locations across the USA. Seventy-five percent of flies tested positive for galbut virus. We selected 155 flies for RNA sequencing to recover galbut virus sequences. This yielded 373 coding-complete or near complete galbut virus segment sequences. Galbut virus sequences clustered into three major clades. Two clades, termed clade A and B, contained *D. melanogaster*-infecting viruses and the third corresponded to *D. simulans*-infecting viruses. Galbut virus diversity within clades was low but higher between clades, with pairwise identities between clades approaching 75%. A third of samples exhibited evidence of coinfection which did not require complete sets of all three galbut virus segments. *D. melanogaster* clade A viruses showed a bimodal distribution of infection levels while clade B viruses showed similar infection levels in all samples. There was evidence of galbut virus reassortment within clades but not between clades, indicating that viruses in these clades are reproductively isolated. Chaq virus, a likely satellite virus of galbut virus, appears to be associated only with clade A infections. Together this data explores the sequence diversity of galbut virus and describes genotype-phenotype associations of these infections.

4.2 Introduction

Galbut virus is a tri-segmented double-stranded (ds) RNA virus from the partitivirus family that infects *Drosophila melanogaster*[1,2]. Chaq virus is likely a satellite of galbut virus or may represent an optional 4th galbut virus segment[1,2]. Galbut virus appears to have limited impact on its host's fitness, a common trait among partitiviruses[3]. One study found that galbut virus infection resulted in reduced offspring numbers[4]. While another found that in at least one inbred fly line from the *Drosophila* Genetic Reference Panel (DGRP), galbut virus infection did not reduce offspring numbers and even sped up time to pupation and eclosion[5,6]. Another inbred fly line infected with galbut virus exhibited decreased survival following infection by some bacterial and fungal pathogens[5]. However, how these potentially offsetting fitness traits described in the lab contribute to overall fitness of flies in the wild is yet to be elucidated.

Galbut virus is exceptionally common and globally distributed: every wild *D. melanogaster* population examined thus far has infected flies and ~60% of individual flies are infected. Galbut virus is vertically transmitted with high efficiency, and its global spread likely coincided with *D. melanogaster's spread* across the world as a human commensal[1,7].

Galbut virus has been found to infect flies across species barriers. Wild *Drosophila simulans* flies from Australia and *Drosophila suzukii* flies from the UK and Japan have also been shown to be infected by galbut virus[8,9]. The presence of an endogenous viral element (EVE) in some European populations of flies[10] and that it infects multiple hosts suggests that galbut virus has been infecting flies for a long time. The recent sequencing of *D. melanogaster* from museum collections suggests that

galbut virus evolves slower than expected[11], a phenomenon seen in other partitiviruses[12,13].

Despite galbut viruses' presence in several *Drosophila* species, questions about general galbut virus biology and its impact on host biology in the wild remain unanswered. Three clades of galbut virus have been identified, two that infect *D. melanogaster* and one that infects *D. simulans*[2,11]. The biological characteristics of these clades such as their abundance in natural populations, whether different genotypes circulate within the same populations and if there are defining traits associated with each clade are unknown.

To study galbut virus diversity and evolution, we collected *D. melanogaster* from Colorado, Maine, Pennsylvania and Ohio. We screened 974 flies for galbut virus and selected 155 for sequencing. This resulted in the recovery of 373 coding-complete or near-complete galbut virus and chaq virus sequences. We found that overall galbut virus sequence diversity within genotypes was low but relatively high between genotypes. We investigated genotype-phenotype associations and found that viral RNA levels and chaq association differed between genotypes. We also found that coinfection and reassortment are common features of galbut virus infection. Collectively, these data describe the sequence diversity and genotype-phenotype associations of galbut virus which provides insight into the biology and evolution of common persistent viral infections of arthropods.

4.3 Results

4.3.1 Galbut virus prevalence varies by location and time

To better understand galbut virus prevalence and diversity, we sampled flies from across the United States. We collected flies from nine locations in Larimer County, Colorado, and from one location each in Maine, Pennsylvania, and Ohio (**Fig. 1A & 1B**). We used buckets baited with banana and yeast to collect flies in Colorado and volunteers used 3D printed fly traps to collect flies from Maine, Ohio and Pennsylvania[14]. Overall, we collected and tested 974 individual flies from all locations during the summer of 2023.

RNA from individual flies was screened for galbut virus via RT-qPCR (**Fig. 1C**). Ribosomal protein L32 (RpL32) mRNA levels were used to normalize galbut virus RNA levels and served as a positive control. Of the 974 flies captured, 952 were positive for RpL32 mRNA, indicating successful RNA extraction and sufficient relatedness to *D. melanogaster* for our RpL32-targeting primers to work. Seven hundred and fourteen (75%) were positive for galbut virus RNA by RT-qPCR. We observed a multi-modal distribution of galbut virus levels, with levels in individual flies ranging from $2x10^{-6}$ to $2x10^{4}$ relative to levels of RpL32 mRNA (**Fig. 1C**).

Among Colorado locations, galbut virus prevalence varied from 40% to 100% (**Fig. 1D**). In Maine, Ohio and Pennsylvania prevalence was 77%, 33% and 97%, respectively (**Fig. 1D**). We performed a Kruskal-Wallis test with pairwise comparisons using Dunn's test to determine if there were statistically significant differences in prevalence by location. Overall, there was statistically significant evidence that prevalence varied by location (**Supp. Table 1**). Interestingly, James and Ohio had no

statistically significant differences in prevalence with any location. This could be a result of collecting low numbers of flies from these locations which decreased our power to detect differences in prevalence although we did correct for multiple comparisons.

We repeatedly sampled flies from the Wabash and Rampart locations over three months. A Kruskal-Wallis test found significant differences in galbut virus prevalence at the Rampart and Wabash locations over time. Galbut virus prevalence significantly rose from 24% in August to 89% ($p < 0.0005$). There was not a significant difference between September and October although the prevalence fell to 82% in October ($p = 1$). Similarly, at the Wabash location, there was a significant decrease in prevalence between the first two months of sampling, July and August ($p < 0.0005$) but not between August and September ($p = 0.08$) (**Fig. 1E & F**). Thus, galbut virus prevalence varied significantly over time, but in one location prevalence rose while the other location fell.

Galbut virus was highly prevalent in all populations except Ohio where recovery of *D. melanogaster* was limited. Of the 17 samples collected from Ohio, only three were positive for RpL32 mRNA and one was positive for galbut virus, a prevalence of 33%. The RpL32 mRNA Ct values for the other two samples were higher than expected for *D. melanogaster,* it is possible that these samples, which were morphologically similar to *D. melanogaster* were different species.

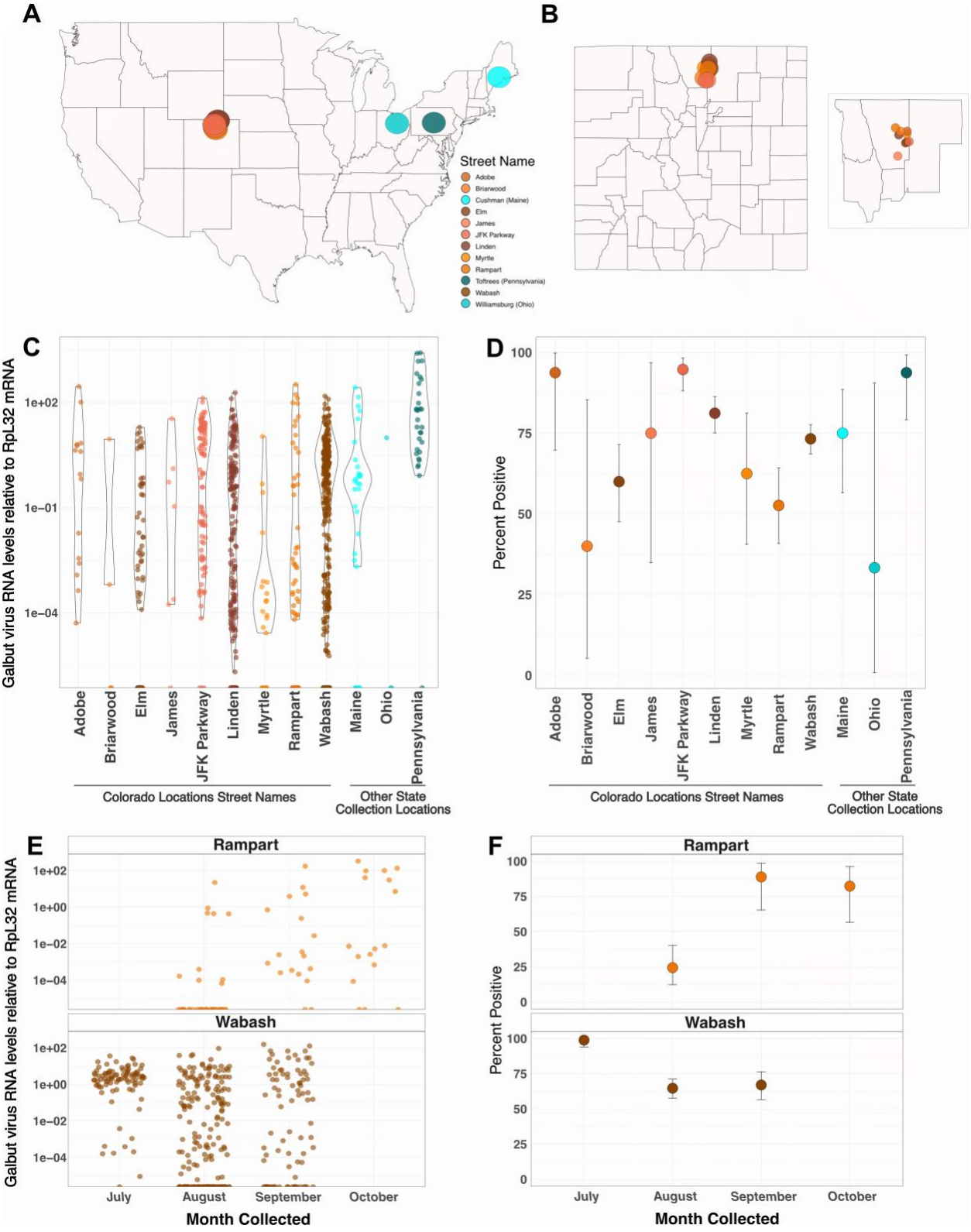


Figure 1: Galbut virus infection is widespread in populations studied in Colorado, Maine, and Pennsylvania. (A) Locations in the United States where wild *Drosophila*

were collected in 2023. (B) A map of collection locations in Colorado with Larimer county and surrounding counties (inset). (C) Galbut virus RNA levels in individual flies from Colorado (orange/brown colors) and other states (blue colors). Negative samples are plotted at the minimum Y axis value. Colorado locations are labeled using street names. (D) Galbut virus prevalence each location expressed as percent positive of RpL32-positive flies also positive for galbut virus. Binomial confidence intervals are shown. (E) Galbut virus RNA levels in individual flies over time at the Rampart and Wabash locations. (F) Galbut virus prevalence over time at the Rampart and Wabash locations. Binomial confidence intervals are shown. RT-qPCR data is normalized to levels of RpL32 mRNA.

4.3.2 Recovery of coding complete galbut virus genome sequences from 101 infected flies.

We selected 146 of the 974 samples for recovery of galbut virus genome sequences using total RNA sequencing. Samples were selected to provide a broad representation of geography, infection level, and galbut virus genotype, with genotype inferred from the pattern of qPCR positivity from our two galbut virus-targeting primer sets and qPCR melting curve differences (**Table 1, Supplemental Table 3**). We also sequenced nine samples that had been collected in 2020 and 2021 from Colorado.

We set a cutoff of 10 reads mapping to any galbut virus segment or chaq virus to identify a sample as positive by sequencing for galbut virus. Under these conditions, 16 samples had partial galbut virus sequences and 38 did not have any galbut virus mapping reads. From the 101 samples with sufficient galbut virus-mapping reads, we recovered 373 coding complete or near coding complete galbut virus and chaq virus sequences.

Of the samples that yielded no coding-complete galbut virus sequences, eight were from samples collected in Ohio that were not *D. melanogaster* and for which

species identity could not be determined. These samples were sequenced to determine if they were infected by a virus similar to galbut virus. Some of these samples were infected with partiti-like viruses but the exploration of these viruses was beyond the scope of this manuscript. An additional 15 samples from other locations had more than 10 galbut or chaq virus reads but were too low coverage to assemble coding-complete genome sequences (**Fig. 1**). Five other samples had been unexpectedly positive for only one of our galbut virus RT-qPCR primer pairs. This could be an artifact of off-target amplification of this primer pair or reflect evidence of an EVE. The remaining 10 samples were detected as low positive samples via RT-qPCR. The low positive status of these samples likely resulted in low enough virus levels that the sequences were simply not recovered.

We examined the correlation between galbut virus RNA 1 levels measured by RT-qPCR and the number of galbut-virus RNA 1 mapping reads per million (RPM) total reads. We found that these measures of galbut virus abundance were correlated, with a coefficient of determination of 0.69 (**$p < 0.001$, Fig. 2**).

Table 1: Sequenced wild *D. melanogaster* metadata.

Sample ID	State Collected	Street Collected	*Drosophila* species (1)	Galbut virus sequence detected	Coding-complete galbut virus sequence	Chaq virus sequence detected	Coding-complete chaq virus sequence	NCBI nucleotide accession	SRA accession (2)	Number of galbut and chaq virus reads (RPM)
1020_A7	Colorado	Wabash		Yes	Yes	Yes	Yes	PQ625166, PQ625079, PQ624976, PQ624867, PQ624868	SRR30712331	3599.27
1020_A8	Colorado	Wabash		No	-	No	-		SRR30712330	-
1020_A8_2	Colorado	Wabash		No	-	No	-		SRR30712232	-
1020_B11	Colorado	Wabash		No	-	No	-		SRR30712317	-
1020_B6	Colorado	Wabash		Yes	Yes	Yes	Yes	PQ625167, PQ625080, PQ624977, PQ624869	SRR30712306	4676.62
1020_C1	Colorado	Wabash		No	-	No	-		SRR30712231	-
1020_C2	Colorado	Wabash		Yes	Yes	No	-	PQ625081, PQ624978, PQ624870	SRR30712220	300.66
1020_D1	Colorado	Wabash		Yes	Yes	No	-	PQ625082, PQ624979, PQ624871	SRR30712209	390.97
1020_D3	Colorado	Wabash		Yes	Yes	Yes	Yes	PQ625168, PQ625083, PQ624980, PQ624872	SRR30712198	73.35
1020_D9	Colorado	Wabash		No	-	No	-		SRR30712187	-
1020_F_0818_D11	Colorado	Wabash		Yes	Yes	No	-	PQ625091, PQ624988, PQ624881	SRR30712329	249.08
1020_F_0818_D4	Colorado	Wabash		Yes	Yes	Yes	Yes	PQ625172, PQ625090, PQ624987, PQ624880	SRR30712287	1044.98
1020_F_0818_E4	Colorado	Wabash		Yes	No	No	-		SRR30712276	1.07
1020_F_31	Colorado	Wabash		Yes	No	Yes	Yes	PQ625170, PQ625086, PQ624983, PQ624875	SRR30712265	17538.20
1020_F_36	Colorado	Wabash		Yes	Yes	No	-	PQ625087, PQ624984, PQ624876	SRR30712254	198.88
1020_F_40	Colorado	Wabash		Yes	Yes	Yes	Yes	PQ625171, PQ625088, PQ624985, PQ624877, PQ624878	SRR30712243	546.17
1020_F_44	Colorado	Wabash		Yes	Yes	No	-	PQ625089, PQ624986, PQ624879	SRR30712237	493.13
1020_F1	Colorado	Wabash		Yes	Yes	No	-	PQ625084, PQ624981, PQ624873	SRR30712243	82.54
1020_F3	Colorado	Wabash		Yes	No	No	No		SRR30712235	5.35
1020_F6	Colorado	Wabash		Yes	Yes	Yes	Yes	PQ625169, PQ625085, PQ624982, PQ624874	SRR30712234	3661.32
1020_F8	Colorado	Wabash		No	-	No	-		SRR30712327	-
1020_G1	Colorado	Wabash		Yes	No	No	-		SRR30712326	-
1020_G3	Colorado	Wabash		Yes	Yes	No	-	PQ625092, PQ624989, PQ624882	SRR30712325	611.15
1020_H7	Colorado	Wabash		No	-	No	-		SRR30712324	-
1020_M_0818_A10	Colorado	Wabash		Yes	Yes	No	-	PQ625097, PQ624994, PQ624887	SRR30712323	553.01
1020_M_0818_A5	Colorado	Wabash		Yes	No	Yes	Yes	PQ625174, PQ625095, PQ624992, PQ624885	SRR30712322	45347.21
1020_M_0818_A7	Colorado	Wabash		Yes	Yes	Yes	Yes	PQ625175, PQ625096, PQ624993, PQ624886	SRR30712321	59.72
1020_M_19	Colorado	Wabash		Yes	Yes	Yes	Yes	PQ625173, PQ625094, PQ624991, PQ624884	SRR30712320	49691.43
1020_M_8	Colorado	Wabash		Yes	Yes	No	-	PQ625093, PQ624990, PQ624883	SRR30712319	467.89
1428_F_15	Colorado	Elm		No	-	No	-		SRR30712318	-
1428_F_8	Colorado	Elm		Yes	Yes	Yes	Yes	PQ625176, PQ625098, PQ624995, PQ624888	SRR30712316	1277.30
1428_M_3	Colorado	Elm		Yes	Yes	No	-	PQ625100, PQ624997, PQ624890, PQ624891	SRR30712315	1099.89
1428_M_4	Colorado	Elm		Yes	Yes	Yes	Yes	PQ625177, PQ625101, PQ624998, PQ624892	SRR30712314	2466.44
1428_M_6	Colorado	Elm		Yes	Yes	No	-	PQ624893	SRR30712313	47.14
1428-F-21	Colorado	Elm		Yes	Yes	No	-	PQ625099, PQ624996, PQ624889	SRR30712312	529.82
1428-M-21	Colorado	Elm		Yes	Yes	No	-	PQ625102, PQ624999, PQ624894	SRR30712311	2475.24
1428-M-36	Colorado	Elm		Yes	No	No	-		SRR30712310	-
1428-M-42	Colorado	Elm		Yes	Yes	Yes	Yes	PQ625178, PQ625103, PQ625000, PQ624895	SRR30712309	19814.01
1428-M-43	Colorado	Elm		Yes	Yes	No	No		SRR30712308	5.24
500_F_11	Colorado	Linden		Yes	Yes	Yes	Yes	PQ625155, PQ625060, PQ624956, PQ624846	SRR30712307	1246.09
500_F_12	Colorado	Linden		No	-	No	-		SRR30712305	-
500_F_21	Colorado	Linden		Yes	Yes	No	-	PQ625061, PQ624957, PQ624958, PQ624847	SRR30712304	135.92
500_F_41	Colorado	Linden		Yes	Yes	Yes	Yes	PQ625156, PQ625062, PQ624959, PQ624848, PQ624849	SRR30712303	47796.27
500_F_42	Colorado	Linden		Yes	No	No	-		SRR30712302	0.91
500_F_67	Colorado	Linden		Yes	Yes	No	-	PQ625063, PQ624960, PQ624850	SRR30712301	1203.06
500_M_100	Colorado	Linden		Yes	No	No	-		SRR30712300	1.10
500_M_16	Colorado	Linden		Yes	Yes	No	-	PQ625064, PQ624961, PQ624851	SRR30712299	603.21
500_M_17	Colorado	Linden		Yes	Yes	Yes	Yes	PQ625157, PQ625065, PQ624962, PQ624852	SRR30712298	3975.30
500_M_40	Colorado	Linden		No	-	No	-		SRR30712297	-
500_M_54	Colorado	Linden		Yes	Yes	No	-	PQ625066, PQ624963, PQ624853	SRR30712233	2435.91
500_M_55	Colorado	Linden		Yes	No	Yes	-		SRR30712230	-
500_M_57	Colorado	Linden		Yes	Yes	Yes	Yes	PQ625158, PQ625067, PQ624964, PQ624854	SRR30712229	2802.57
500_M_6	Colorado	Linden		No	-	No	-		SRR30712228	-
500_M_60	Colorado	Linden		Yes	Yes	No	-	PQ625068, PQ624965, PQ624855, PQ624856	SRR30712227	642.06
500_M_61	Colorado	Linden		Yes	Yes	Yes	Yes	PQ625160, PQ625070, PQ624967, PQ624858	SRR30712226	5206.74
500_M_61_2	Colorado	Linden		Yes	Yes	Yes	Yes	PQ625159, PQ625069, PQ624966, PQ624857	SRR30712225	954.02
500_M_65	Colorado	Linden		Yes	Yes	Yes	Yes	PQ625162, PQ625072, PQ624969, PQ624860	SRR30712224	3958.83
500_M_65_2	Colorado	Linden		Yes	Yes	Yes	Yes	PQ625161, PQ625071, PQ624968, PQ624859	SRR30712223	1645.09
500_M_71	Colorado	Linden		Yes	No	No	-		SRR30712222	1.92
500_M_71_2	Colorado	Linden		Yes	No	Yes	Yes	PQ625163, PQ625073, PQ624970, PQ624861	SRR30712221	18123.15
500_M_83	Colorado	Linden		Yes	Yes	No	-	PQ625074, PQ624971, PQ624862	SRR30712219	2917.61
500_M_84	Colorado	Linden		Yes	No	Yes	Yes	PQ625164, PQ625075, PQ624972, PQ624863	SRR30712218	56683.29
500_M_F10	Colorado	Linden		Yes	Yes	Yes	Yes	PQ625165, PQ625076, PQ624973, PQ624864	SRR30712217	2914.68
500_M_F11	Colorado	Linden		Yes	Yes	No	-	PQ625077, PQ624974, PQ624865	SRR30712216	237.42
500_M_F9	Colorado	Linden		Yes	No	Yes	No		SRR30712215	0.98
500_M_G2	Colorado	Linden		Yes	Yes	No	-	PQ625078, PQ624975, PQ624866	SRR30712214	205.36
Adobe-B4	Colorado	Adobe		Yes	Yes	No	-	PQ625104, PQ625001, PQ624896	SRR30712213	489.13
Adobe-B6	Colorado	Adobe		Yes	Yes	No	-	PQ625105, PQ625002, PQ624897	SRR30712212	653.93
Adobe-B8	Colorado	Adobe		Yes	Yes	Yes	Yes	PQ625179, PQ625106, PQ625003, PQ624898	SRR30712214	62982.91

Adobe-C1	Colorado	Adobe		No	-	No	-		SRR30712210	-
Adobe-C2	Colorado	Adobe		Yes	Yes	No	-	PQ625107, PQ625004, PQ624899	SRR30712208	1357.89
Briarwood_D10	Colorado	Briarwood		No	-	No	-		SRR30712207	-
Briarwood_D9	Colorado	Briarwood		Yes	Yes	Yes	Yes	PQ625180, PQ625108, PQ625005, PQ624900	SRR30712206	4506.78
CVID_0825_B1	Colorado	Rampart		Yes	Yes	No	-	PQ625109, PQ625006, PQ624901	SRR30712205	269.69
CVID_0825_C10	Colorado	Rampart		Yes	Yes	No	-	PQ625110, PQ625007, PQ624902	SRR30712204	1130.62
CVID_0911_E3	Colorado	Rampart		No	-	No	-		SRR30712203	-
CVID_0911_F3	Colorado	Rampart		No	-	No	-		SRR30712202	-
CVID_0911_F4	Colorado	Rampart		Yes	Yes	Yes	Yes	PQ625181, PQ625111, PQ625008, PQ624903	SRR30712201	2960.22
CVID_1006_G6	Colorado	Rampart		Yes	Yes	Yes	Yes	PQ625182, PQ625112, PQ625009, PQ624904	SRR30712200	1692.98
CVID_1006_G7	Colorado	Rampart		Yes	Yes	Yes	Yes	PQ625183, PQ625113, PQ625010, PQ624905	SRR30712199	2576.28
CVID_1006_H2	Colorado	Rampart		Yes	Yes	Yes	Yes	PQ625184, PQ625114, PQ625011, PQ624906	SRR30712197	4147.74
CVID_1006_H3	Colorado	Rampart		Yes	Yes	Yes	Yes	PQ625185, PQ625115, PQ625012, PQ624907	SRR30712196	4751.59
CVID_1006_H4	Colorado	Rampart		Yes	No	No	-		SRR30712195	-
hydei-20-TD-1	Colorado	Laurel	Drosophila affinis	No	-	No	-		SRR30712194	-
James-A4	Colorado	James		Yes	Yes	No	-	PQ625116, PQ625013, PQ624908	SRR30712193	1173.77
James-A5	Colorado	James		Yes	Yes	Yes	Yes	PQ625186, PQ625117, PQ625014, PQ624909	SRR30712192	83.23
ME-F-1	Maine	Cushman		Yes	Yes	Yes	Yes	PQ625187, PQ625188, PQ625118, PQ625015, PQ624910	SRR30712291	5491.59
ME-F-2	Maine	Cushman		Yes	Yes	Yes	Yes	PQ625189, PQ625016, PQ624911	SRR30712290	35.69
ME-F-3	Maine	Cushman		Yes	Yes	Yes	Yes	PQ625190, PQ625119, PQ625017, PQ624912	SRR30712289	15162.59
ME-F-9	Maine	Cushman		Yes	Yes	No	No	PQ625018, PQ624913	SRR30712288	206.79
ME-M-1	Maine	Cushman		Yes	Yes	Yes	Yes	PQ625191, PQ625120, PQ625019, PQ624914	SRR30712286	12582.52
ME-M-10	Maine	Cushman		Yes	Yes	No	-	PQ625127, PQ625026, PQ624922	SRR30712285	3059.10
ME-M-11	Maine	Cushman		Yes	Yes	Yes	Yes	PQ625196, PQ625128, PQ625027, PQ624923	SRR30712284	12009.36
ME-M-13	Maine	Cushman		Yes	Yes	No	No		SRR30712283	2.93
ME-M-14	Maine	Cushman		Yes	Yes	No	-	PQ625129, PQ625028, PQ624924	SRR30712282	3291.81
ME-M-17	Maine	Cushman		Yes	Yes	No	No	PQ625130, PQ625029, PQ624925	SRR30712281	3262.84
ME-M-18	Maine	Cushman		Yes	Yes	No	-	PQ625131, PQ625030, PQ624926	SRR30712280	1268.57
ME-M-19	Maine	Cushman		Yes	Yes	No	-	PQ625132, PQ625031, PQ624927	SRR30712279	8923.71
ME-M-2	Maine	Cushman		Yes	Yes	Yes	Yes	PQ625192, PQ625193, PQ625121, PQ625020, PQ624915	SRR30712269	11838.56
ME-M-20	Maine	Cushman		Yes	Yes	No	No	PQ625133, PQ625032, PQ624928	SRR30712268	2589.53
ME-M-3	Maine	Cushman	undetermined	Yes	Yes	Yes	Yes	PQ625214, PQ625153, PQ625056, PQ624952	SRR30712267	702.89
ME-M-4	Maine	Cushman		Yes	Yes	No	-	PQ625122, PQ625021, PQ624916	SRR30712266	2168.10
ME-M-6	Maine	Cushman		Yes	Yes	No	-	PQ625123, PQ625022, PQ624917	SRR30712264	5739.65
ME-M-7	Maine	Cushman		Yes	Yes	Yes	Yes	PQ625194, PQ625124, PQ625125, PQ625023, PQ625024, PQ624918, PQ624919	SRR30712263	3142.71
ME-M-8	Maine	Cushman		Yes	Yes	Yes	Yes	PQ625195, PQ625126, PQ625025, PQ624920, PQ624921	SRR30712262	26.20
ME-M-9	Maine	Cushman		Yes	No	No	-		SRR30712261	0.75
mel-20-AG-4	Colorado	Birch		No	-	No	-		SRR30712260	-
mel-20-AG-5	Colorado	Birch		No	-	No	-		SRR30712259	-
mel-20-TD-3	Colorado	Laurel		Yes	Yes	No	-	PQ625057, PQ624953, PQ624841, PQ624842	SRR30712258	141.44
mel-21-LK-5	Colorado	Heatheridge		Yes	No	No	-		SRR30712257	0.83
mel-21-MDS-9	Colorado	Michener		Yes	Yes	Yes	Yes	PQ625154, PQ625059, PQ624955, PQ624845	SRR30712256	2286.00
Myrtle_E10	Colorado	Myrtle		Yes	No	Yes	No		SRR30712240	-
Myrtle_E9	Colorado	Myrtle		Yes	Yes	Yes	Yes	PQ625197, PQ625136, PQ625035, PQ624931	SRR30712239	2417.33
Myrtle-D1	Colorado	Myrtle		No	-	No	-		SRR30712238	-
Myrtle-D11	Colorado	Myrtle		Yes	Yes	No	-	PQ625135, PQ625034, PQ624930	SRR30712191	665.50
Myrtle-D5	Colorado	Myrtle		Yes	Yes	No	-	PQ625134, PQ625033, PQ624929	SRR30712190	178.32
Myrtle-D8	Colorado	Myrtle		Yes	No	No	-		SRR30712189	0.94
Myrtle-E4	Colorado	Myrtle		No	-	No	-		SRR30712188	-
OH_M_3	Ohio	Williamsburg	undetermined	No	-	No	-		SRR30712186	-
OH_M_4	Ohio	Williamsburg	undetermined	No	-	No	-		SRR30712185	-
OH_M_5	Ohio	Williamsburg		Yes	Yes	Yes	Yes	PQ625198, PQ625137, PQ625036, PQ624932	SRR30712184	11008.06
OH_M_6	Ohio	Williamsburg	undetermined	No	-	No	-		SRR30712183	-
OH_M_7	Ohio	Williamsburg		No	-	No	-		SRR30712182	-
OH-F-1	Ohio	Williamsburg	undetermined	No	-	No	-		SRR30712181	-
OH-F-2	Ohio	Williamsburg	undetermined	No	-	No	-		SRR30712180	-
OH-F-3	Ohio	Williamsburg	undetermined	No	-	No	-		SRR30712179	-
Peach_1_B2	Colorado	JFK Parkway		Yes	No	No	-		SRR30712178	-
Peach_10_F10	Colorado	JFK Parkway		Yes	Yes	No	-	PQ625140, PQ625039, PQ624935	SRR30712177	1807.48
Peach_11_F11	Colorado	JFK Parkway		No	-	No	-		SRR30712328	-
Peach_12_H1	Colorado	JFK Parkway		No	-	No	-		SRR30712296	-
Peach_13_H7	Colorado	JFK Parkway		Yes	Yes	Yes	Yes	PQ625199, PQ625141, PQ625040, PQ624936, PQ624937	SRR30712295	37.68
Peach_2_B4	Colorado	JFK Parkway		Yes	No	Yes	No		SRR30712294	3.23
Peach_3_B5	Colorado	JFK Parkway		Yes	No	No	-		SRR30712293	18.77
Peach_4_B7	Colorado	JFK Parkway		Yes	No	No	-		SRR30712292	5.49
Peach_5_B9	Colorado	JFK Parkway		Yes	No	Yes	No		SRR30712278	4.15
Peach_6_E5	Colorado	JFK Parkway		No	-	No	-		SRR30712277	-
Peach_7_F6	Colorado	JFK Parkway		No	-	No	-		SRR30712275	-
Peach_8_F7	Colorado	JFK Parkway		Yes	Yes	No	-	PQ625138, PQ625037, PQ624933	SRR30712274	2310.65
Peach_9_F8	Colorado	JFK Parkway		Yes	Yes	No	No	PQ625139, PQ625038, PQ624934	SRR30712273	1969.55
Penn-F-1	Pennsylvania	Toftrees		Yes	No	Yes	Yes	PQ625200, PQ625041, PQ624938	SRR30712272	52.47
Penn-F-12	Pennsylvania	Toftrees		Yes	Yes	Yes	Yes	PQ625207, PQ625147, PQ625049, PQ624944	SRR30712271	475.17
Penn-F-2	Pennsylvania	Toftrees		Yes	No	Yes	Yes	PQ625201, PQ625202, PQ625142, PQ625042, PQ624939	SRR30712270	52234.45
Penn-F-3	Pennsylvania	Toftrees		Yes	Yes	Yes	Yes	PQ625203, PQ625143, PQ625043, PQ624940	SRR30712255	10590.83
Penn-F-4	Pennsylvania	Toftrees		Yes	Yes	Yes	Yes	PQ625204, PQ625144, PQ625145, PQ625044, PQ625045, PQ624941, PQ624942	SRR30712253	1254.28
Penn-F-6	Pennsylvania	Toftrees		Yes	Yes	Yes	Yes	PQ625205, PQ625046, PQ625047	SRR30712252	18.44
Penn-F-9	Pennsylvania	Toftrees		Yes	Yes	Yes	Yes	PQ625206, PQ625146, PQ625048, PQ624943	SRR30712251	42525.03
Penn-M-10	Pennsylvania	Toftrees		Yes	Yes	Yes	Yes	PQ625211, PQ625150, PQ625052, PQ624948	SRR30712250	110903.51
Penn-M-11	Pennsylvania	Toftrees		Yes	Yes	Yes	Yes	PQ625212, PQ625053, PQ624949	SRR30712276	80.31
Penn-M-14	Pennsylvania	Toftrees		Yes	Yes	Yes	Yes	PQ625213, PQ625151, PQ625054, PQ624950	SRR30712248	84.19
Penn-M-15	Pennsylvania	Toftrees		Yes	No	Yes	No	PQ625152, PQ625055, PQ624951	SRR30712247	58329.87
Penn-M-2	Pennsylvania	Toftrees		Yes	Yes	Yes	Yes	PQ625208, PQ625209, PQ625148, PQ625050, PQ624945, PQ624946	SRR30712246	899.05
Penn-M-8	Pennsylvania	Toftrees		Yes	Yes	Yes	Yes	PQ625210, PQ625149, PQ625051, PQ624947	SRR30712245	43.38
sim-20-AG-3	Colorado	Birch		Yes	No	Yes	No		SRR30712244	0.29
sim-20-TD-4	Colorado	Laurel		Yes	Yes	No	-	PQ625058, PQ624954, PQ624843, PQ624844	SRR30712242	230.64
sim-20-TD-5	Colorado	Laurel		No	-	No	-		SRR30712241	-

(1) D. melanogaster unless otherwise noted

(2) Sequence Read Archive

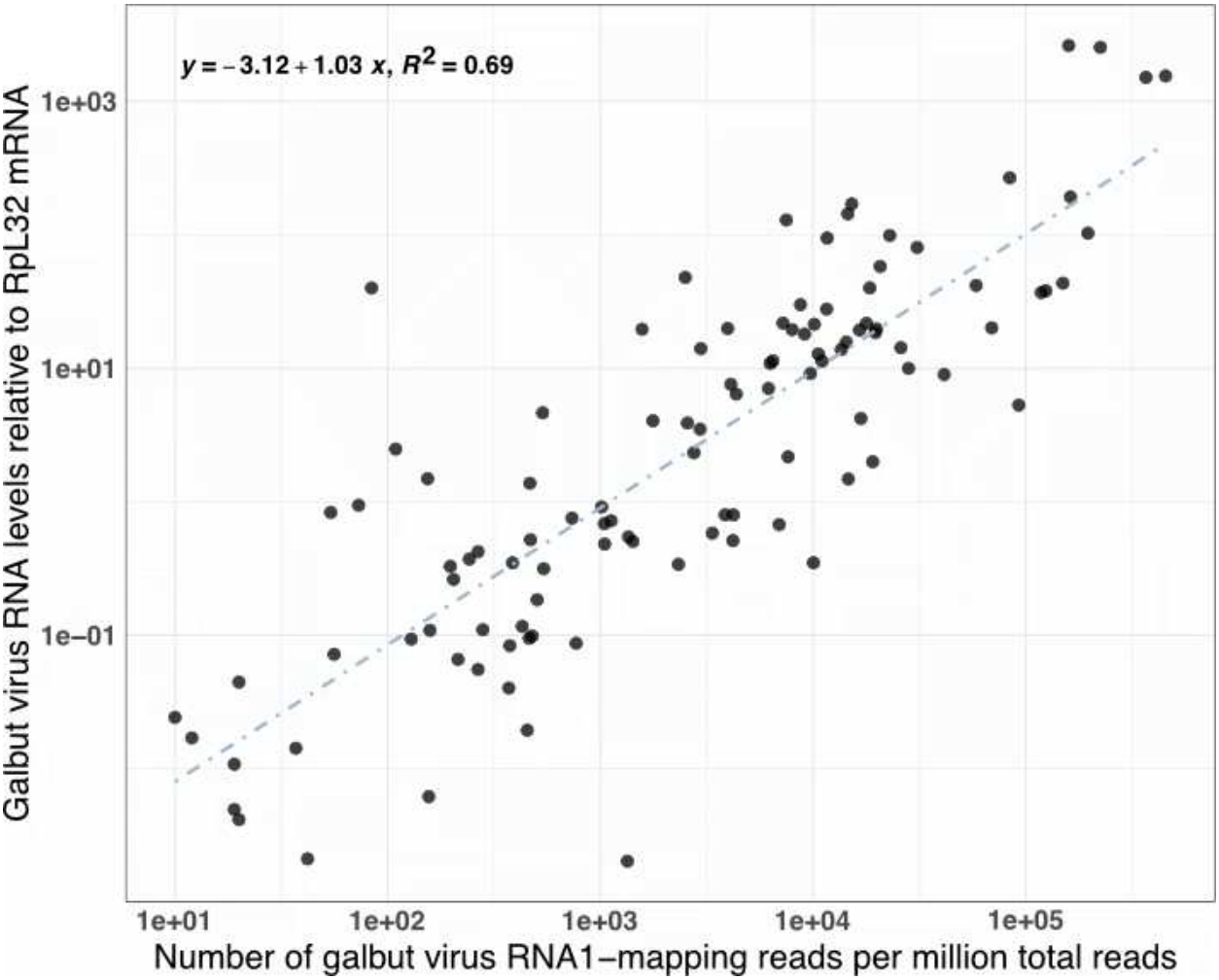


Figure 2: Correlation between galbut virus RNA levels measured by RT-qPCR and sequencing. RNA levels measured by RT-qPCR and by the number of reads to galbut virus RNA 1 are plotted. Linear regression line is plotted. Samples with fewer than 10 galbut virus or chaq virus mapping reads were omitted.

4.3.3 Galbut virus coinfection is common and does not require complete sets of all three segments

Many samples contained evidence of co-infection by more than one galbut virus genotype. Forty-one samples (33%) exhibited evidence of coinfection by various numbers of galbut virus and chaq virus segments (**Fig. 3**). Three samples corresponded to simple co-infections, with two complete sets of all three galbut virus segments (**Fig.**

3). These were the only samples with two coding-complete RNA1 genotypes. Other co-infections yielded variable numbers of coding-complete sequences for the different segments. For instance, from sample 1020-A7, we assembled 1 genotype for RNA 1, RNA 2, and chaq virus, but two distinct RNA 3 genotypes that shared 98.8% pairwise nucleotide identity. Sample ME-F-1 produced one RNA 1, two RNA 2s, three RNA 3s and two chaq virus sequences. Only one of the RNA 2 and RNA 3 sequences were complete. The two chaq virus sequences shared 91.8% pairwise nucleotide identity. This was the only sample to show evidence of three genotypes of one segment (**Fig. 3**). This type of complex co-infection has been observed for other segmented RNA viruses[15–17].

Some samples with fewer galbut virus-mapping reads did not produce a complete set of galbut virus segments (**Fig. 3B**). RNA 1 was the most common coding-incomplete sequence (**Fig. 3**). Some datasets, including datasets with numerous galbut virus-mapping reads, did not produce chaq virus sequences (**Fig. 3**). This reflects chaq virus's likely status as a satellite of galbut virus or an optional 4th galbut virus segment[1,2,8,18].

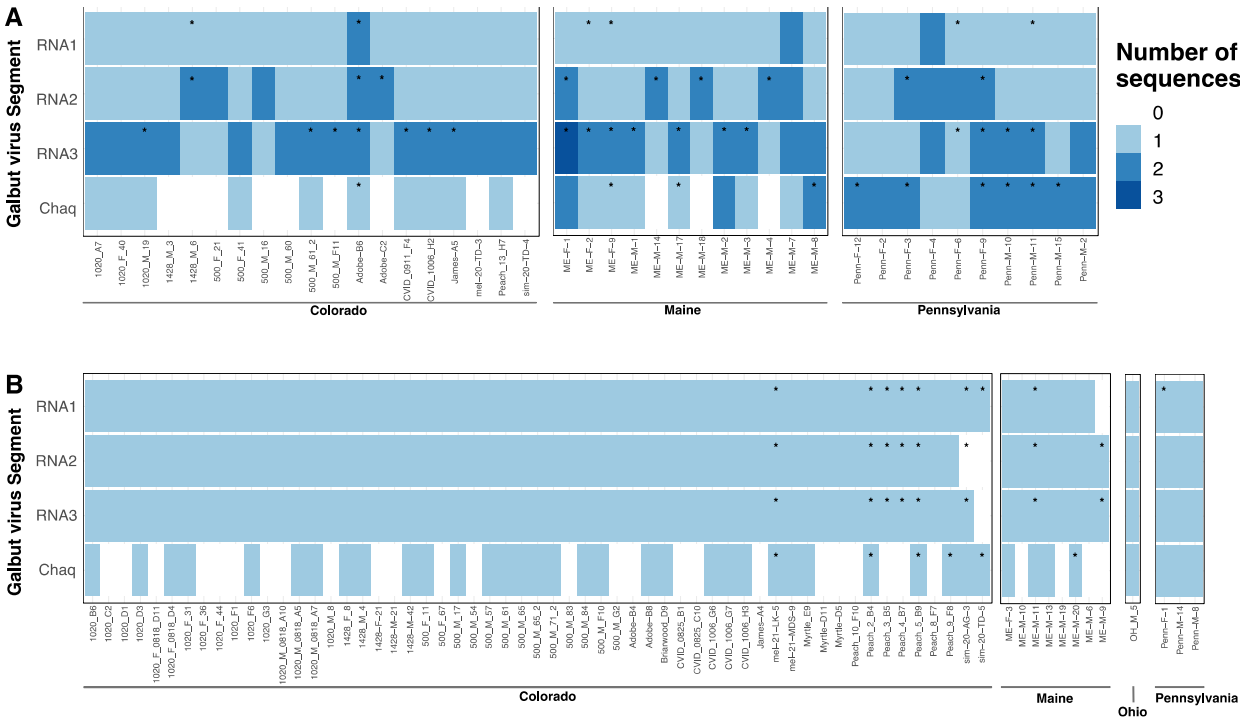


Figure 3: Galbut virus coinfection is common and does not require complete sets of all three segments. Heatmaps depicting the number of sequences recovered for each segment for (A) samples with evidence of coinfection, or (B) samples with no evidence of coinfection. Asterisks (*) indicate partial sequences where coverage was not sufficient to recover coding-complete sequences.

4.3.4 Galbut virus diversity and evolutionary history

We generated maximum likelihood phylogenetic trees using the galbut virus sequences generated here and those available in the National Center for Biotechnology Information (NCBI) nucleotide database. To be included, existing sequences had to derive from a single fly (not a pool) and have a known collection year and location. We also generated galbut virus RNA 2 and 3 sequences from Sequence Read Archive (SRA) datasets from Australian *D. simulans* for which galbut virus RNA 1 and Chaq virus sequences were already available[8]. These new Australian sequences have been deposited to the NCBI nucleotide database under accessions PV185866-PV185894.

Galbut virus sequences clustered into several major genotypes. We defined galbut virus genotypes as groups of sequences that shared > ~95% pairwise nucleotide identity and that formed monophyletic clades. Galbut virus RNA 1 and RNA 2 sequences grouped into three major genotypes, two containing viruses from *D. melanogaster* and the third containing viruses from *D. simulans*. Following existing convention, we termed the two *D. melanogaster* clades clade A and clade B (**Figs. 4 & 5**)[11]. For RNA 1 and RNA 2, the clade B sequences further subdivided into clades we named B.1 and B.2 (**Figs. 4 & 5**). RNA 3 also had three major genotypes, including subclades B.1 and B.2. Additionally, RNA 3 clade A subdivided into two clades, A.1 and A.2 (**Fig. 6**). Chaq virus clustered into three clades, which we termed A.1, A.2, and *D. simulans* (**Fig. 7**).

Overall, galbut virus diversity within clades was low: many sequences shared over 99% pairwise nucleotide identity (**Figs. 4, 5, 6 & 7**). However, galbut virus diversity between clades was higher. Between clades pairwise identity values were ~87%, ~88%, ~76% and ~80% for RNA 1, RNA 2 RNA 3 and chaq virus, respectively. Full maximum likelihood trees without collapsed tips can be found in **Supplemental Figures 1, 2, 3, and 4**.

Geography partially explained phylogenetic clustering (**Figs. 4, 5, 6 & 7**). Subsets of samples from the same location and the same date tended to form clusters of identical or nearly identical sequences. For example, most RNA 1 sequences from flies from Pennsylvania formed a monophyletic cluster (**Fig. 4**). Other sequences from Pennsylvania were spread across the RNA 1 tree. In general, sequences from different locations, including Australia, were intermixed throughout trees. These patterns support

the existence of location-specific micro-lineages that probably reflected shared ancestry through vertical transmission. Individual locations did harbor galbut virus diversity from across the phylogeny, which likely reflects global dispersal of *D. melanogaster*. Samples obtained from museum specimens clustered near contemporary sequences but were not identical to contemporary sequences (**Figs. 4, 5 & 6**)[11].

There were differences in the relative branching order of the *D. melanogaster* and *D. simulans* clades in the different trees (**Figs. 4, 5 & 6**). In the galbut virus RNA 2 and RNA 3 trees, the *D. simulans* clade was more closely related to galbut virus clade A. In the RNA 1 tree, the *D. simulans* clade was more closely related to clade B (**Figs. 4, 5 & 6**). It is possible that these topological differences reflect a historic reassortment event. Alternatively, the difference could result from an artifact of rooting on tree topology.

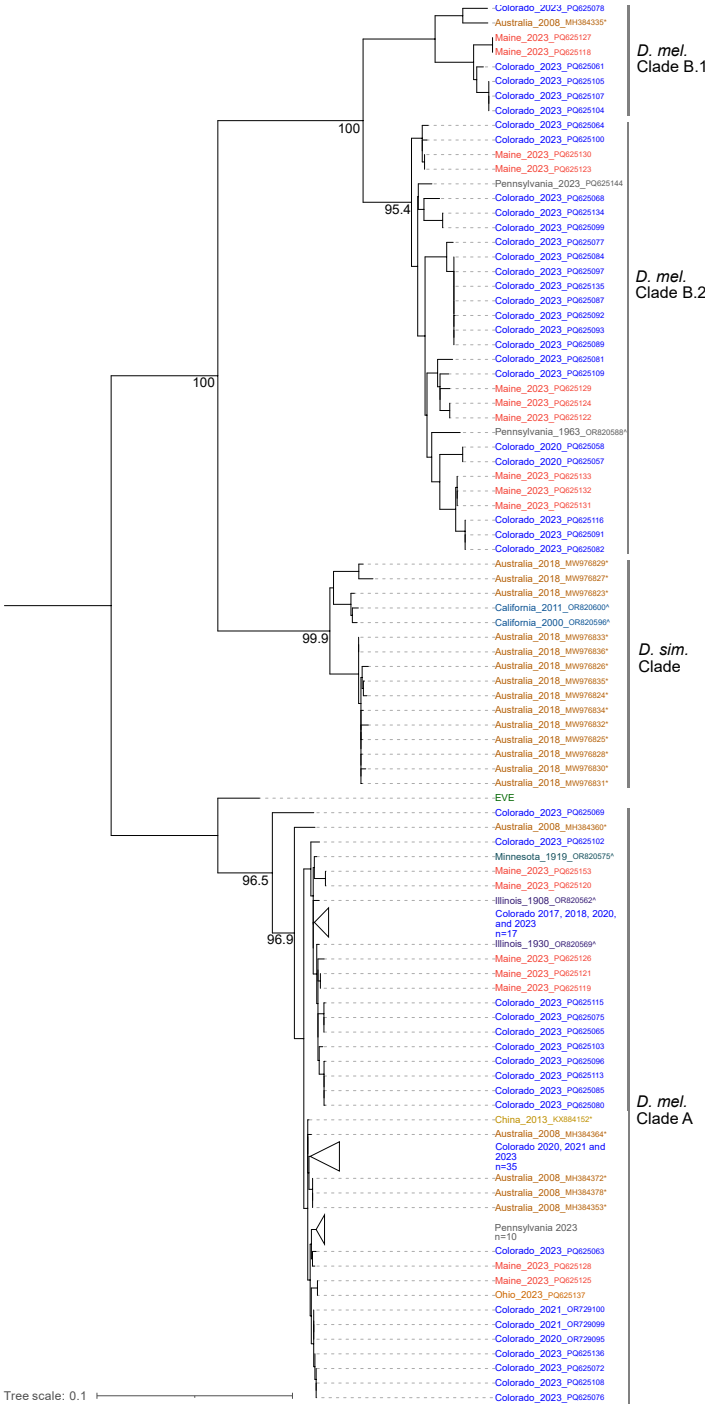


Figure 4: Midpoint rooted maximum likelihood phylogenetic tree of galbut virus RNA 1 nucleotide sequences. Sample names are colored by location. Sequences generated by others are indicated with an asterisk (*). Sequences from museum specimens are indicated by a caret (^)[11]. EVE indicates an endogenized galbut virus sequence present in some *D. melanogaster* genomes[10]. Support values for select nodes are indicated. Accession numbers are noted except where groups of closely related sequences were collapsed.

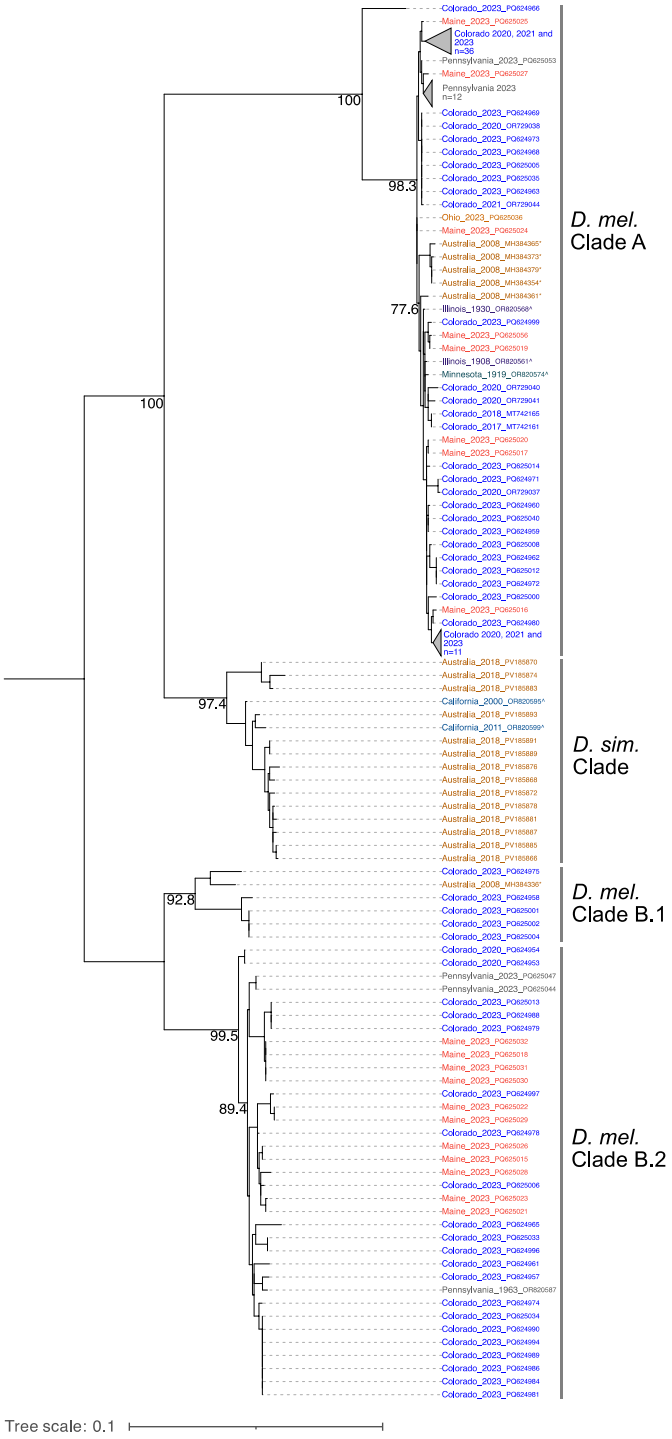


Figure 5: Midpoint rooted maximum likelihood phylogenetic tree of galbut virus RNA2 nucleotide sequences. Sample names are colored by location. Sequences generated by others are indicated with an asterisk (*). Sequences from museum specimens are indicated by a caret (^)[11]. Support values for select nodes are indicated. Accession numbers are noted except where groups of closely related sequences were collapsed.

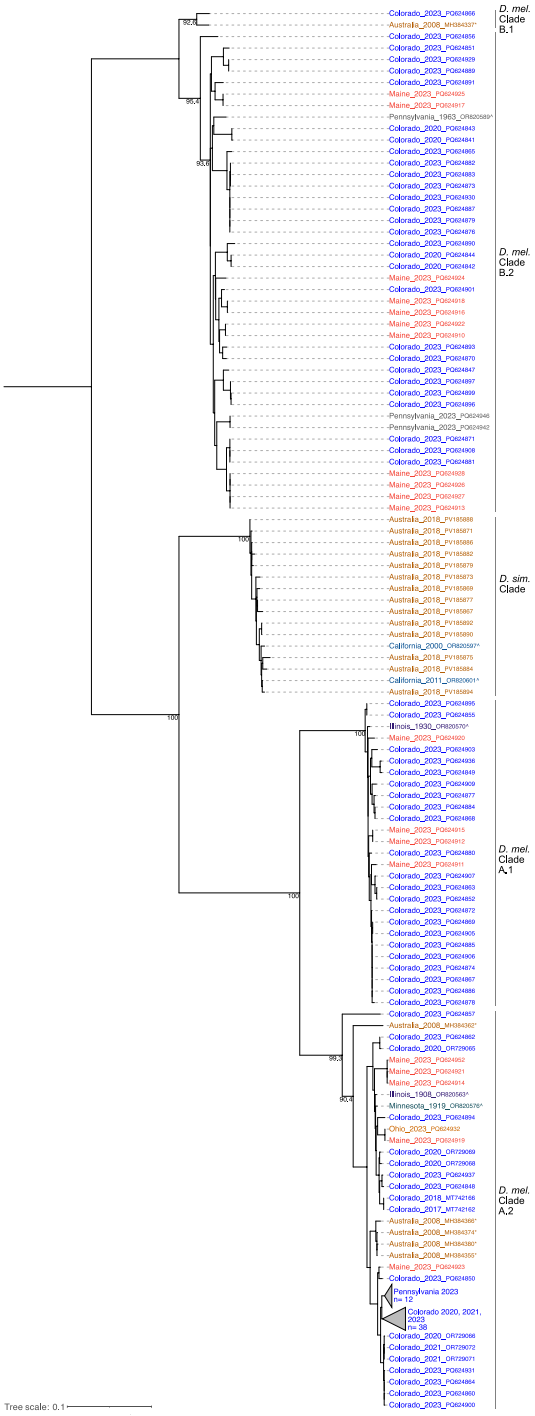


Figure 6: Midpoint rooted maximum likelihood phylogenetic tree of galbut virus RNA3 nucleotide sequences. Sample names are colored by location. Sequences generated by other investigators are indicated with an asterisk (*). Sequences from museum specimens are indicated by a caret (^)[11]. Support values for select nodes are indicated. Accession numbers are noted except where groups of closely related sequences were collapsed.

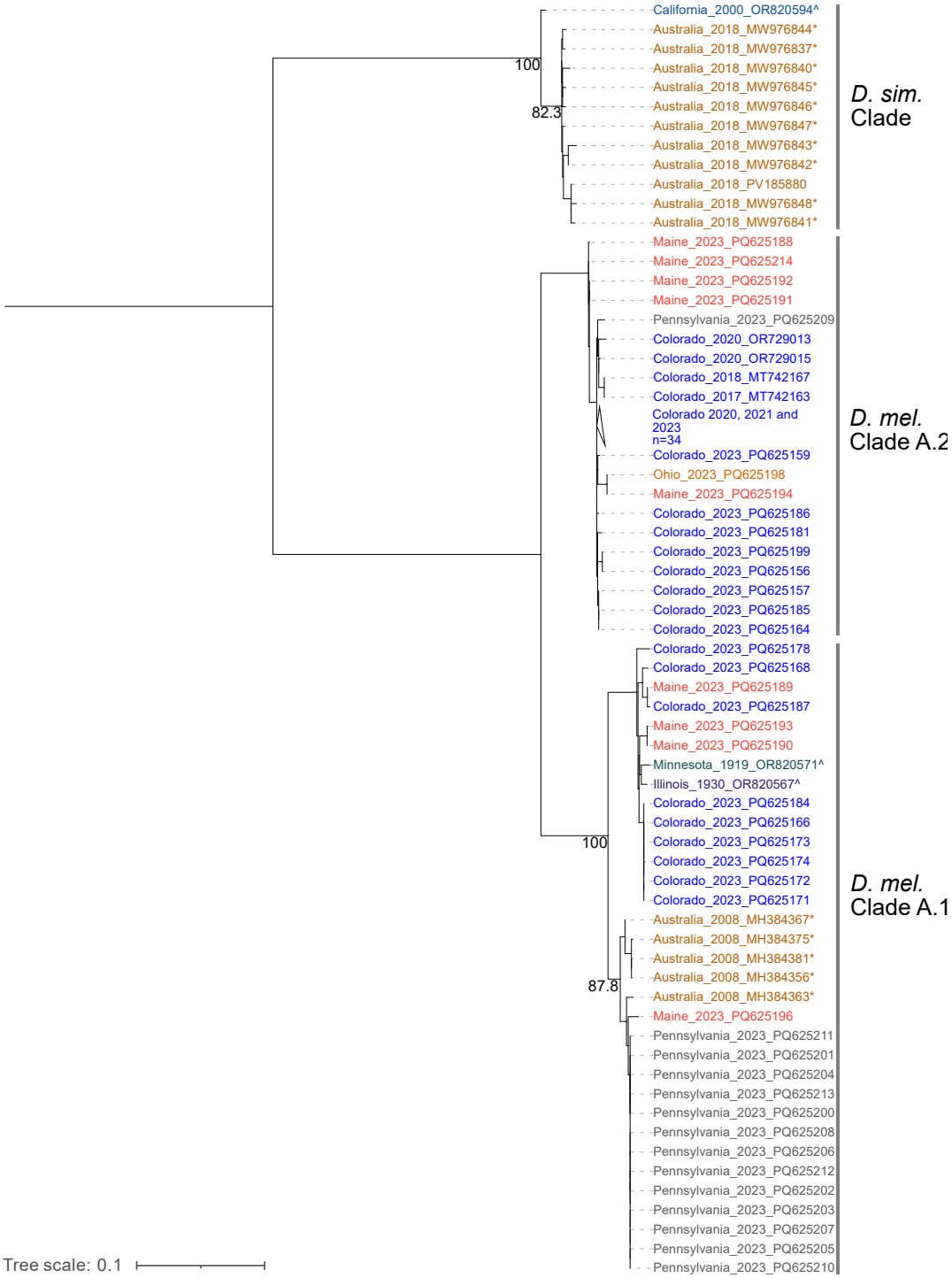


Figure 7: Midpoint rooted maximum likelihood phylogenetic tree of chaq virus. Sample names are colored by location. Sequences generated by other investigators are indicated with an (*). Sequences from museum specimens are indicated by a (^)[11]. Support values for select nodes are indicated. Accession numbers are noted except where groups of closely related sequences were collapsed.

We quantified the within-segment diversity of galbut virus and chaq virus sequences. Galbut virus segments exhibited varying levels of sequence diversity (Kruskal-Wallis test, $p < 2.2x10^{-16}$). RNA 3 sequences exhibited the most diversity and RNA 2 the least (**Figs. 4, 5, 6, and 8**). The median percent identity between RNA 1, RNA 2, RNA 3, and chaq sequences was 89.7%, 91.7%, 83.0%, 91.1%, respectively (**Fig. 8**). Because there are no clade B chaq virus sequences (see below), the diversity of chaq sequences is underestimated by this measure.

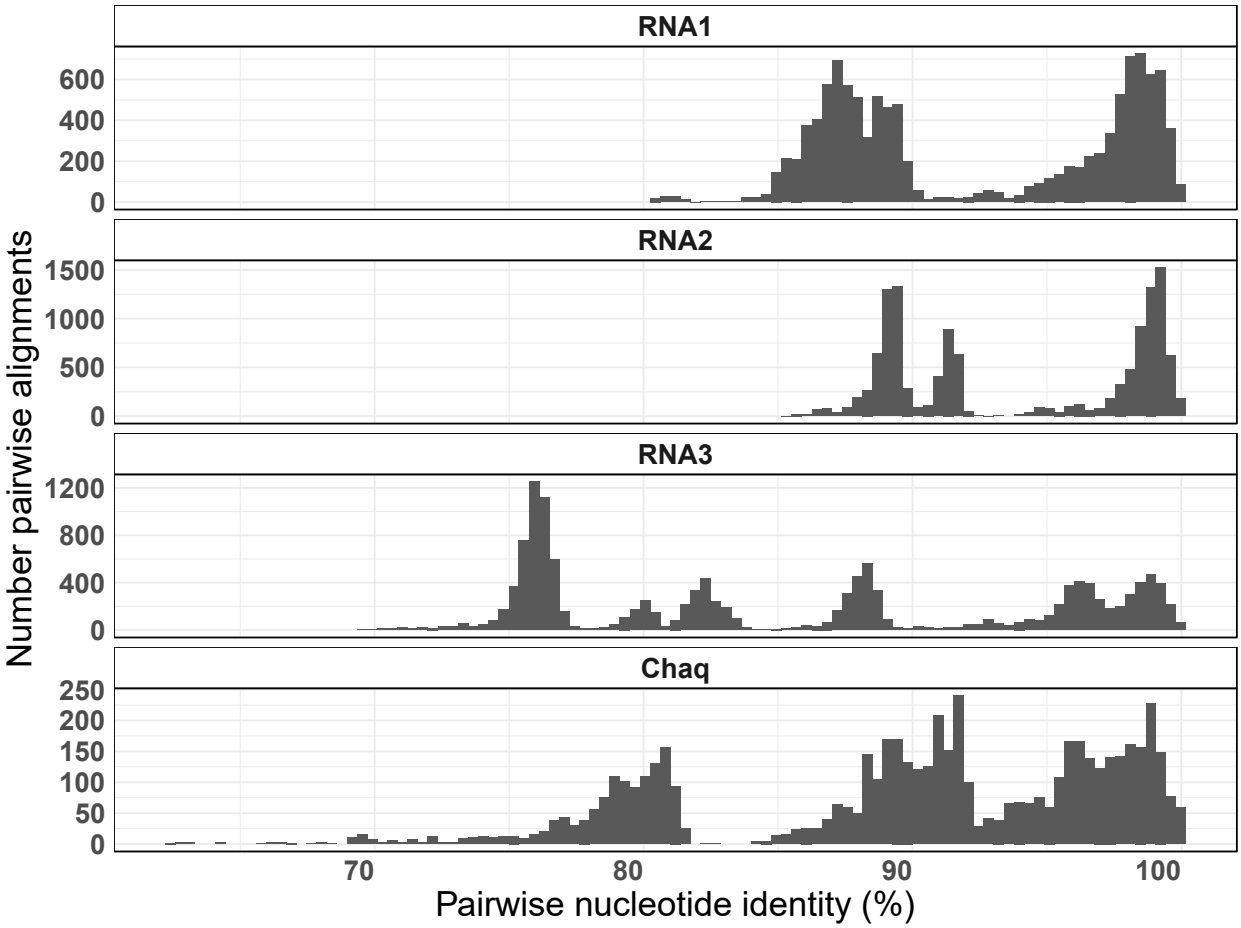


Figure 8: Galbut virus RNA 3 is more diverse than galbut virus RNAs 1 and 2 or chaq virus. The pairwise nucleotide identity (%) is shown as a histogram of the number of pairwise alignments for each nucleotide at each position for individual galbut virus segments and chaq virus.

We examined galbut virus sequences for evidence of natural selection. Botella et al reported purifying selection in the RNA 1 and RNA 2 segments of a fungus-infecting partitivirus[19]. We employed the Fixed Effects Likelihood model to quantify evidence of selection in each galbut virus segment and chaq virus[20]. There was little evidence of diversifying selection in galbut virus RNA 1, RNA 2, or chaq virus with 2, 4 and 2 diversifying sites identified, respectively (**Fig. 9 & Table 2**). However, galbut virus RNA 3 showed more evidence of diversifying selection, with 20 sites identified (**Fig. 9 & Table 2**). We observed that some sites of diversifying selection appeared to cluster together. This is particularly true for sites in RNA 3 (**Figure 9**). For all segments there was more evidence of purifying selection, with 135, 58, 51, and 21 sites experiencing purifying selection identified in galbut virus segments RNA 1, 2, 3 and chaq virus respectively (**Figure 9 & Supplemental Table 1**).

Table 2: Sites experiencing diversifying selection.

Segment	Codon	Alpha (1)	Beta (2)	Alpha=Beta (3)	LRT (4)	p-value (5)	Total Branch Length (6)
Chaq	24	0	2.63	0.58	2.89	0.09	0.01
Chaq	109	0	12.61	0.7	6.41	0.01	0.01
RNA1	43	0	14.17	1.02	5.88	0.02	0.1
RNA1	412	0	2.18	0.61	3.01	0.08	0.05
RNA2	78	0	3.07	1.35	3.17	0.08	0.02
RNA2	136	0	3.45	1.11	4.3	0.04	0.02
RNA2	191	0	3.74	1.17	4.48	0.03	0.02
RNA2	431	0	2.39	0.52	2.99	0.08	0.06
RNA3	14	0	1.85	0.41	2.71	0.1	0
RNA3	17	0.8	43.03	1.57	4.32	0.04	0.01
RNA3	24	0	26.57	2.51	8.02	0	0.01
RNA3	31	0	10.96	1.66	6.54	0.01	0
RNA3	55	0	15.99	0.41	6.14	0.01	0.01
RNA3	57	0	41.82	1.27	3.55	0.06	0.01
RNA3	91	0.7	65.34	1.94	5.08	0.02	0.01
RNA3	94	0.89	6542.52	1.81	7.43	0.01	0.01
RNA3	96	0	2.91	0.36	3.76	0.05	0
RNA3	98	0	2.15	0.36	3.14	0.08	0
RNA3	106	0.58	4.42	1.47	2.8	0.09	0
RNA3	111	0.5	5.16	1.32	3.62	0.06	0
RNA3	147	0	2.05	0.35	3.17	0.08	0.01
RNA3	159	0	2.5	0.72	4.15	0.04	0.01
RNA3	233	0	13.82	1.3	7.08	0.01	0.01
RNA3	255	0.72	10.24	2.08	3.88	0.05	0.01
RNA3	260	0	2.33	0.39	3.14	0.08	0
RNA3	442	0	2.64	0.71	4.66	0.03	0.01
RNA3	448	0	2.14	0.35	3.19	0.07	0
RNA3	449	0	20.87	1.92	9.19	0	0.01
(1) Synonymous substitution rate at a site							
(2) Non-synonymous substitution rate at a site							
(3) The rate estimate under the neutral model							
(4) Likelihood ratio test statistic for beta = alpha, versus beta /= alpha							
(5) Asymptotic p-value for evidence of selection							
(6) The total length of branches contributing to inference at this site, used to scale dN-dS							

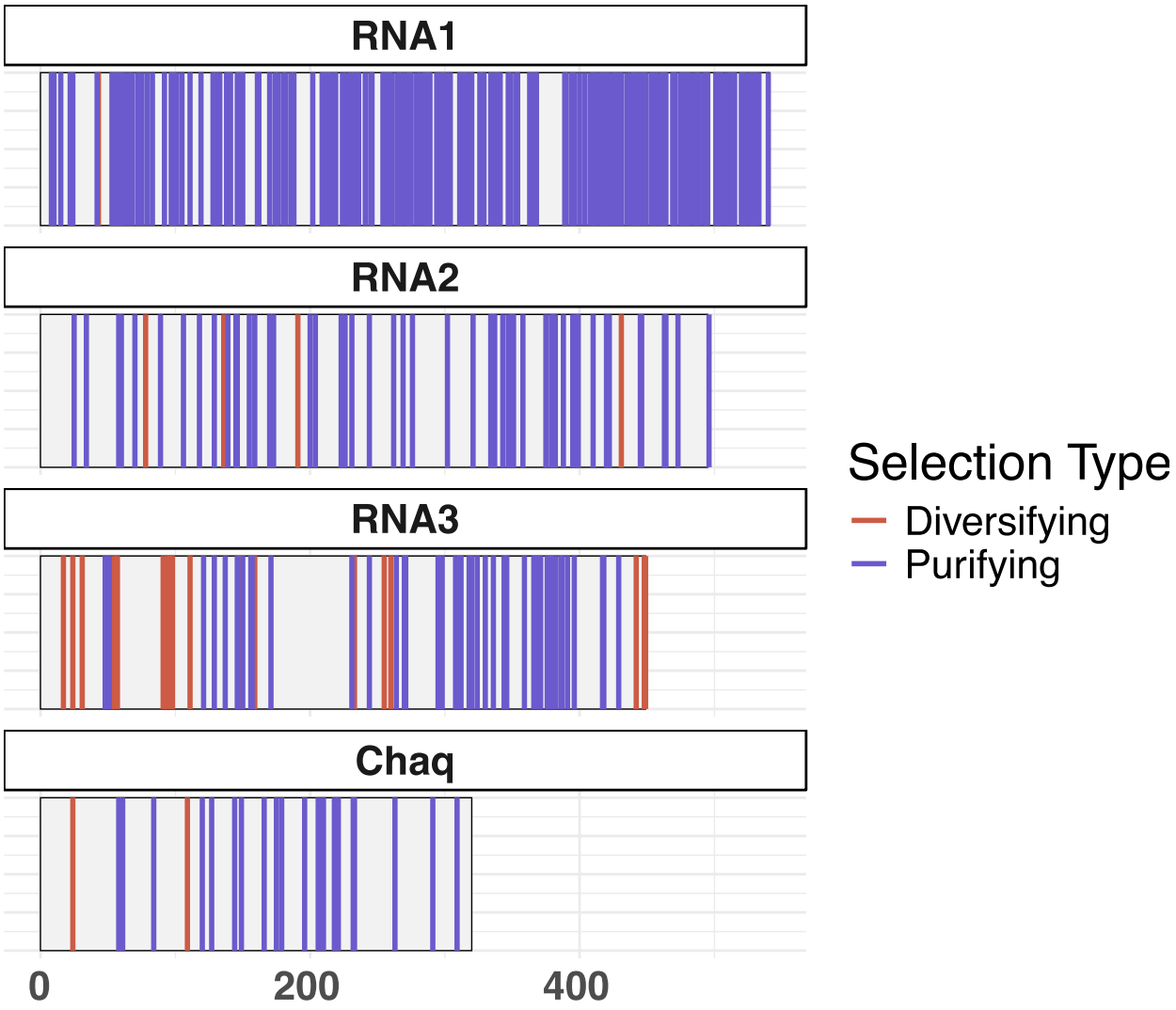


Figure 9: Galbut virus RNA 3 sequences exhibit more evidence of diversifying selection. Location of diversifying (red) and purifying (purple) selection are shown for each galbut virus segment and chaq virus. An asymptotic p-value of 0.09 was used for evidence of selection.

4.3.5 Galbut virus genotype A replicates to higher levels than genotype B regardless of chaq virus infection

Previous observations from our lab suggested that galbut virus genotype A replicated to higher levels than genotype B. We corroborated this observation using galbut virus-mapping read counts normalized to total read counts (galbut virus mapping

reads per million; RPM). To remove the potentially confounding effect of co-infection, we focused this analysis on non-co-infected samples (**Fig. 3**). Due to the multimodal distributions of infection levels observed in galbut virus and chaq virus, the median was used to calculate the fold difference in galbut A and B genotype infection levels (**Fig. 1C & Supp. Fig. 5**). Galbut virus genotype A median RPM values were 5.1 times higher than in genotype B infections (**Fig. 10A**).

Galbut virus genotype A infections generally had high numbers of reads. Galbut virus genotype B infections were generally low to mid positives but could replicate to higher levels. Five clade A samples had fewer than 100 RPM, lower than any clade B sample. Investigation of these samples did not reveal a simple explanation as to why they received fewer galbut virus reads compared to other clade A samples.

We next evaluated whether the presence of chaq virus increased levels of galbut virus RNA. There was no difference in median galbut virus RPM levels between infections with or without chaq virus (**Fig. 10B**; p-value = 0.24). Samples with chaq virus exhibited a much broader distribution of galbut virus RNA levels (Levene's test, p = 0.15). Samples negative for chaq were all clustered together in the middle of the range of chaq virus positive samples (**Fig. 10B**).

Figure 10 and our RT-qPCR data (**Fig. 1**) indicated that clade A galbut virus RNA levels exhibited a bimodal distribution. To investigate this, we plotted a histogram of RPM faceted by segment including partial sequences. There was a bimodal distribution in galbut virus RNA 1 and 2 levels and a trimodal distribution in galbut virus RNA 3 and chaq virus levels. This agrees with the distributions seen in **Figures 1** and **10** (**Supp. Fig. 5**).

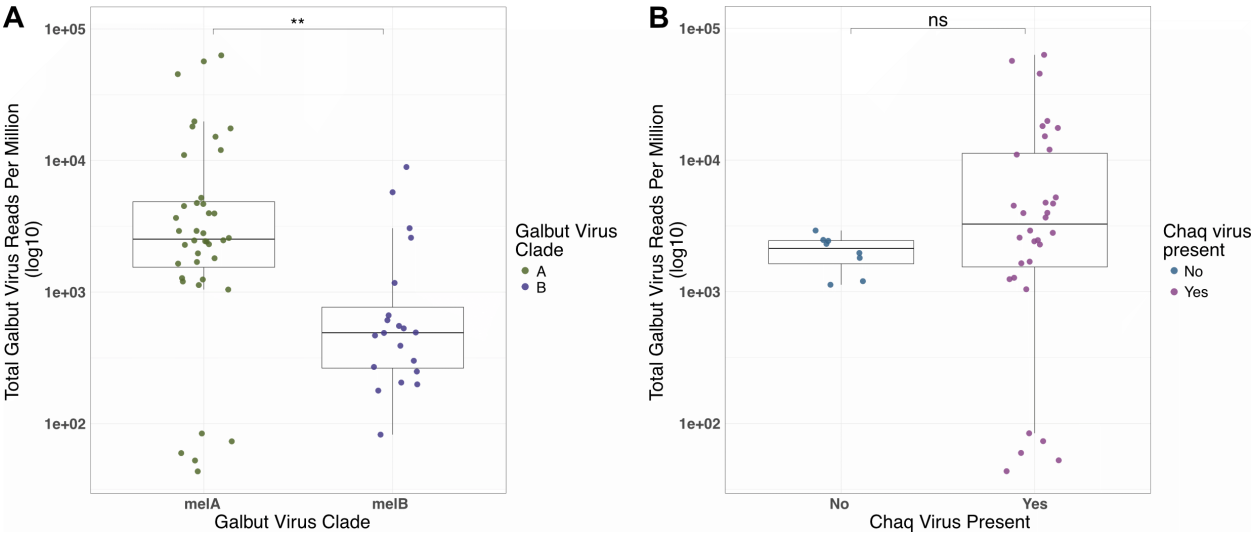


Figure 10: Galbut virus genotype A replicates to higher levels than genotype B. (A) Boxplot of total galbut virus reads (per million) for clades A and B. (B) Boxplot of total galbut virus reads (per million) and the presence of chaq virus. Samples with evidence of coinfection and for which clade could not be determined were omitted.

4.3.6 Galbut virus reassortment occurs within but not between genotypes

Reassortment which is the shuffling of genome segments following co-infection is typical for viruses with segmented genomes[21]. The short branches in galbut virus trees and high frequency of co-infection complicated detection of reassortment. To mitigate the potentially confounding impact of co-infection, we excluded samples with evidence of coinfection from the following reassortment analyses.

There were many well-supported examples of galbut virus reassortment (**Figs. 11, 12 & 13**). Reassortment involved parental viruses from the same clade. For example, sample 500-M-61-2, collected in 2023 from Colorado (Accessions: PQ625159, PQ625069, PQ624966, PQ624857), was genotype A but RNAs 1 and 2 were on differently clustering branches (**Figs. 11, 12 & 13**). Another example of within clade diversity was galbut virus RNA 2 of 500-F-67, collected in Colorado in 2023,

(Accessions: PQ625063, PQ624960, PQ624850) which was most similar to Pennsylvania sequences while the RNA 3 was more similar to samples from Maine and Colorado (**Figs. 11, 12 & 13**).

There was no evidence of reassortment between the major genotypes of galbut virus (**Figs. 11, 12 & 13**). This includes no evidence of reassortment between galbut viruses from *D. melanogaster* or *D. simulans*, and no evidence of reassortment between the *D. melanogaster* clade A and clade B viruses. We observed evidence of individual *D. melanogaster* samples co-infected by clade A and clade B viruses, which would provide an opportunity for reassortment to occur if it were possible.

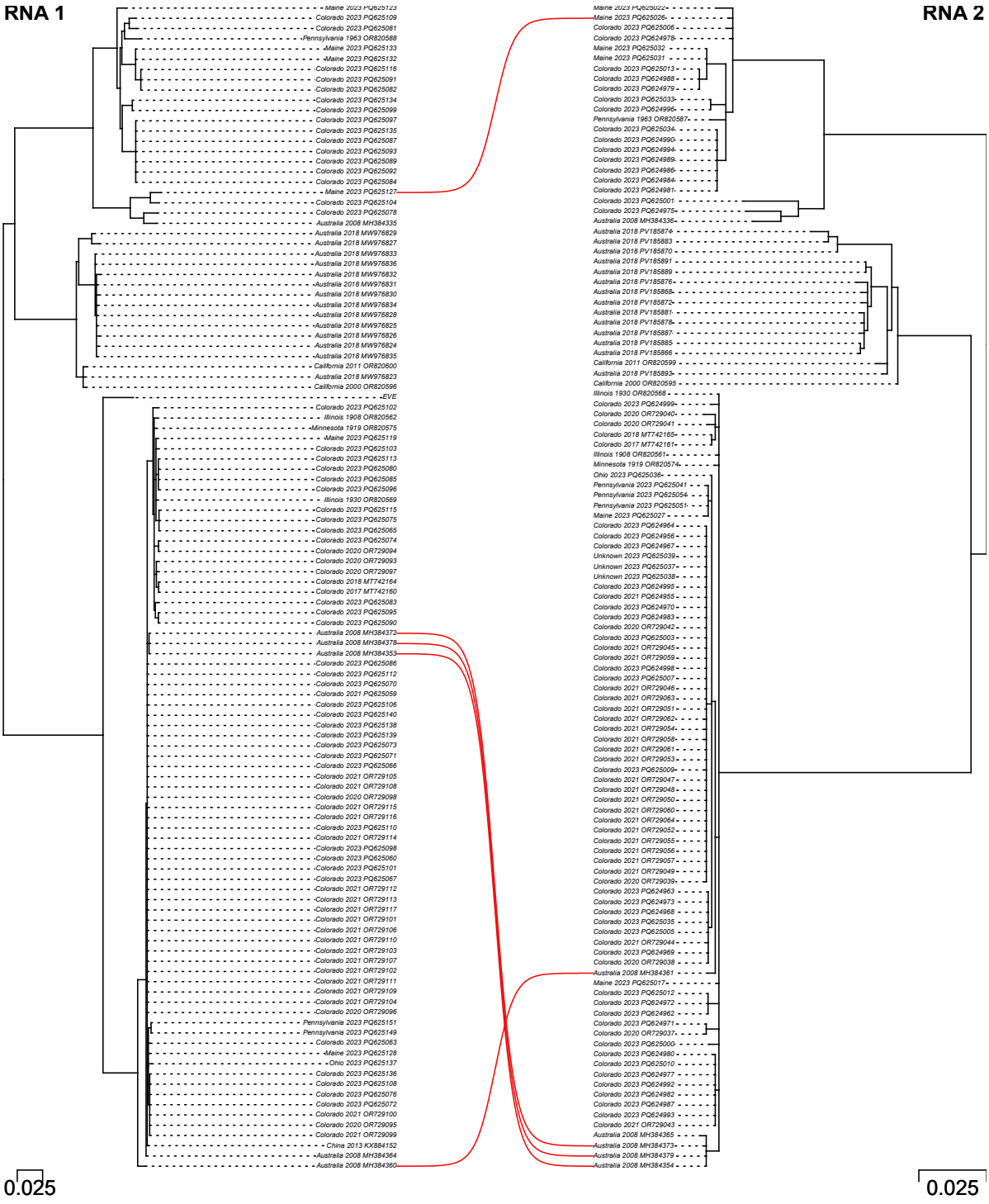


Figure 11: Co-phylogeny of galbut virus RNA1 and RNA 2. Branches with support values < 0.001 were converted to polytomies prior to generation of the phylogeny. Trees were midpoint rooted. Selected examples of reassortment are indicated by red colored connected lines.

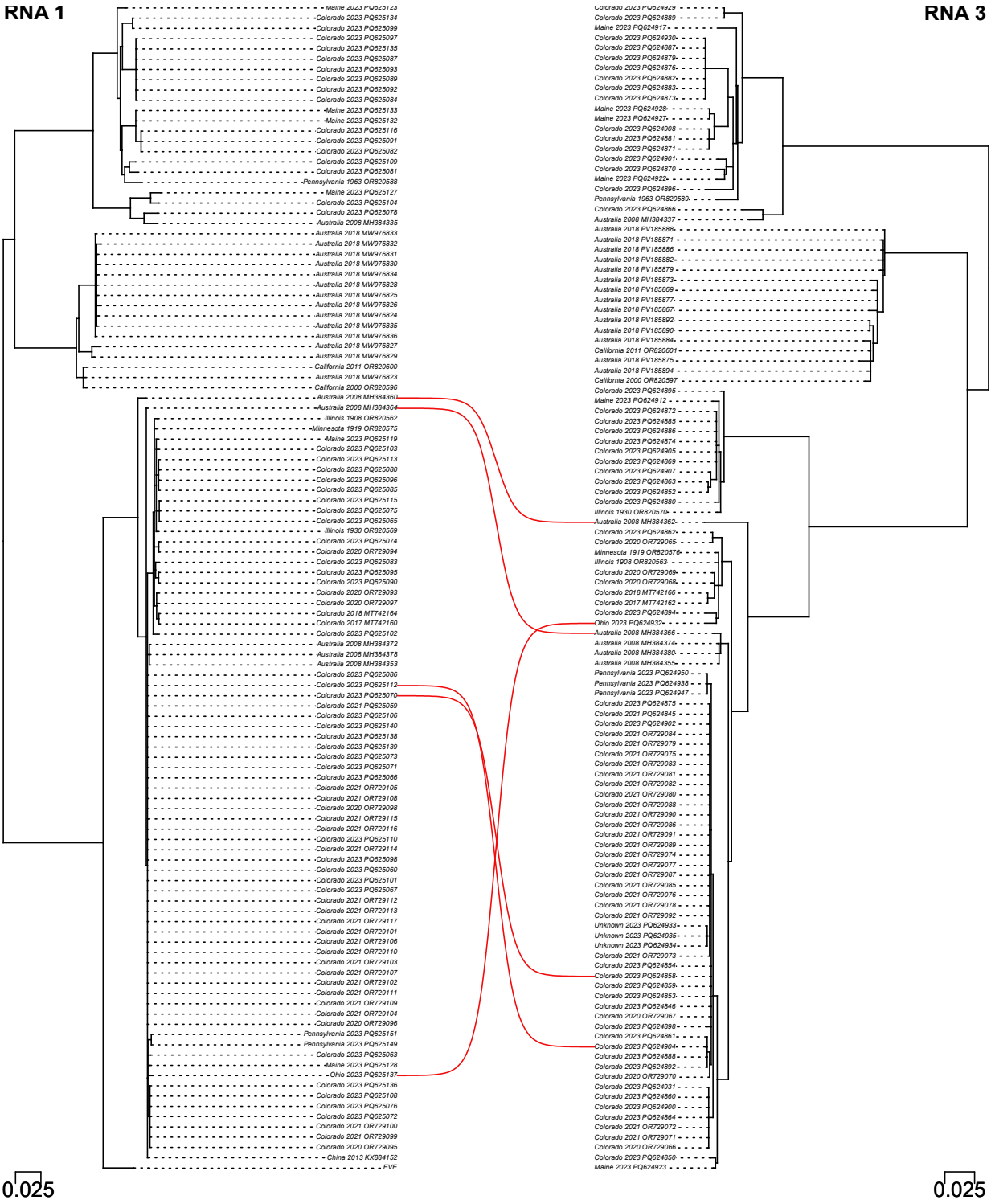


Figure 12: Co-phylogeny of galbut virus RNA1 and RNA 3. Branches with support values < 0.001 were converted to polytomies prior to generation of the phylogeny. Trees were midpoint rooted. Selected examples of reassortment are indicated by red colored connected lines.

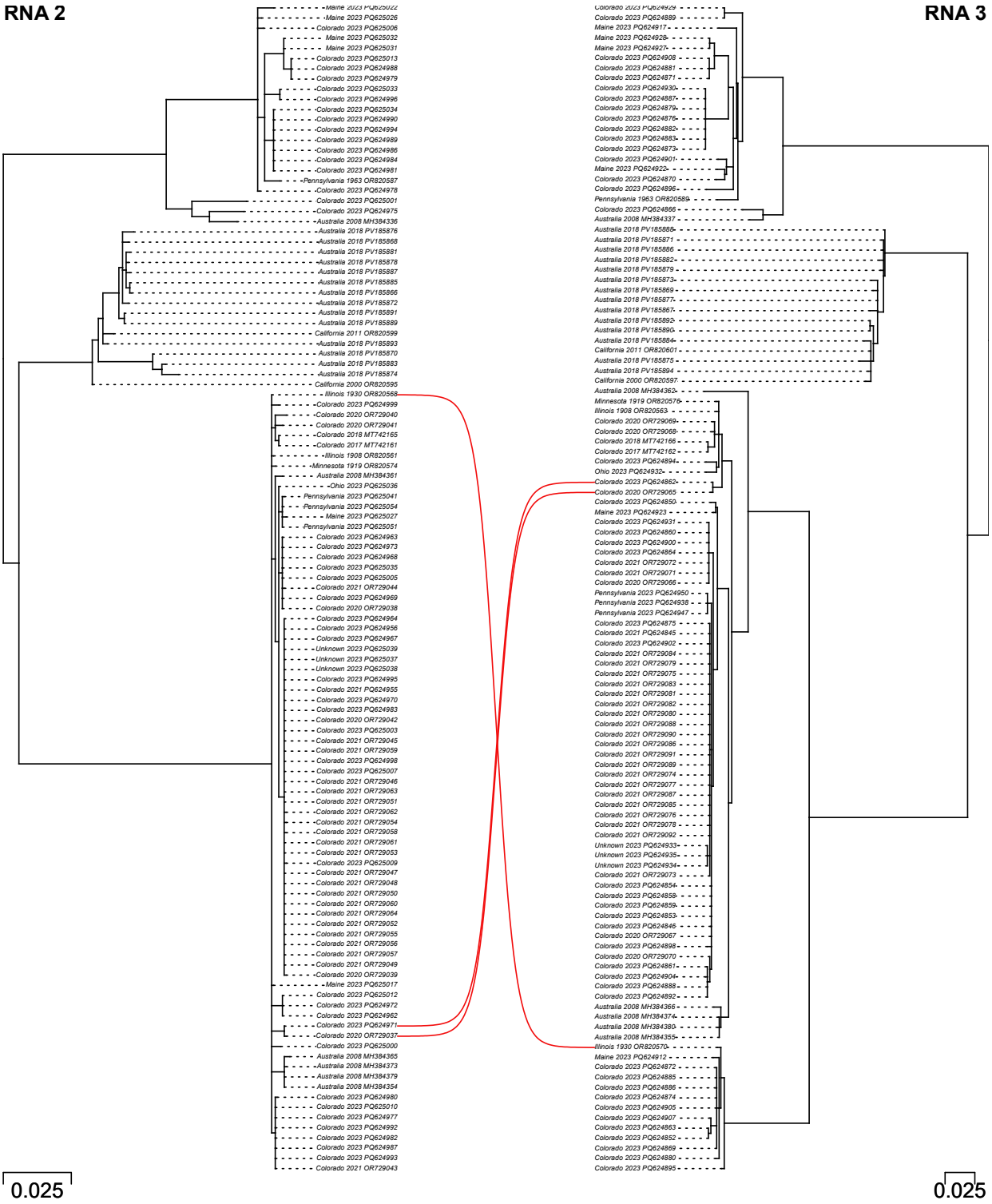


Figure 13: Co-phylogeny of galbut virus RNA 2 and RNA 3. Branches with support values < 0.001 were converted to polytomies prior to generation of the phylogeny. Trees were midpoint rooted. Selected examples of reassortment are indicated by red colored connected lines.

4.3.7 Chaq virus is associated exclusively with galbut virus clade A infection

Chaq virus was exclusively associated with galbut virus clade A infections. There were four instances of flies infected by clade B galbut virus that yielded chaq virus sequences, but in all of these instances, there were co-infecting clade A segments. Two instances were samples with complete galbut virus coinfections of genotypes A and B (ME-M-7 & Penn-F-4; **Fig. 14, 15 & 16**). The third instance was a Pennsylvania sample (Penn-F-6) with two RNA 2 sequences one of each genotype but no RNA 1 or 3 coding complete sequences (**Fig. 15**). The last instance was ME-F-1 which had coding-complete genotype B galbut virus sequences but only partial sequences for a clade A RNA 2 and two partial RNA 3 sequences. Given the coinfection status of all these samples, we concluded that chaq virus is only associated with genotype A infections (**Fig. 14, 15 & 16**).

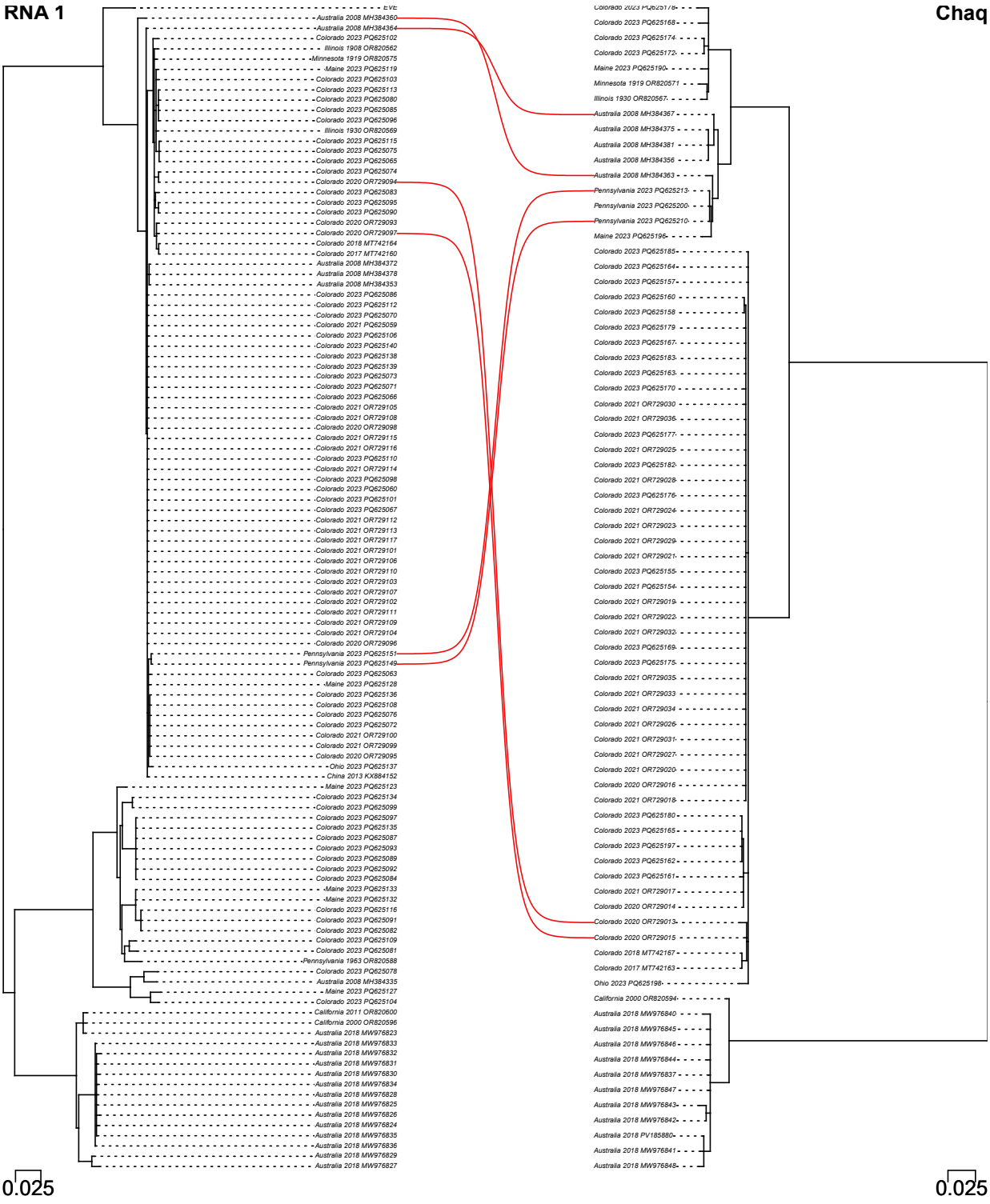


Figure 14: Co-phylogeny of galbut virus RNA1 and chaq virus. Branches with support values < 0.001 were converted to polytomies prior to generation of the phylogeny. Trees were midpoint rooted. Selected examples of reassortment are indicated by red colored connected lines.

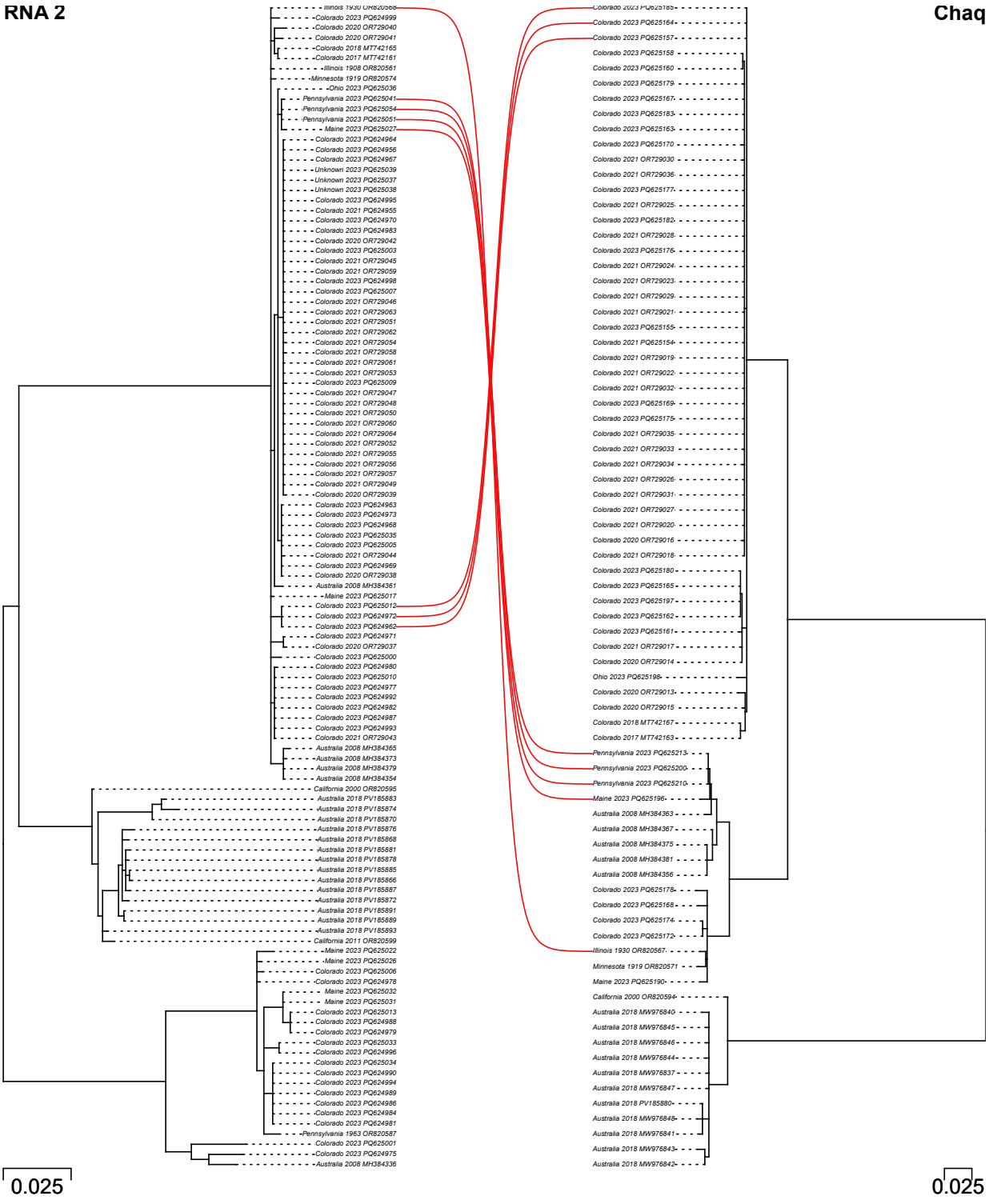


Figure 15: Co-phylogeny of galbut virus RNA 2 and chaq virus. Branches with support values < 0.001 were converted to polytomies prior to generation of the phylogeny. Trees were midpoint rooted. Selected examples of reassortment are indicated by red colored connected lines.

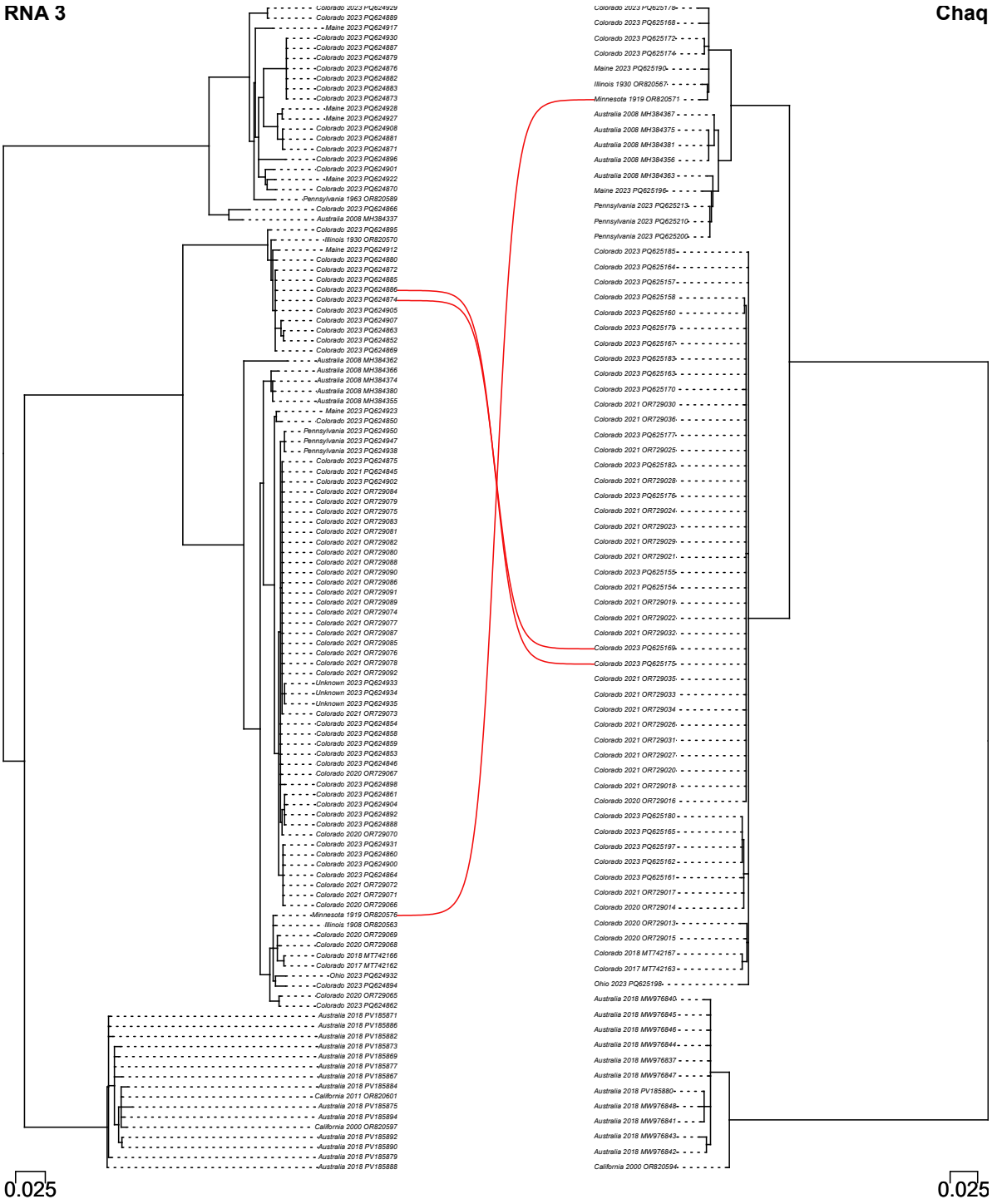


Figure 16: Co-phylogeny of galbut virus RNA 3 and chaq virus. Branches with support values < 0.001 were converted to polytomies prior to generation of the phylogeny. Trees were midpoint rooted. Selected examples of reassortment are indicated by red colored connected lines.

4.4 Discussion

In this manuscript we describe the genetic diversity and evolutionary patterns of galbut virus, a ubiquitous persistent viral infection of *D. melanogaster*[2]. We screened 974 flies from Colorado, Pennsylvania, Maine, and Ohio for galbut virus RNA using RT-qPCR. Galbut virus prevalence varied from 33% to 100% across locations (**Fig. 1**). Prevalence varied as a function of time and location, but no consistent patterns to explain the fluctuating prevalence were evident (**Fig. 1**). Longitudinal sampling of more locations could shed light on the factors that determine galbut virus prevalence.

One hundred and fifty-five samples were selected for recovery of galbut virus sequences. We aimed to sample a range of galbut virus diversity. One-hundred and one samples yielded coding complete or near complete galbut virus sequences that in combination with other available sequences were used to investigate galbut virus diversity and evolution.

We generated maximum likelihood phylogenetic trees to explore galbut virus sequence diversity and how they relate to observed phenotypes. The overall structure of our trees were similar to our previous study[11]. With the addition of more sequences, we discovered that some clades further subdivided into subclades (**Figs. 4, 5, 6 & 7 and Supp. Figs. 1, 2, 3, & 4**). We found that within clades, galbut virus sequence diversity was relatively low with some sequences sharing 100% pairwise identity. However, between clades, sequences were more diverse with pairwise identity reaching just 76% between genotypes A and *D. simulans*. Finally, our trees supported the observation that clade B infections were not associated with chaq virus. The only samples that had galbut virus clade B sequences and chaq virus sequences were also co-infected with clade A galbut virus (**Figs. 14, 15 & 16**).

How a host jump with a vertically transmitted virus would occur is unclear. Our trees indicated that the galbut virus *D. simulans* clade derived from clade A for RNAs 2, 3 and chaq virus but clade B for RNA 1. Although this could be an artifact of rooting, it is possible that the *D. simulans* clade is the result of reassortment between the two clades. Perhaps there is more successful inbreeding in the wild than previously thought[22]. An effort to capture and sequence diverse *Drosophila* species from many regions would expand our knowledge of galbut virus evolution across species.

There is evidence of geographically structured micro-lineages in our trees that reflect shared ancestry. For example, all clade A Pennsylvania sequences were highly similar. Samples from different locations were also found throughout the phylogenies. Maine sequences showed less evidence of geographic isolation as these sequences were dispersed throughout the phylogenies. Similarly, the *D. melanogaster* sequences from Australia were not isolated. These sequences were generally clustered together, but several sequences were more similar to sequences from the USA (**Figs. 4, 5, 6 & 7**). This could reflect the ease with which *D. melanogaster* moves over large geographic areas as a human commensal[23]. It would be interesting to look at the genetic similarity of the host flies. Maybe the flies with highly similar galbut virus sequences were more genetically similar to each other than those that were more similar to American sequences.

When we investigated the patterns of diversifying or purifying selection few loci were identified in galbut virus RNA 1, RNA 2 and chaq virus to be under diversifying selection. RNA 3 however had 20 sites of diversifying selection (**Table 2**). The higher number of diversifying selection sites in RNA 3 could be explained by determining its

role. Perhaps RNA 3 is involved in the RNA interference response and host immune modualtion[24,25].

Our RT-qPCR data suggested that there were low and high infection phenotypes. When we investigated this with the sequencing data, we saw that genotype A infections did replicate to higher levels but that there were genotype A infections at low levels (**Fig. 10A**). Investigation into these low-level infections did not reveal any easily defined characteristics that could explain this phenotype. When we specifically asked whether chaq virus infection could explain the low- and high- infection levels in genotype A we found that chaq virus presence did not impact galbut virus replication level (**Fig. 10B**). All clade A samples that did not have chaq virus exhibited a mid-level infection phenotype similar to clade B infection levels. Perhaps chaq virus is a satellite virus that plays a role in modulating infection levels of galbut virus[26,27].

We found evidence of galbut virus coinfection in 33% of flies. Only two of these cases corresponded to simple co-infection by 2 viruses (2 sets of 3 segments). Instead, coinfections were generally based on segment. RNA 3 was the most common segment to be present as more than one distinct sequence. It is possible that maintaining multiple RNA 3 genotypes in a single infection confers some kind of advantage, perhaps related to counteracting immune factors. Alternatively, molecular details of how the different RNA segments are replicated or transmitted may contribute to preferential persistence of RNA 3. RNA 2 and chaq virus were the next most common. Multiple RNA 1 segments were only found in complete coinfections (**Fig. 3**).

We identified evidence that galbut virus, like other segmented viruses reassorts[21,28]. In all cases of potential reassortment between genotypes, there was

coinfection of both clades (**Figs. 11, 12, 13, 14, 15 & 16**). Infections like this do suggest that replication of segments does not require all sequences to be from the same genotype. There was also evidence of within genotype reassortment. RNA 1 and 2 of sample 500-M-61-2 (Accessions: PQ625159, PQ625069, PQ624966, PQ624857) and RNA 1 and 3 of MEL191G (Accession: MH384360, MH384361, MH384362, MH384363) from Australia, are on divergent branches within clade A but the RNA 3 and RNA 2 sequences are similar to other RNA 3 and 2 sequences respectively. When we looked for evidence of recombination among the segments, the two probable recombination events could be explained by natural nucleotide variation or low-level coinfection.

Here we investigate the genotype-phenotype associations of galbut virus using sequences collected across large geographic areas. Galbut virus infection levels vary widely. This variation could be the result of galbut virus genotype and the presence/absence of chaq virus. We show that it is common for galbut virus genotypes to cocirculate within the same populations and that coinfections of the same or different clades are possible. We also find that reassortment within clades but not between clades is also common. Studies that bridge the biology and genetics of persistently infecting viruses like galbut virus will benefit from subsequential wet lab experiments.

4.5 Materials & Methods

4.5.1 Sample collection

Nine hundred and seventy-four flies were collected over the summer of 2023 from Maine, Pennsylvania, Ohio, and Colorado, USA. Flies collected in Colorado were trapped in buckets baited with banana and yeast with organdy mesh (Oriole Textile, 2060) on top of the fermenting fruit. Flies captured in Maine, Pennsylvania and Ohio

were trapped using 3D printed fly traps[14]. A subset of flies collected in Colorado were isolated from California-grown peaches bought at a grocery store in Fort Collins, CO. It is unclear whether the flies arrived with the peaches or were local. We assigned these flies as being from Colorado, corresponding to where they were collected.

4.5.2 RNA extraction

RNA was extracted from individual flies using a magnetic bead-based RNA/DNA capture method modified for use on the KingFisher (ThermoFisher Scientific). Briefly, flies were placed in an assay block (ThermoFisher Scientific, 9761116) with a small BB (McMaster Carr, 1598K22) and 100µL lysis buffer (20mM DTT (added fresh), 5M guanidine thiocyanate, 0.1M Tris-HCl (pH 7.5), 0.01M Na_2EDTA (pH8)). Flies were homogenized at 30 Hz for three minutes in a TissueLyszer II (Qiagen, 20.747.0001). The homogenate was moved to a new deep well plate (Thermofisher, 9504050) and combined with a master mix that included: 90µL SpeedBeads prepared according to manufacturer recommendations (Sigma, GE65152105050250), 10µL lysis binding enhancer (Proteinase K 200 µg/mL, 20% glycerol, 0.5% SDS) and 60µL 100% isopropanol. In separate 96 well standard plates (ThermoFisher, 97002540), 150µL of wash buffer one (20% Ethanol, 900 mM guanidine thiocyanate, 10 mM Tris-HCl (pH 7.5)), wash buffer two (200mM TE, 80% Ethanol) and 50µL of sterile water was added. The extraction was completed on a KingFisher liquid handling robot (Thermo Scientific). Eluted nucleic acids were then stored at -80°C until further processing.

4.5.3 Galbut virus positivity

To determine which samples were positive for galbut virus RNA, cDNA was generated using 5.5µL RNA, 0.75µM random hexamer primers (IDT), 0.75µM dNTPs and water to 13µL and incubated at 65°C for 5 minutes. Next, 4µL 5X FS buffer, 10M DTT and 1µL mashup reverse transcriptase was added and incubated at 50°C for 60 minutes followed by a heat inactivation step at 80°C for 10 minutes after which each sample was diluted in 90µL nuclease-free water. qPCR was conducted using NEB Luna Universal qPCR mix following manufacturer recommendations and cycling conditions (NEB, M3003). Each sample was screened for galbut virus using two primer sets (**Table 3**). Galbut virus 1 primers only detect clade A sequences. The discovery of galbut virus clade B led us to develop the galbut virus 2 set that detects both clade A and B. All wild collected samples were screened using both primers however the data shown in Figure 1 only uses data from the galbut virus 2 primers.

Table 3: Primer sequences used in this study

Primer pair name (#s in lab collection)	Galbut virus clade detection	Forward	Reverse	Reference
Galbut virus 1 (#1600/1601)	A	CCGTGAAGCAAG GAATCAAT	TGCCGATTTTCTG CTCTTTT	Cross 2020[1]
Galbut virus 2 (#2165/2170)	A and B	GTTCTTATCCGCA TCGCCTA	AATTCACTACTTC CTTGACACCTC	This study

RpL32 mRNA (1012/1013)	NA	TGCTAAGCTGTCG CACAAATGG	TGCGCTTGTTCGA TCCGTAAC	Cao 2016[29]

4.5.4 NGS library preparation

From the 641 galbut virus positive samples, 146 were selected for total RNA sequencing to recover galbut virus sequences. With the addition of the galbut virus 2 primers, we discovered additional clade B positive samples collected in 2020 and 2021. Nine of these samples were also sequenced. Samples were selected to represent diverse locations, galbut virus RNA levels (ranging from low to high), and qPCR melting temperature (ranging from identical to the positive control to 1˚C away). Sample concentration was obtained using the Qubit high-sensitivity RNA reagent (ThermoFisher, Q32852) and DNA was removed via a DNAse I treatment (NEB B0303S). Prior to library preparation, samples were normalized to 5 ng/μL except where sample concentrations were lower than 5 ng/μL. In these cases, either a different sample was selected, or the maximum volume of input RNA was used. Samples were split into two batches of ≤ 96 samples and prepared for sequencing using the Kapa RNA Hyper Prep kit following manufacturer recommendations (Roche, 08098093702). Eight cycles of library amplification were performed. Adapters contained unique dual indices (Integrated DNA Technologies xGen UDI/UMI 10005903). HeLa cell total RNA was used as a positive control and water was used as a negative control for both library preparations and sequencing. Sample pooling was determined using High Sensitivity DNA Qubit reagents and final library quality control was done using an Agilent D1000 HS Tapestation and KapaQuant reagents following manufacturers recommendations

(Roche, 07960140001). Both pools were sent to Azenta/Genewiz where they were sequenced on a NovaSeq 6000 with 2X150 paired end sequencing.

4.5.5 Data analysis

Identification of galbut virus sequences: We used our lab's previously described virus taxonomy pipeline (https://github.com/stenglein-lab/taxonomy_pipeline) to identify and validate galbut virus sequences from our NGS datasets[11]. In brief, adapters and low-quality reads were trimmed using cutadapt 3.5[30]. Fastqc 0.11.9[31] was used to assess post-collapsed read quality. Host reads were removed using bowtie2 2.4.5[32] and the remaining reads were assembled using spades 3.15.4[33]. Virus sequences were identified using BLASTn 2.12.0. Draft sequences were validated by remapping of trimmed reads using bwa mem aligner version 0.7.17[34]. Final sequences were submitted to the NCBI nucleotide database and raw NGS data to the NCBI sequence read archive repository (see Data Availability statement below).

Molecular species identification: To confirm species assignment of samples, we used our lab's species identification pipeline (https://github.com/stenglein-lab/species_id)[11]. In brief, this pipeline takes trimmed reads and competitively maps them to a set of 529 cytochrome c oxidase subunit 1 sequences from across the Drosophilidae family. This pipeline identified one *Drosophila affinis*, one *Drosophila simulans* and 147 *D. melanogaster*. Six samples remained unidentified.

Maximum Likelihood trees: All available galbut virus and chaq virus sequences with a known location, date of collection and were from individual flies were downloaded from GenBank. Alignments were generated using MAFFTT v7.490 with default parameters[35]. IQ-TREE v2.3.6 was used to generate maximum likelihood trees[36]. IQ-TREE

parameters were: singularity exec iqtree:2.3.6--h503566f_1 iqtree2 -s $fasta --threads-max $num_threads -T AUTO --mem 33% -B 1000 -alrt 1000. iTol was used for tree visualization[37]. Trees were midpoint rooted. Bootstrap support values at large branches were retained.

Selection analysis: We used the Fixed Effects Likelihood model described by Kosakovsk-Pond and Frost[20]. This analysis was run on Datamonkey with the HyPhy application[38,39]. We explored the FUBAR method (fast, unconstrained Bayesian approximation) and had similar results for galbut virus RNAs 1 and 2 and chaq virus however considerably less sites of positive selection were identified in RNA 3[40]. Given that the RNA 3 alignment had sequence similarities between -76% and 100%, the results from the FEL method were more appropriate.

Recombinant detection analysis: We used the Recombination Detection Program[41,42] to evaluate alignments for evidence of recombination. This tool identified one candidate recombination event in chaq virus and another in galbut virus RNA 3. We visualized the pairwise percent identity between the candidate recombination events. This visualization led us to conclude that the chaq virus recombinant was the result of a chimera of two partial chaq virus sequences. We were unable to recover a coding complete sequence of these sequences. Inspection of pairwise alignments of the potential RNA 3 recombination event was not convincing. The similarity between the recombination regions could be explained by natural variations in similarity.

Cophylogeny analysis: Tanglegrams were generated in R/RStudio using the phytools and ape packages[43,44]. Branches shorter than 0.001 were turned into polytomies to avoid false positive signals of reassortment from short branches.

4.6 Acknowledgements

The authors would like to acknowledge the individuals that volunteered to collect wild flies: Lisa Harrington and Paul Keene, Khwannarin Khemsom, Anastazia Jablumovsky, Elle Irons and Michael Martzahn, Kelly Wallace and Ben Ryan, Claire Stewart and Natali Wickencamp and the packaging department at New Belgium Brewing Company. We thank them for their effort in collecting flies themselves or for the opportunity for us to collect flies from their property. We would also like to acknowledge Eddie Holmes, Ayda Susana Ortiz-Baez, Mang Shi, Ary A Hoffmann for their work in Australian D. simulans from which their SRA datasets were used to recover galbut virus sequences.

4.7 Data availability

Sequencing data has been deposited in the National Center for Biotechnology and Information Sequence Read Archive (SRA) database under accessions SRR30712177-SRR30712331 and in the Nucleotide database under accessions PQ624841-PQ625194, PQ625196-PQ625214, and PV185866-PV185894. The scripts used to analyze the data, raw RT-qPCR data statistical analyses and metadata can be found at this GitHub repository: https://github.com/LKeene/DiverseCollections.

4.8 Funding

NSF IOS 2048214 (AHK, ALB and MDS), NIH T32GM132057 (AHK), and NIH/NCATS Colorado CTSA Grant Number UL1 TR002535

CHAPTER 4: REFERENCES

1. Cross, S. T. *et al.* Partitiviruses Infecting Drosophila melanogaster and Aedes aegypti Exhibit Efficient Biparental Vertical Transmission. *J Virol* **94**, (2020).
2. Webster, C. L. *et al.* The Discovery, Distribution, and Evolution of Viruses Associated with Drosophila melanogaster. *PLoS Biol* **13**, e1002210 (2015).
3. Ghabrial, S. A., Ochoa, W. F., Baker, T. S. & Nibert, M. L. Partitiviruses: General Features. in *Encyclopedia of Virology (Third Edition)* (eds. Mahy, B. W. J. & Van Regenmortel, M. H. V) 68–75 (Academic Press, Oxford, 2008). doi:https://doi.org/10.1016/B978-012374410-4.00573-2.
4. Wallace, M. A. & Obbard, D. J. Naturally occurring viruses of *Drosophila* reduce offspring number and lifespan. *Proceedings of the Royal Society B: Biological Sciences* **291**, (2024).
5. Cross, S. T. *et al.* Galbut Virus Infection Minimally Influences Drosophila melanogaster Fitness Traits in a Strain and Sex-Dependent Manner. *Viruses* **15**, 539 (2023).
6. Mackay, T. F. C. *et al.* The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**, 173–8 (2012).
7. Keller, A. Drosophila melanogaster's history as a human commensal. *Current Biology* **17**, PR77-R81 (2007).
8. Ortiz-Baez, A. S., Shi, M., Hoffmann, A. A. & Holmes, E. C. RNA virome diversity and Wolbachia infection in individual Drosophila simulans flies. *J Gen Virol* **102**, 001639 (2021).
9. Medd, N. C. *et al.* The virome of Drosophila suzukii, an invasive pest of soft fruit. *Virus Evol* **4**, (2018).
10. Wallace, M. A. *et al.* The discovery, distribution, and diversity of DNA viruses associated with *Drosophila melanogaster* in Europe. *Virus Evol* **7**, veab031 (2021).
11. Keene, A. H. & Stenglein, M. D. Sequencing RNA from old, dried specimens reveals past viromes and properties of long-surviving RNA. Preprint at https://doi.org/10.1101/2024.10.03.616531 (2024).
12. Xiao, X. *et al.* A Novel Partitivirus That Confers Hypovirulence on Plant Pathogenic Fungi. *J Virol* **88**, 10120–10133 (2014).
13. Thapa, V. *et al.* Using a Novel Partitivirus in Pseudogymnoascus destructans to Understand the Epidemiology of White-Nose Syndrome. *PLoS Pathog* **12**, e1006076 (2016).
14. Keene-Snickers, A. H., Dunham, T. J. & Stenglein, M. D. Experimental assessment of 3D-printed traps and chemical attractants for the collection of wild Drosophila melanogaster. Preprint at https://doi.org/10.1101/2025.01.28.635319 (2025).
15. Albrecht, T. *et al.* Occurrence of Wheat Curl Mite and Mite-Vectored Viruses of Wheat in Colorado and Insights into the Wheat Virome. *Plant Dis* **106**, 2678–2688 (2022).
16. Kapuscinski, M. L. *et al.* Genomic characterization of 99 viruses from the bunyavirus families Nairoviridae, Peribunyaviridae, and Phenuiviridae, including 35 previously unsequenced viruses. *PLoS Pathog* **17**, e1009315 (2021).
17. Stenglein, M. D. *et al.* Widespread recombination, reassortment, and transmission of unbalanced compound viral genotypes in natural arenavirus infections. *PLoS Pathog* **11**, e1004900 (2015).

18. Shi, M. *et al.* No detectable effect of Wolbachia wMel on the prevalence and abundance of the RNA virome of Drosophila melanogaster. *Proceedings of the Royal Society B: Biological Sciences* **285**, (2018).
19. Botella, L., Vainio, E. J., Hantula, J., Diez, J. J. & Jankovsky, L. Description and prevalence of a putative novel mycovirus within the conifer pathogen Gremmeniella abietina. *Arch Virol* **160**, 1967–1975 (2015).
20. Kosakovsky Pond, S. L. & Frost, S. D. W. Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. *Mol Biol Evol* **22**, 1208–1222 (2005).
21. Lowen, A. C. It's in the mix: Reassortment of segmented viral genomes. *PLoS Pathogens* vol. 14 Preprint at https://doi.org/10.1371/journal.ppat.1007200 (2018).
22. CASTLE, W. E. INBREEDING, CROSS-BREEDING AND STERILITY IN DROSOPHILA. *Science (1979)* **23**, 153–153 (1906).
23. Arguello, J. R., Laurent, S. & Clark, A. G. Demographic History of the Human Commensal *Drosophila melanogaster*. *Genome Biol Evol* **11**, 844–854 (2019).
24. Swevers, L., Broeck, J. Vanden & Smagghe, G. The possible impact of persistent virus infection on the function of the RNAi machinery in insects: A hypothesis. *Front Physiol* **4 NOV**, (2013).
25. Karlikow, M., Goic, B. & Saleh, M. C. RNAi and antiviral defense in Drosophila: Setting up a systemic immune response. *Dev Comp Immunol* **42**, 85–92 (2014).
26. Qiu, W. & Karen-Beth, S. G. Defective Interfering RNAs of a Satellite Virus. *J Virol* **75**, 5429–5432 (2001).
27. Parks, W. P., Casazza, A. M., Alcott, J. & Melnick, J. L. ADENO-ASSOCIATED SATELLITE VIRUS INTERFERENCE WITH THE REPLICATION OF ITS HELPER ADENOVIRUS. *J Exp Med* **127**, 91–108 (1968).
28. McDonald, S. M., Nelson, M. I., Turner, P. E. & Patton, J. T. Reassortment in segmented RNA viruses: Mechanisms and outcomes. *Nature Reviews Microbiology* vol. 14 448–460 Preprint at https://doi.org/10.1038/nrmicro.2016.46 (2016).
29. Cao, C., Magwire, M. M., Bayer, F. & Jiggins, F. M. A Polymorphism in the Processing Body Component Ge-1 Controls Resistance to a Naturally Occurring Rhabdovirus in Drosophila. *PLoS Pathog* **12**, e1005387 (2016).
30. Martin, M. Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads. *EMBnet J* **17**, 10–12 (2011).
31. Andrews, S. *et al.* FastQC: a quality control tool for high throughput sequence data. https://qubeshub.org/resources/fastqc (2012).
32. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
33. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. *Curr Protoc Bioinformatics* **70**, e102 (2020).
34. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
35. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* **30**, 772–780 (2013).
36. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268–274 (2015).

37. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res* **52**, W78–W82 (2024).
38. Weaver, S. *et al.* Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. *Mol Biol Evol* **35**, 773–777 (2018).
39. Kosakovsky Pond, S. L. *et al.* HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Mol Biol Evol* **37**, 295–299 (2020).
40. Murrell, B. *et al.* FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. *Mol Biol Evol* **30**, 1196–1205 (2013).
41. Martin, D. & Rybicki, E. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562–563 (2000).
42. Martin, D. P. *et al.* RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol* **7**, (2021).
43. Revell, L. J. phytools: Phylogenetic Tools for Comparative Biology (and Other Things). *CRAN: Contributed Packages* Preprint at https://doi.org/10.32614/CRAN.package.phytools (2011).
44. Paradis, E. *et al.* ape: Analyses of Phylogenetics and Evolution. *CRAN: Contributed Packages* Preprint at https://doi.org/10.32614/CRAN.package.ape (2002).

CHAPTER 5: CONCLUSION AND FUTURE DIRECTIONS

My research spans time and space in an effort to study the evolutionary patterns of a common, persistent viral infection of *D. melanogaster*. While this dissertation was mainly focused on one virus and host, I believe the techniques I used here can be applied to study diverse viral species on a large timescale and across many geographic regions. In this chapter, I suggest future research directions based on my findings. I hope that these suggestions lead to an increased use of museum specimens as a source to study RNA. I also hope that they provide more evidence for the importance of leveraging citizen scientists to collect specimens. Finally, I suggest future research directions that address questions around the basic biology of galbut virus and how this knowledge in combination with my genomic data could lead to a better understanding of persistent viruses and their relationship with their host.

The study of old RNA has generally been limited to specimens that have been preserved chemically or stored frozen on purpose or by chance[1–3]. This is because RNA is believed to be unstable in the environment due to its chemical structure and the ubiquity of RNases[4]. In Chapter 2, I used dried insects stored in museums to recover viral and host RNA. Sequences derived from this RNA showed that RNA can persist in the absence of preservation for longer than expected (in some cases over 100 years!) and that this RNA could be used to study virus evolution.

In submitting this paper for peer review and publication there was resistance to the data due to the similarity between old and new virus sequences (some sequences were >99% identical to contemporary sequences). There is a strong belief that viruses,

particularly RNA ones, evolve incredibly fast and that my sequences could only be the result of contamination. This belief was held despite great effort on our part to provide evidence (strict laboratory protocols, inclusion of all positive and negative control data-RT-qPCR and sequencing and host species identification using nuclear and mitochondrial RNA markers) arguing against contamination from our lab being the source of these sequences. To address these concerns directly, I acquired more insect specimens (other random species of flies and beetles that nobody in our building works with), performed all wet bench work in an ancient DNA (aDNA) laboratory and sequenced them using a commercial sequencing company. This data corroborated our original findings that RNA and RNA virus sequences can be obtained from old insect specimens. This is true even in samples that are over 100 years old. I am currently revising the Chapter 2 paper to include this data.

In expanding my Chapter 2 work to include non-*Drosophila* insects, it became clear that more research needs to be done to determine what specimen types work with my methods. It is unclear how specimen size and insect species impact the successful isolation of RNA and subsequent library preparation. For instance, could this technique be successfully applied to large insects like grasshoppers or butterflies? The size of the insect likely plays a role in how quickly it dries out. During this time, it is possible that RNases remain active and continue to degrade the RNA resulting in lower RNA abundance over time. However, larger insects have substantially more RNA than small ones which could overcome the increased time of RNA degradation. It would also be interesting to explore whether these methods could be extended to mammals or reptiles stored in museum collections? If so, which tissues would be appropriate? Skin may be

the most stable because it is already comprised of dead tissue, but it is also exposed to more surface contamination. Organs likely have more RNA to begin with, but how does decomposition affect RNA survival? These are all questions that could be explored in the future.

Future work could also assess how specimen storage impacts RNA preservation. Of the 12 specimens I received from Hawai'i, only two were sequenced because the RNA concentrations were low, and library generation was not successful. These samples were collected from 1950-1965 and were therefore not the oldest samples I used for this work. I hypothesized that humidity may play more of a role in RNA stability than temperature as RNA from 1,000-year-old corn has been isolated from a region that experiences extreme temperature fluctuations but remains arid year-round[5]. Although most museum specimens are now stored in temperature- and humidity-controlled environments, knowing how environmental conditions impact RNA stability is worth investigating, particularly for work with very old specimens.

In Chapter 3, I developed two fly traps for the collection of wild *D. melanogaster*. These traps were then used to collect flies for the work described in Chapter 4. A benefit of using 3D printed components in these traps is that they can be customized to what consumables are available. By using different foods, these traps can also be used to collect different *Drosophila* species. For example, raspberry pulp mixed with sugar can be used to target *Drosophila suzukii*, an agricultural pest species[6]. It will be interesting to see how effective these traps are when different species are the target for collection.

Finally, the work in Chapter 4 explored the genotypic and phenotypic associations of different galbut virus strains. This work has further corroborated my

findings in Chapter 2 that galbut virus evolves slowly. The goal of collecting many flies from many places was to capture the genetic diversity of the virus to perform time-resolved phylogenetic analyses. Unfortunately, even with old sequences and sequences from across the U.S., there was not enough diversity to confidently perform these analyses. Galbut virus is simply not evolving fast enough. Perhaps this is due to interactions with the host that allow it to maintain itself as a vertically transmitted infection? This close association could certainly constrain the virus's ability to change in line with the prisoner of war hypothesis[7]. However, the exact biological mechanisms that lead to this constraint have yet to be understood.

In many ways the work presented in Chapter 4 has led to more questions about the overall biology of galbut virus than it has resolved. Galbut virus RNA 3 is the most diverse segment and is found as coinfections more often than the other segments. The questions around why this occurs could be better understood if we knew the role this segment plays. In my preliminary exam I hypothesized that RNA 3 interacts with the host RNA interference response to modulate the hosts immune response[8,9]. Work by an undergraduate mentee of mine is currently ongoing to determine if this is true. Similarly, the role of chaq virus is unknown however, it follows similar genetic patterns to galbut virus RNA 3. Preliminary work using AlphaFold to predict the structure of chaq virus by members of the Stenglein Lab suggested that it could be a capsid protein, which would suggest that chaq virus is a satellite of galbut virus. Further work needs to be done to confirm this finding.

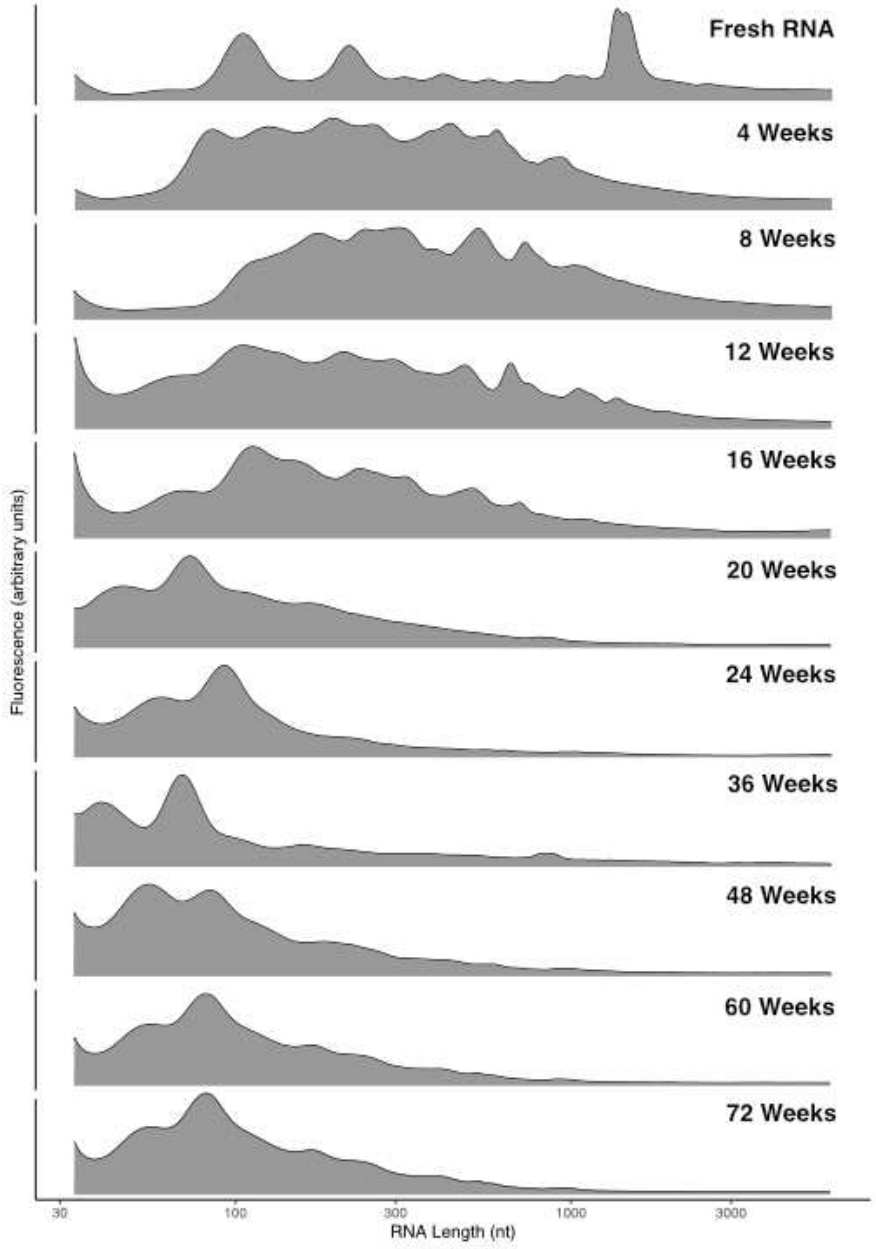
In this dissertation, I presented rationale for the study of persistent viral infections that would otherwise be ignored because they do not cause noticeable disease. Most

organisms are infected by multiple viruses at one time, but most virology effort is focused a few viruses that cause overt disease. These other viruses are still important to the host and the biological and evolutionary implications they have are worth researching. I am grateful for the support and guidance for taking on these projects the last five years. I look forward to seeing how this research continues to evolve.

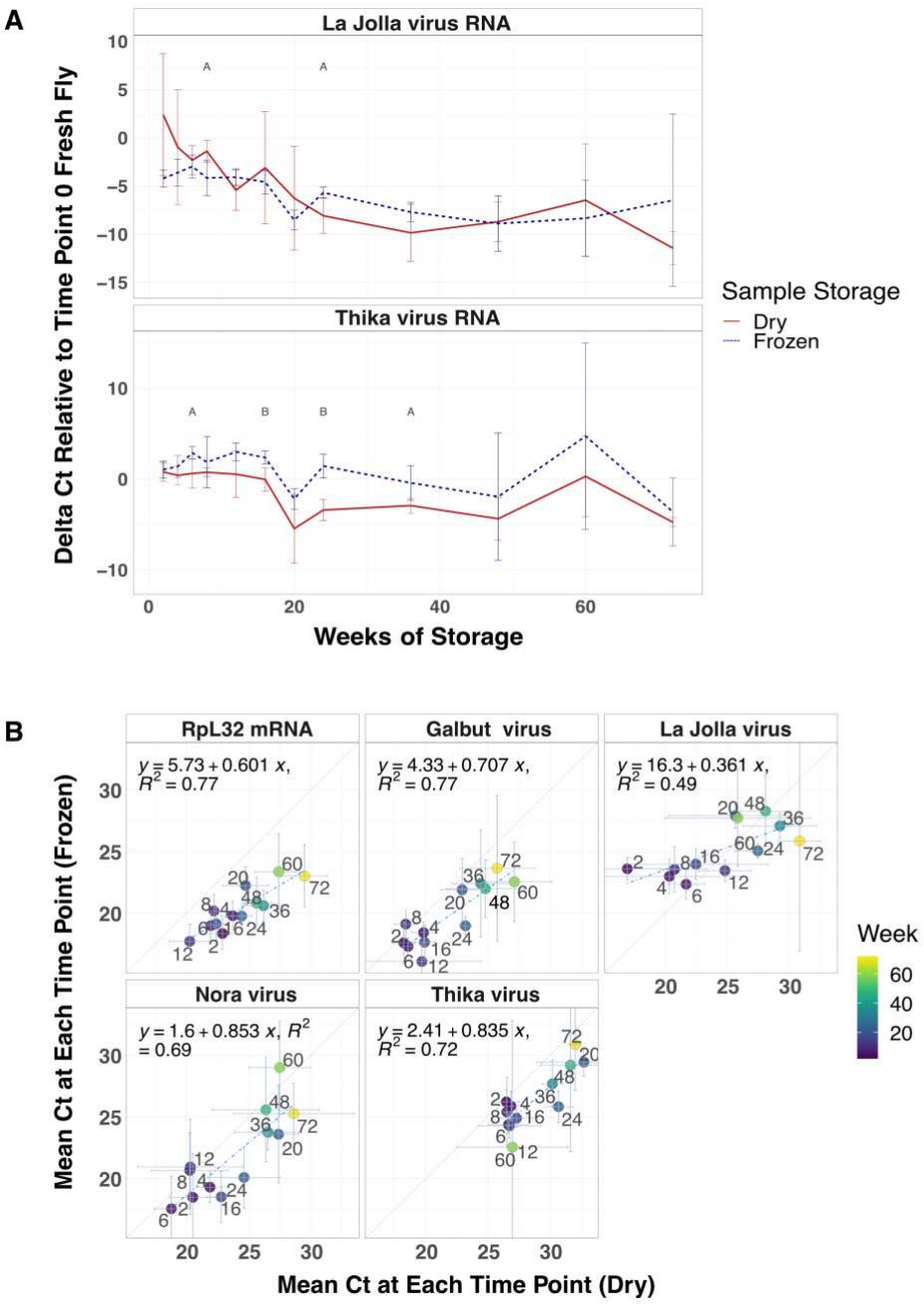
CHAPTER 5 REFERENCES

1. Faria, N. R. *et al.* The early spread and epidemic ignition of HIV-1 in human populations. *Science (1979)* **346**, 56–61 (2014).
2. Smith, A. W., Skilling, D. E., Castello, J. D. & Rogers, S. O. Ice as a reservoir for pathogenic human viruses: Specifically, caliciviruses, influenza viruses, and enteroviruses. *Med Hypotheses* **63**, 560–566 (2004).
3. Speer, K. A. *et al.* A comparative study of RNA yields from museum specimens, including an optimized protocol for extracting RNA from formalin-fixed specimens. *Front Ecol Evol* **10**, (2022).
4. Green, M. R. & Sambrook, J. How to Win the Battle with RNase. *Cold Spring Harb Protoc* **2019**, pdb.top101857 (2019).
5. Peyambari, M., Warner, S., Stoler, N., Rainer, D. & Roossinck, M. J. A 1,000-Year-Old RNA Virus. *J Virol* **93**, e01188-18 (2019).
6. Lasa, R., Toledo-Hernández, R. A., Rodríguez, D. & Williams, T. Raspberry as a Source for the Development of Drosophila suzukii Attractants: Laboratory and Commercial Polytunnel Trials. *Insects* **10**, 137 (2019).
7. Simmonds, P., Aiewsakun, P. & Katzourakis, A. Prisoners of war — host adaptation and its constraints on virus evolution. *Nat Rev Microbiol* **17**, 321–328 (2019).
8. van Cleef, K. W. R., van Mierlo, J. T., van den Beek, M. & van Rij, R. P. Identification of Viral Suppressors of RNAi by a Reporter Assay in Drosophila S2 Cell Culture. In *Antiviral RNAi: Concepts, Methods, and Applications* (ed. Van Rij, R. P.) 201–213 (Humana Press, Totowa, NJ, 2011). Doi:10.1007/978-1-61779-037-9_12.
9. Shimura, H., Kim, H., Matsuzawa, A., Akino, S. & Masuta, C. Coat protein of partitiviruses isolated from mycorrhizal fungi functions as an RNA silencing suppressor in plants and fungi. *Sci Rep* **12**, (2022).

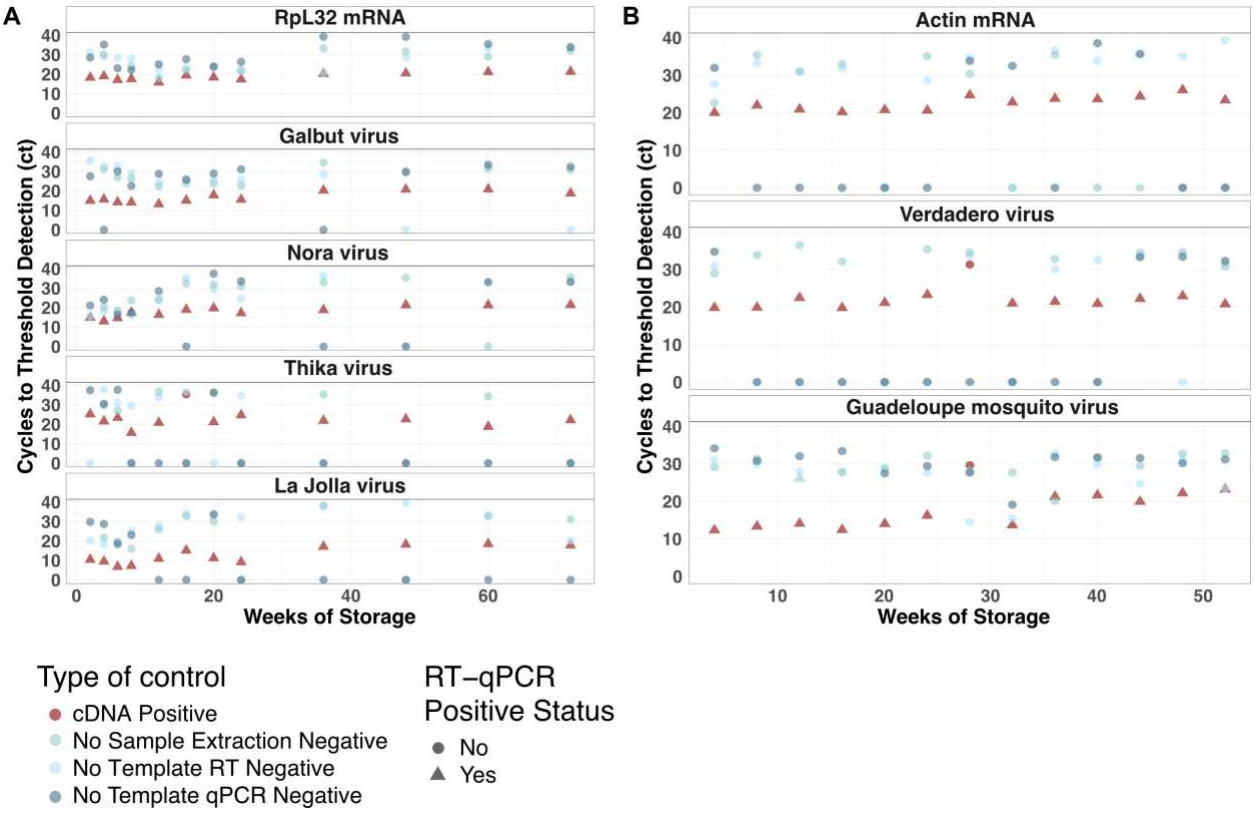
Appendix A: Chapter 2 Supplemental Figures



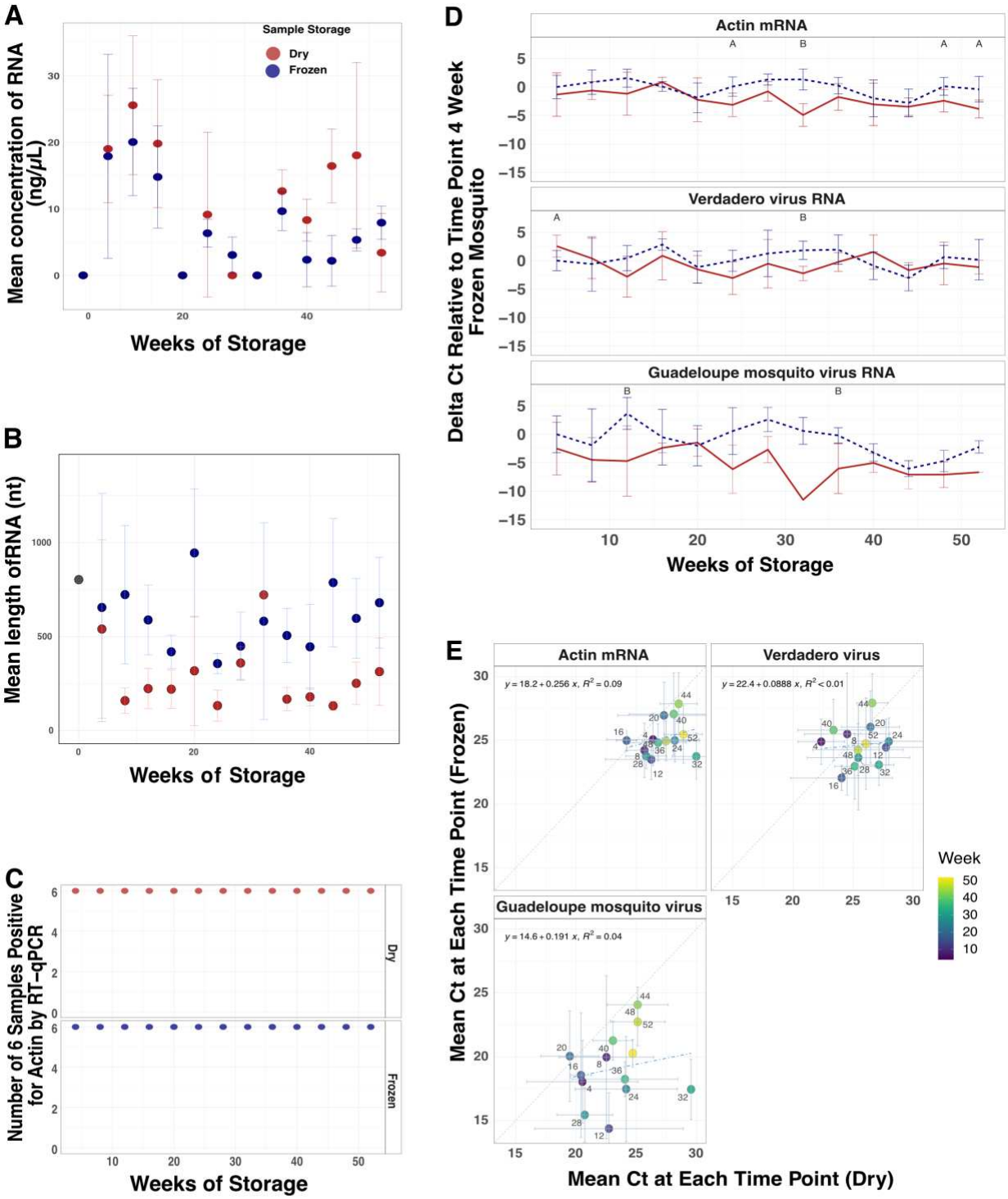
Supplemental Figure 1: RNA grew increasingly fragmented in experimentally dried flies. The RNA length of representative sample from each time point of dried *D. melanogaster* are shown. An Agilent Tapestation was used to measure RNA length distributions. Lengths <33 nt, corresponding to the Tapestation lower marker internal control, are not plotted.



Supplementary Figure 2: Viral RNA and host mRNA persisted in fly specimens over 72 weeks. (A) Average difference between La Jolla virus and Thika virus RNA levels relative to levels in fresh *D. melanogaster*. A two-tailed t-test was used to determine the statistical significance between dried and frozen means at every time point (A < 0.05, B < 0.001). (B) Comparison of mean Ct values at each time point for each RNA target. Linear regression lines are shown as is the diagonal. Error bars represent mean plus and minus the standard deviation. N = 3 male and 3 female at each time point.

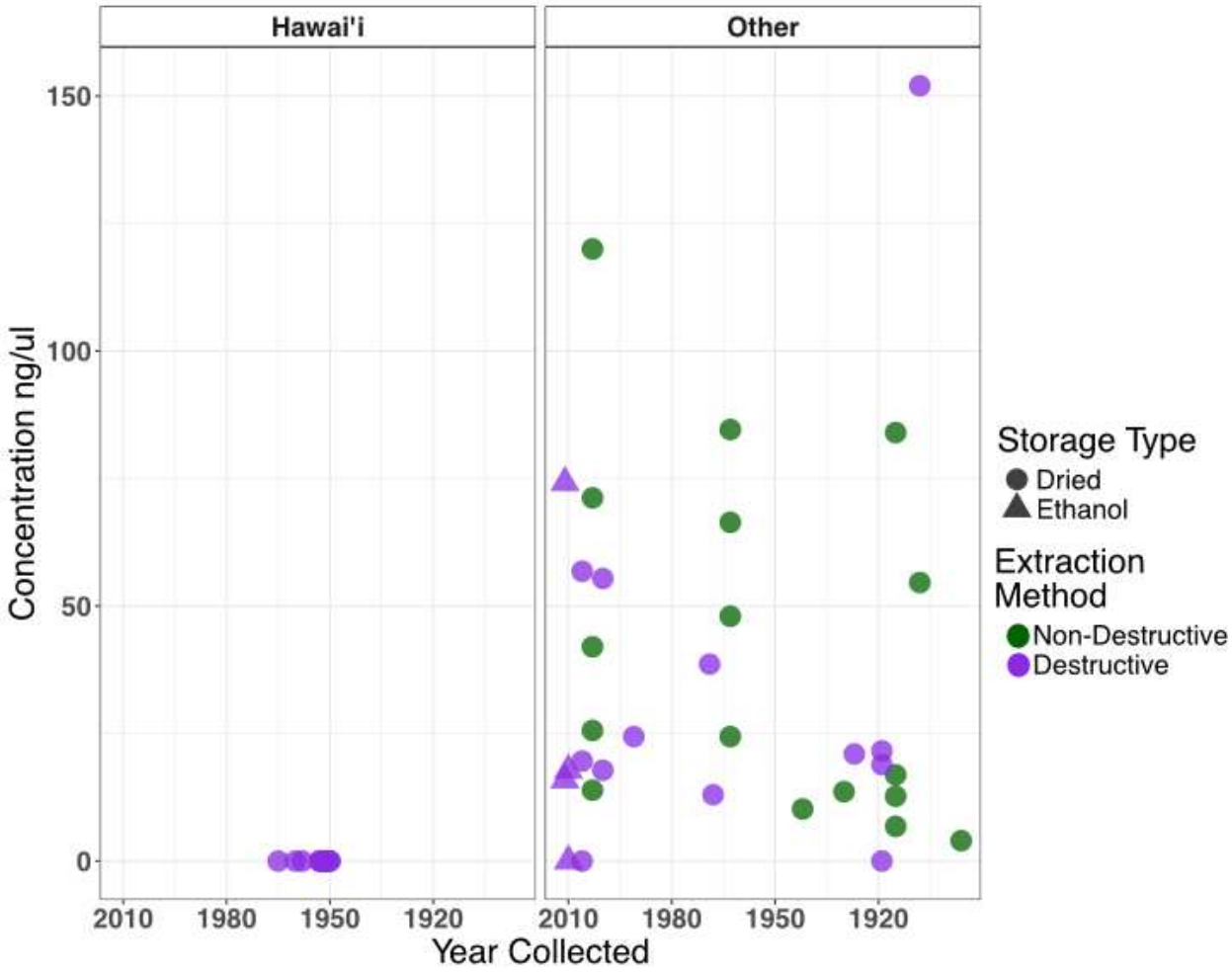


Supplemental Figure 3: Positive and negative controls in experimentally dried flies and mosquitoes showed no evidence of cross-contamination at extraction or RT-qPCR steps. Detection and threshold cycle for the indicated RNA targets are shown for experimentally dried and frozen flies (A) and mosquitoes (B). One positive and 3 negative controls were used at each time point: a cDNA positive control, a no-sample extraction blank control, a no-template RT control, and a no-sample qPCR control. Positive calls required the sample's melting temperature to be within one degree of the cDNA positive control's. Undetermined Cts (no amplification) were set to zero. Samples from negative control samples with called Cts (likely from primer dimer products) were confirmed negative by analysis of product melting temperatures and agarose gel electrophoresis. Color indicates control type and symbols indicate whether samples were called positive or negative.

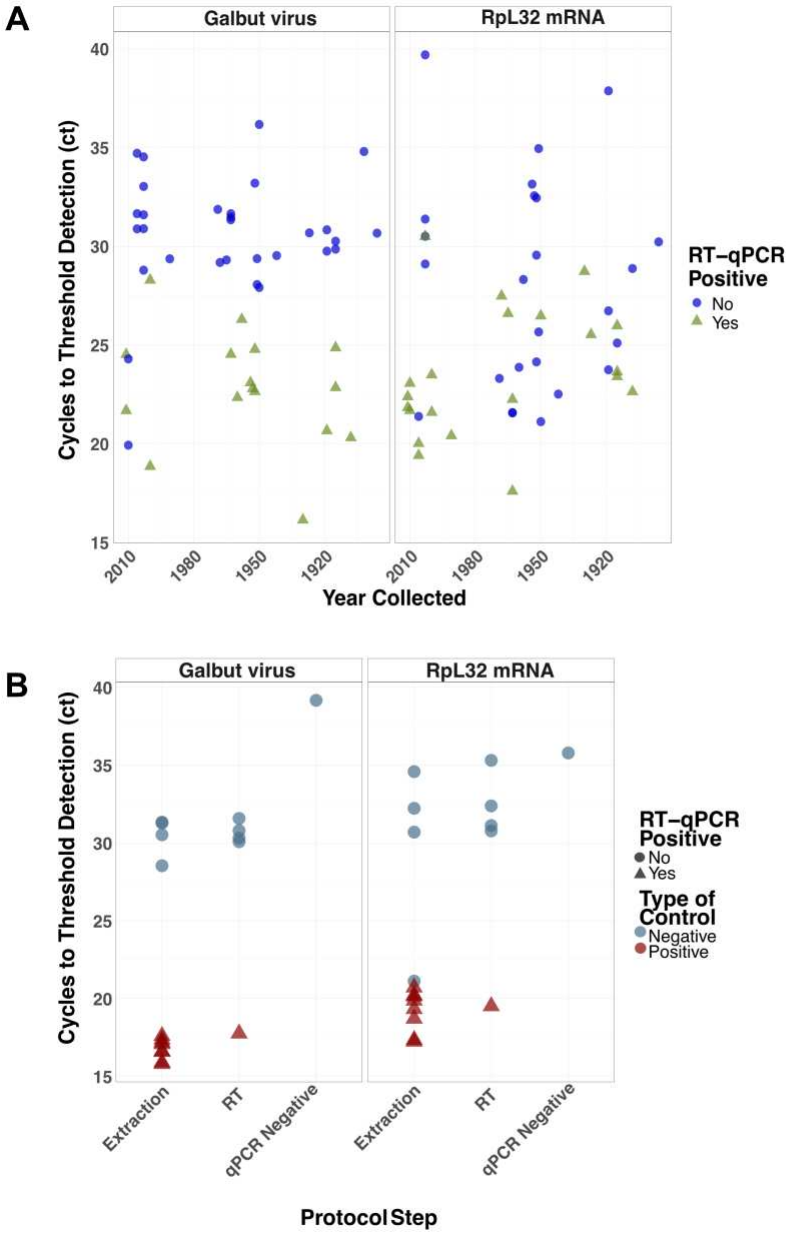


Supplemental Figure 4: Viral RNA and host mRNA persisted in dried mosquito specimens over 52 weeks. (A) Average RNA concentration and (B) RNA length of

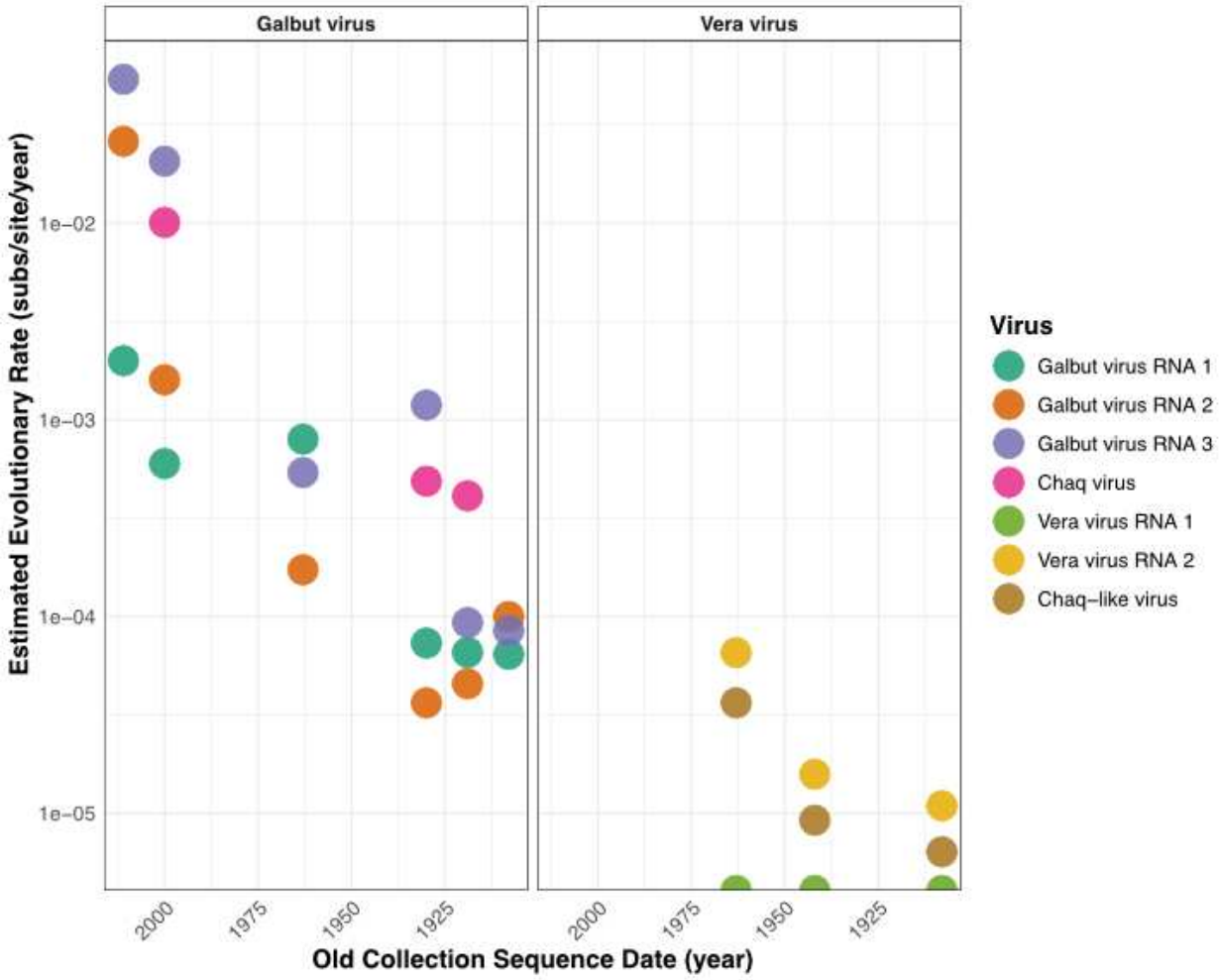
RNA from dried and frozen mosquitoes over 52 weeks. N = 3 female dried or frozen *Ae. aegypti* at each time. (C) Number of samples positive of 3 males and 3 females for actin mRNA via RT-qPCR in dried and frozen specimens over 52 weeks. (D) Average difference of actin mRNA, verdadero virus and Guadeloupe mosquito virus RNA levels relative to time point 4 weeks frozen *Ae. aegypti*. A two-tailed t-test was used to determine the statistical significance between dried and frozen means at every time point ($A < 0.05$, $B < 0.001$). (E) Comparison of mean Ct values at each time point for each RNA target. Linear regression lines are shown as is the diagonal. Error bars represent mean plus and minus the standard deviation. N = 3 male and 3 female.



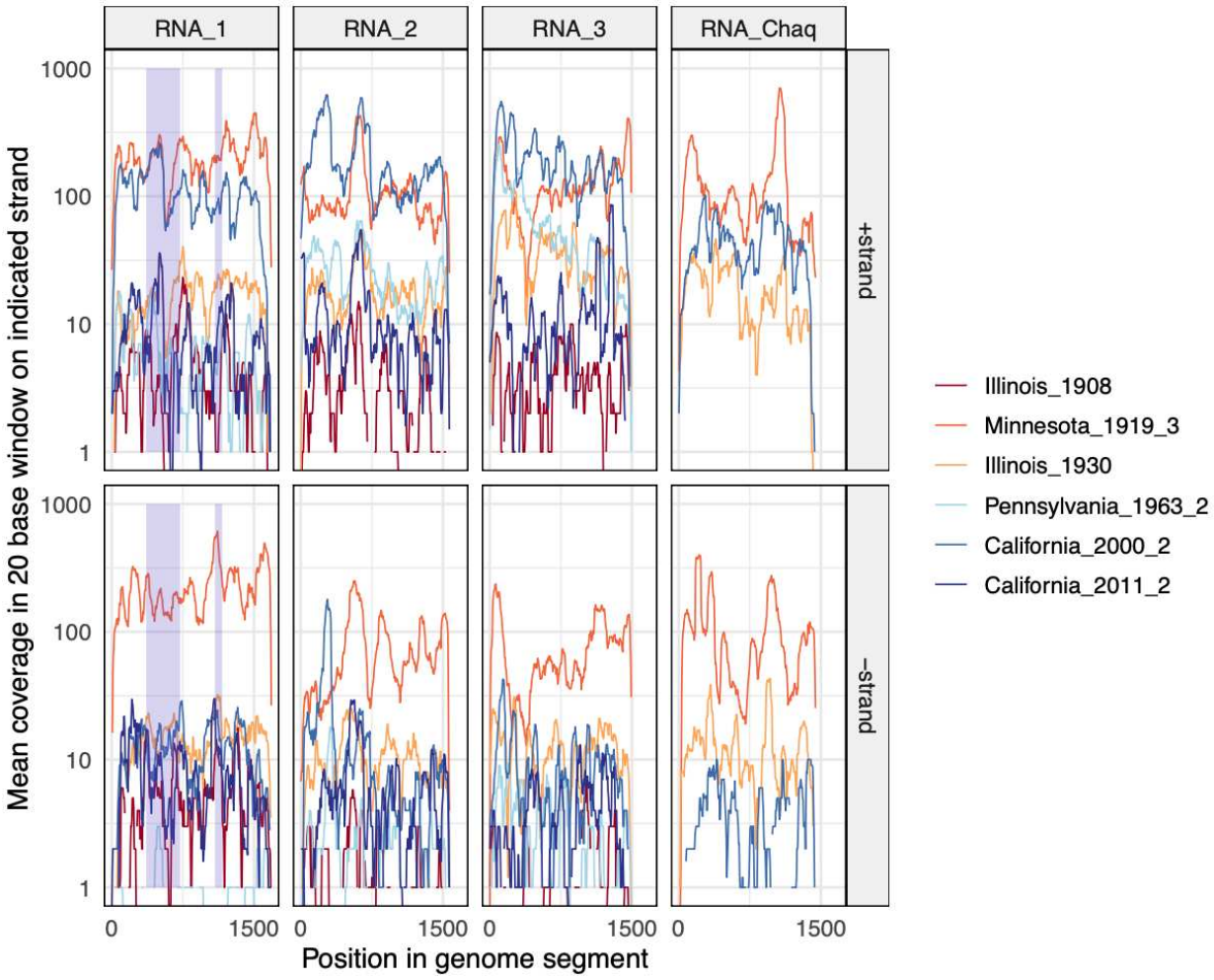
Supplemental Figure 5: RNA yield from Hawai'ian specimens was low. RNA concentrations of samples from Hawai'i or any other location. Purple = destructive RNA extraction and green = non-destructive RNA extraction. Shape indicates whether the sample was dried or stored in ethanol.



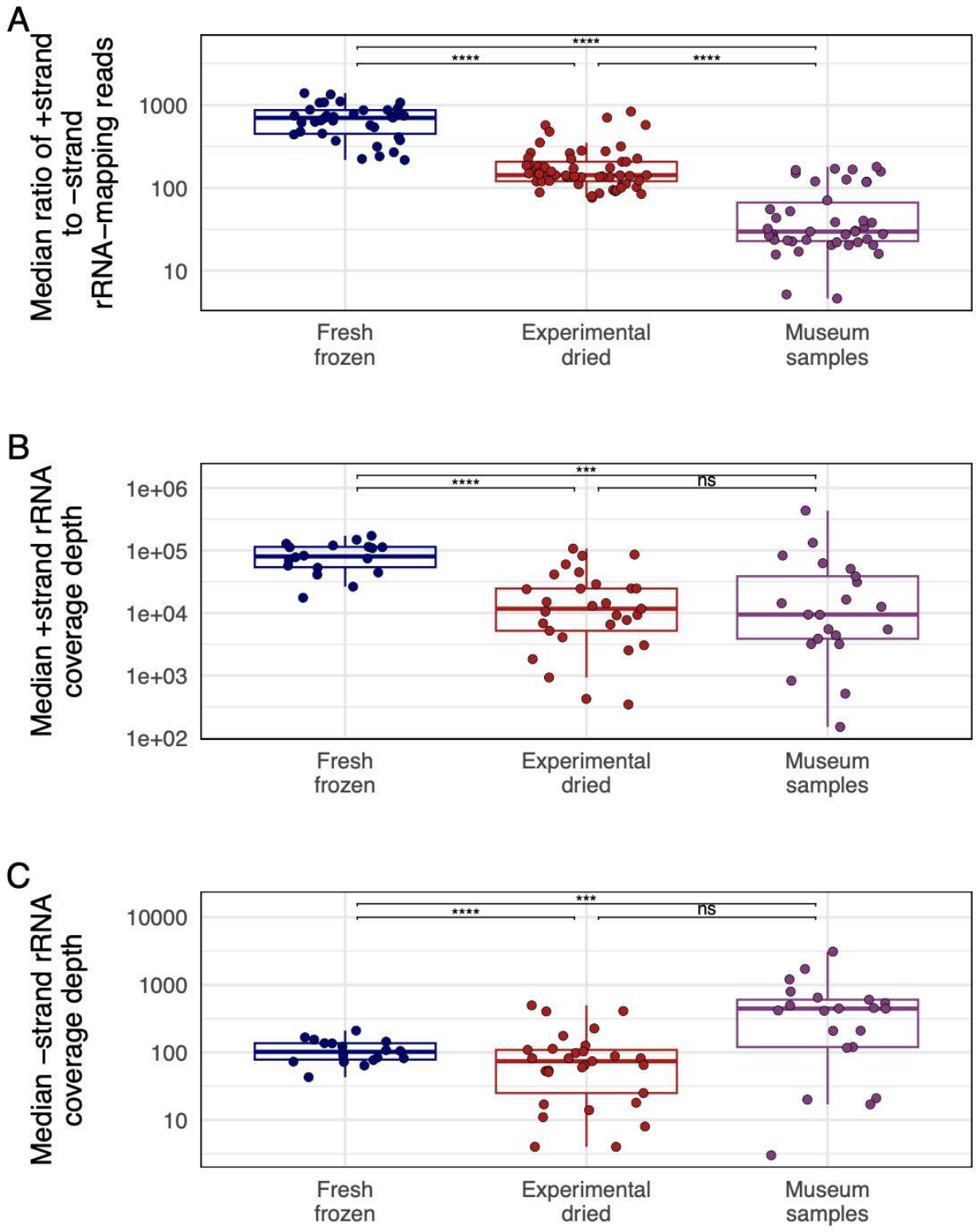
Supplemental Figure 6: Screening of Museum specimens and controls for galbut virus and RpL32 mRNA. (A) RT-qPCR results for museum specimens and (B) controls screened for galbut virus and RpL32 mRNA using short-range primers. Positive calls required that the melting temperature be within one degree of the cDNA positive control and the presence of an appropriately sized band on an agarose gel to distinguish legitimate positives from primer dimers. Undetermined Cts were set to zero. Two extraction positive (fresh FoCo-17 fly) and one negative (extraction blank) control was used for each set of extractions. Four RT no template negative controls were also used to increase the chance of detecting contamination.



Supplemental Figure 7: The estimated evolutionary rate of galbut virus and vera virus from sequences recovered from museum specimens. Rates were calculated as number of substitutions per nucleotide per year relative to the most closely related sequence available in GenBank with a known collection year. Samples without a suitably close contemporary sequence were removed from this analysis. Color indicates virus. Vera virus RNA 1 sequences are 100% identical to their closest available sequence.

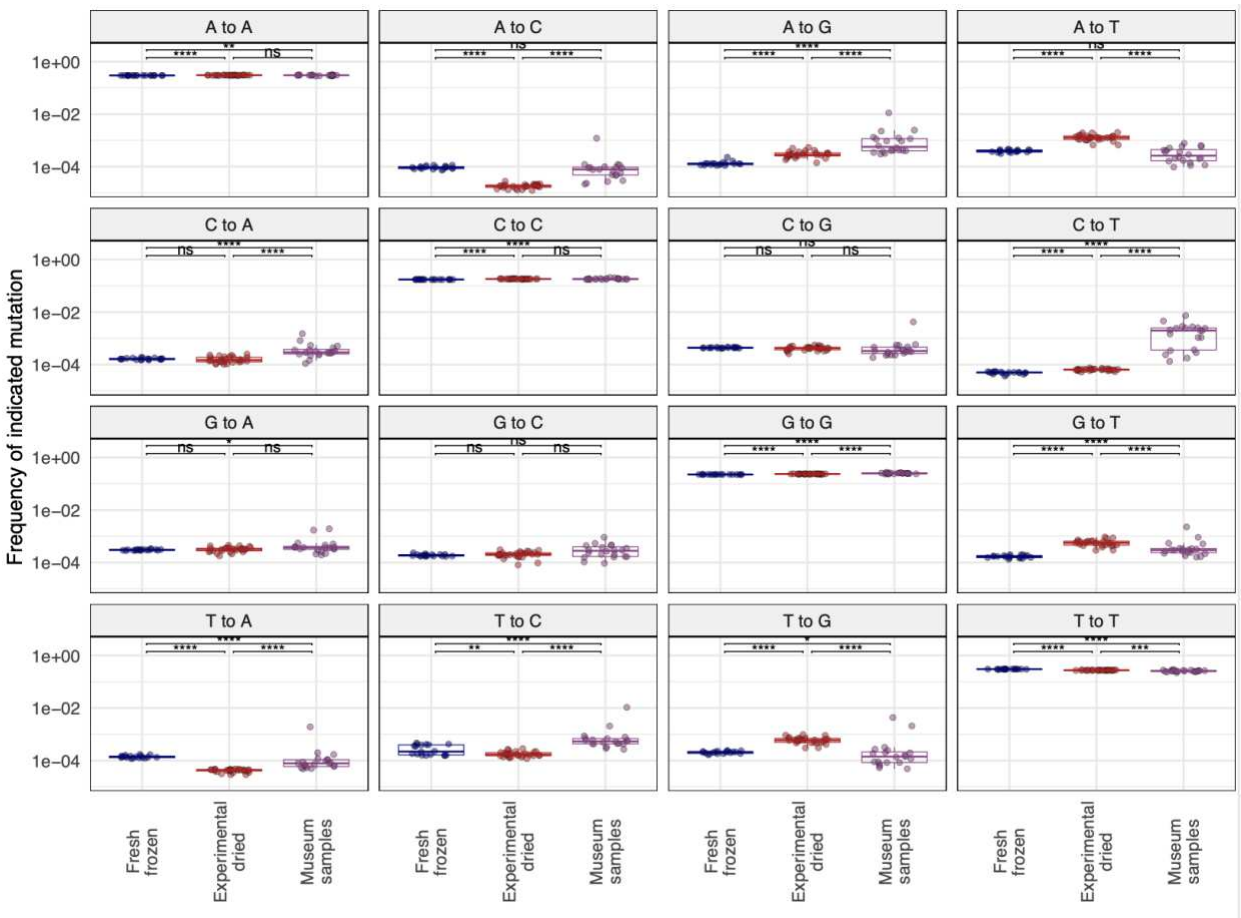


Supplemental Figure 8: Galbut virus coverage derives from both strands of all segments. Coverage levels of +strand and –strand mapping reads across galbut virus segments are plotted. The position of the long and short amplicons targeting RNA 1 are shaded in purple (**Supplemental Table 2**).

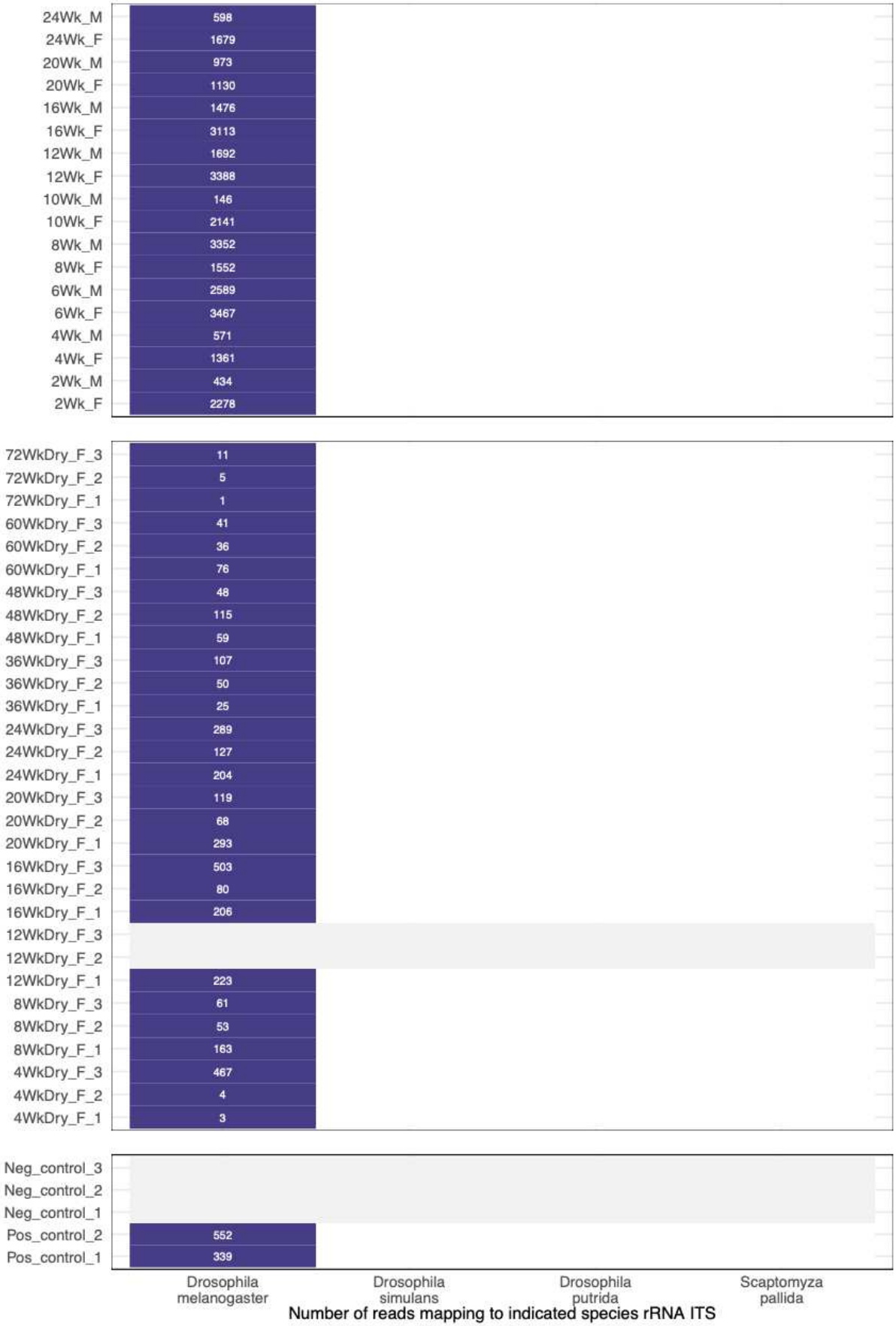


Supplemental Figure 9: Strand ratios of surviving ribosomal RNA are consistent with preferential survival of dsRNA in old samples. (A) The ratio of +strand to –

strand coverage of rRNA-mapping reads in *D. melanogaster* datasets. Each point represents the median ratio from individual-fly datasets. Adjusted *p*-values significance levels from Wilcoxon test are indicated. (B) The depth of coverage of +strand rRNA-mapping reads. (C) The depth of coverage of –strand rRNA mapping reads.

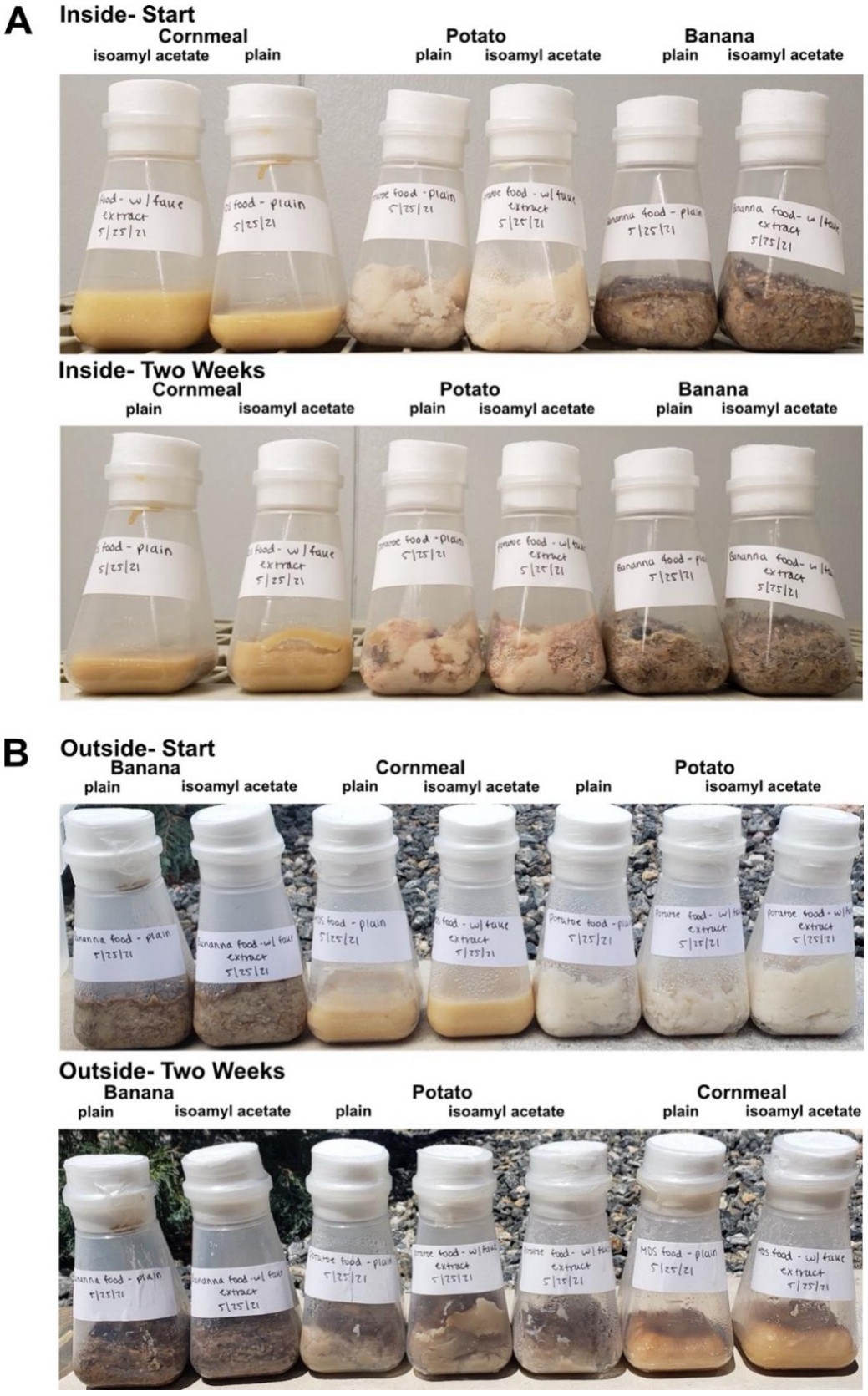


Supplemental Figure 10: Old RNA is chemically damaged and exhibits mismatch patterns consistent with cytosine and adenine deamination. Mismatches in rRNA-mapping reads from *D. melanogaster* datasets were quantified and the frequency of each mismatch type is plotted. Each point represents a dataset from an individual fly. Adjusted *p*-values significance levels from Wilcoxon test are indicated.


24Wk_M 598
24Wk_F 1679
20Wk_M 973
20Wk_F 1130
16Wk_M 1476
16Wk_F 3113
12Wk_M 1692
12Wk_F 3388
10Wk_M 146
10Wk_F 2141
8Wk_M 3352
8Wk_F 1552
6Wk_M 2589
6Wk_F 3467
4Wk_M 571
4Wk_F 1361
2Wk_M 434
2Wk_F 2278
72WkDry_F_3 11
72WkDry_F_2 5
72WkDry_F_1 1
60WkDry_F_3 41
60WkDry_F_2 36
60WkDry_F_1 76
48WkDry_F_3 48
48WkDry_F_2 115
48WkDry_F_1 59
36WkDry_F_3 107
36WkDry_F_2 50
36WkDry_F_1 25
24WkDry_F_3 289
24WkDry_F_2 127
24WkDry_F_1 204
20WkDry_F_3 119
20WkDry_F_2 68
20WkDry_F_1 293
16WkDry_F_3 503
16WkDry_F_2 80
16WkDry_F_1 206
12WkDry_F_3
12WkDry_F_2
12WkDry_F_1 223
8WkDry_F_3 61
8WkDry_F_2 53
8WkDry_F_1 163
4WkDry_F_3 467
4WkDry_F_2 4
4WkDry_F_1 3
Neg_control_3
Neg_control_2
Neg_control_1
Pos_control_2 552
Pos_control_1 339
Drosophila melanogaster
Drosophila simulans
Drosophila putrida
Scaptomyza pallida
Number of reads mapping to indicated species rRNA ITS

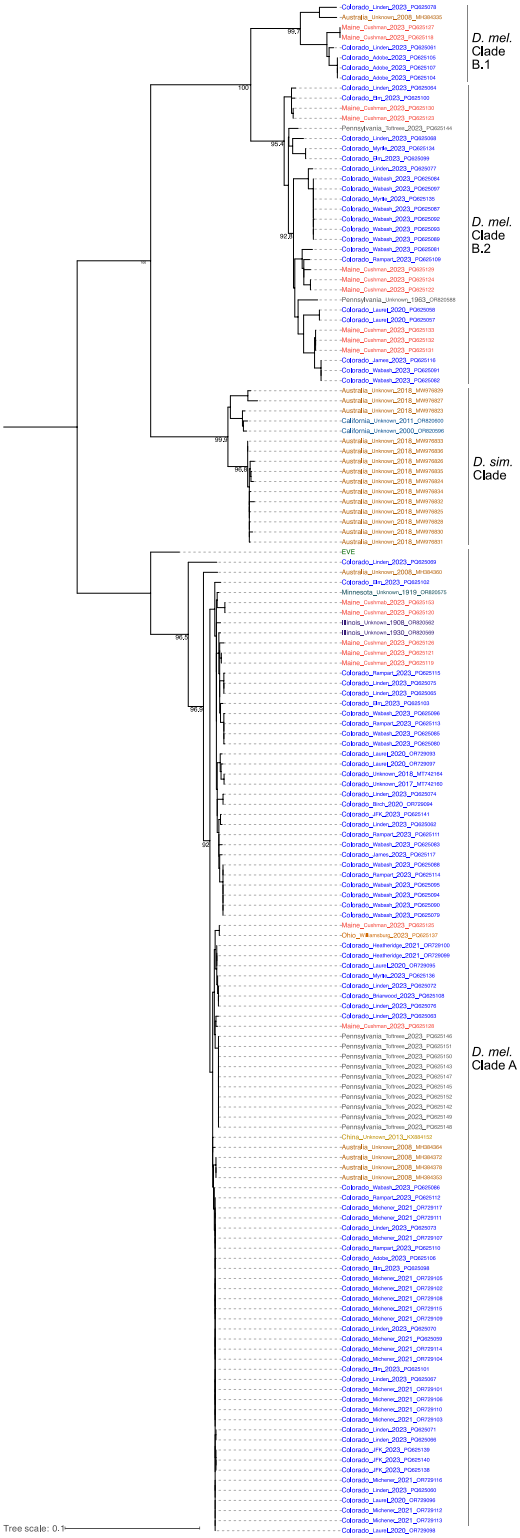
Supplemental Figure 11: Host-mapping reads exhibit species specificity. The number of reads mapping to the indicated species rRNA internal transcribed spacer 1 (ITS) sequence for each fresh-frozen, experimentally dried, or control dataset is shown. Datasets with no ITS-mapping reads, including negative control water libraries, are represented with grey.

Appendix B: Chapter 3 Supplemental Figure

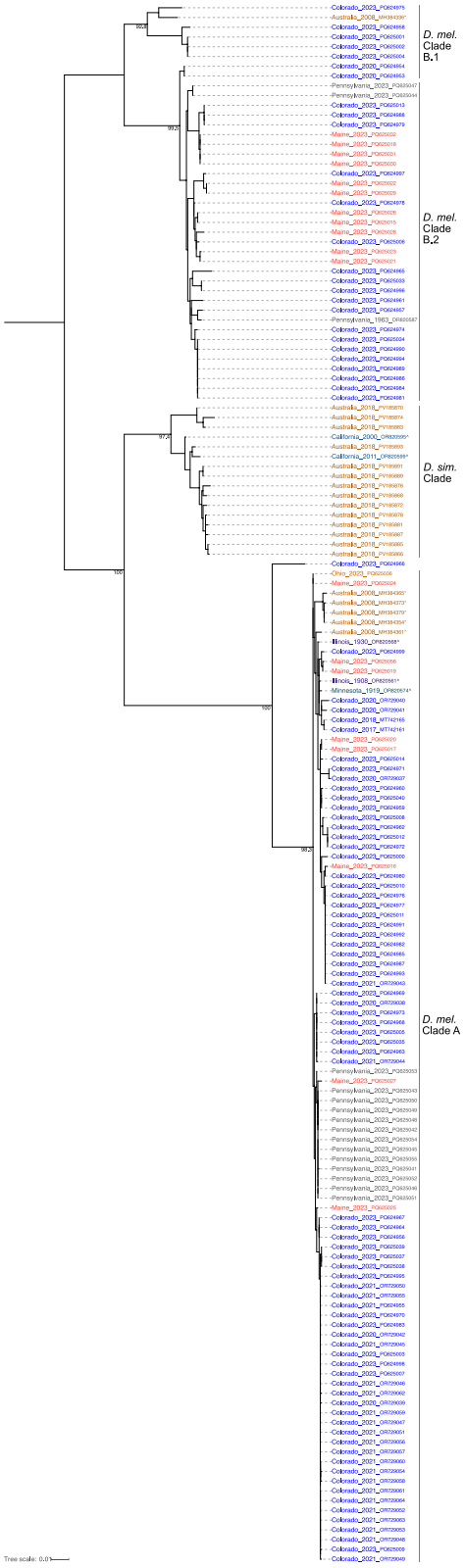


Supplemental Figure 1: Cornmeal-based fly food was stable over two weeks in indoor and outdoor environments. Cornmeal-, potato-, and banana-based fly foods were placed either in our indoor insectary (A) or outside (B). Representative before and after pictures are shown for both experiments. The food was visually inspected daily over the experimental period.

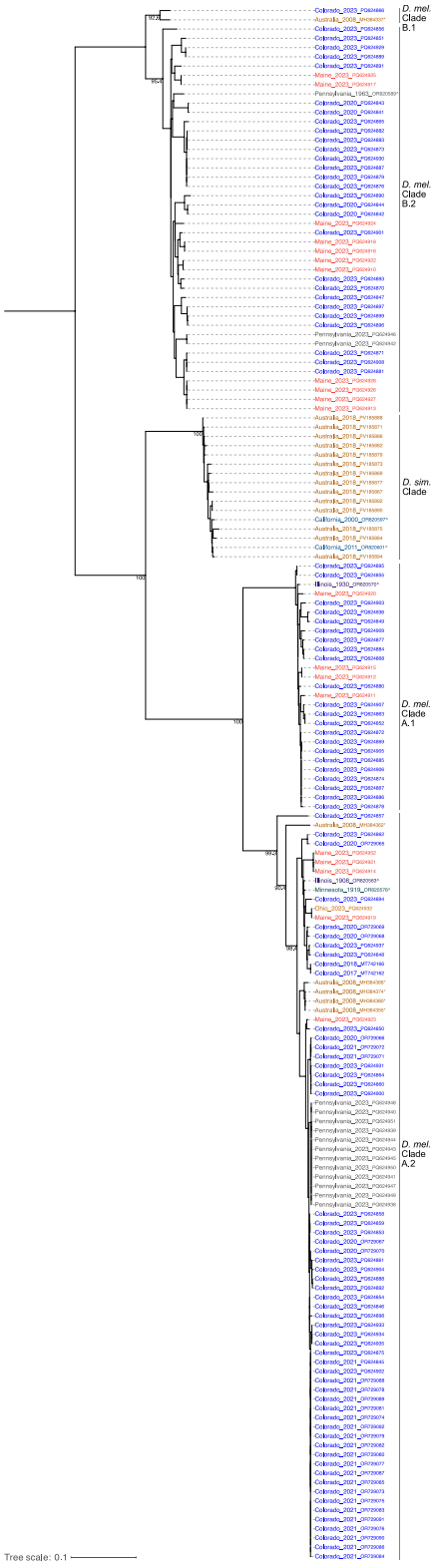
Appendix C: Chapter 4 Supplemental Figures



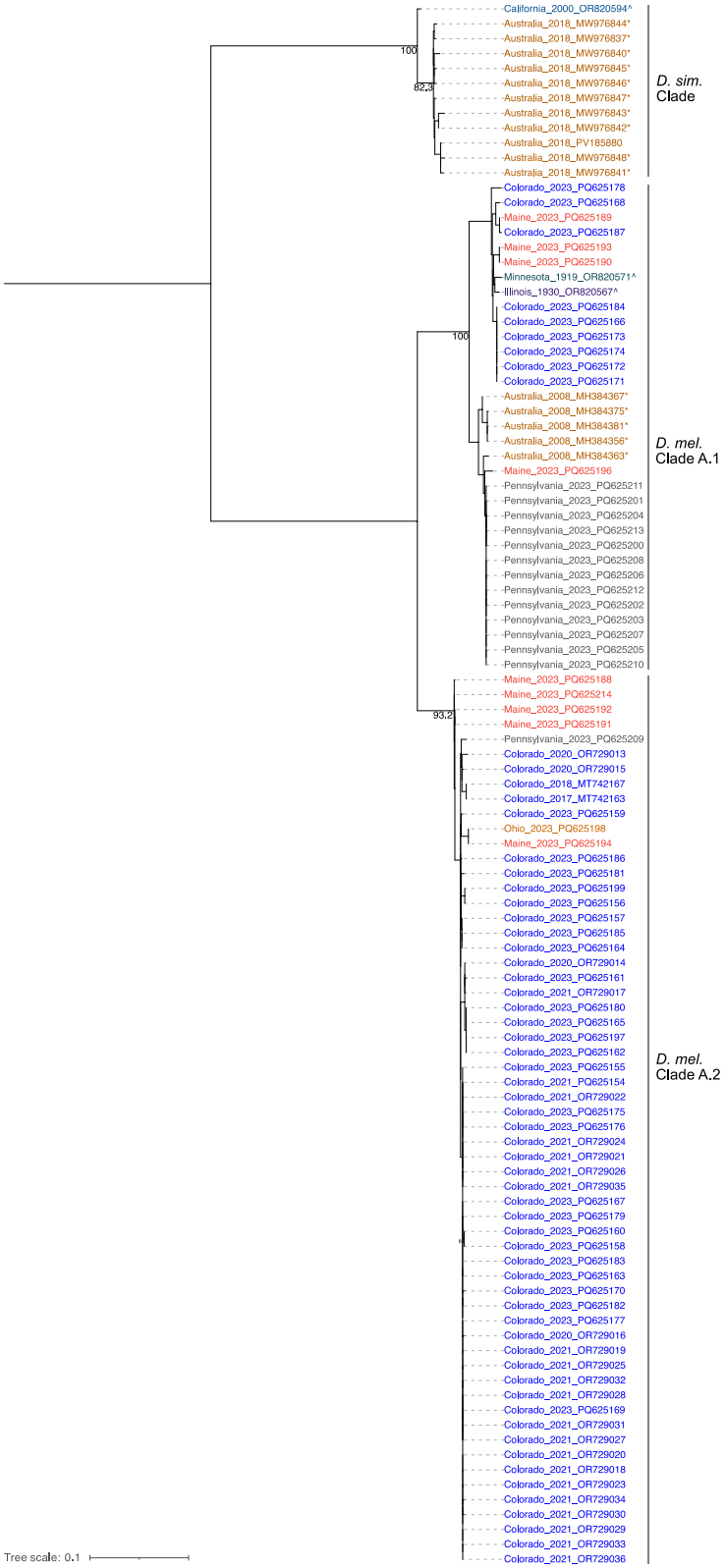
Supplemental Figure 1 Midpoint rooted maximum likelihood tree of galbut virus RNA 1 nucleotide sequences extended data. Data as shown in Figure 4 but with no clades collapsed.



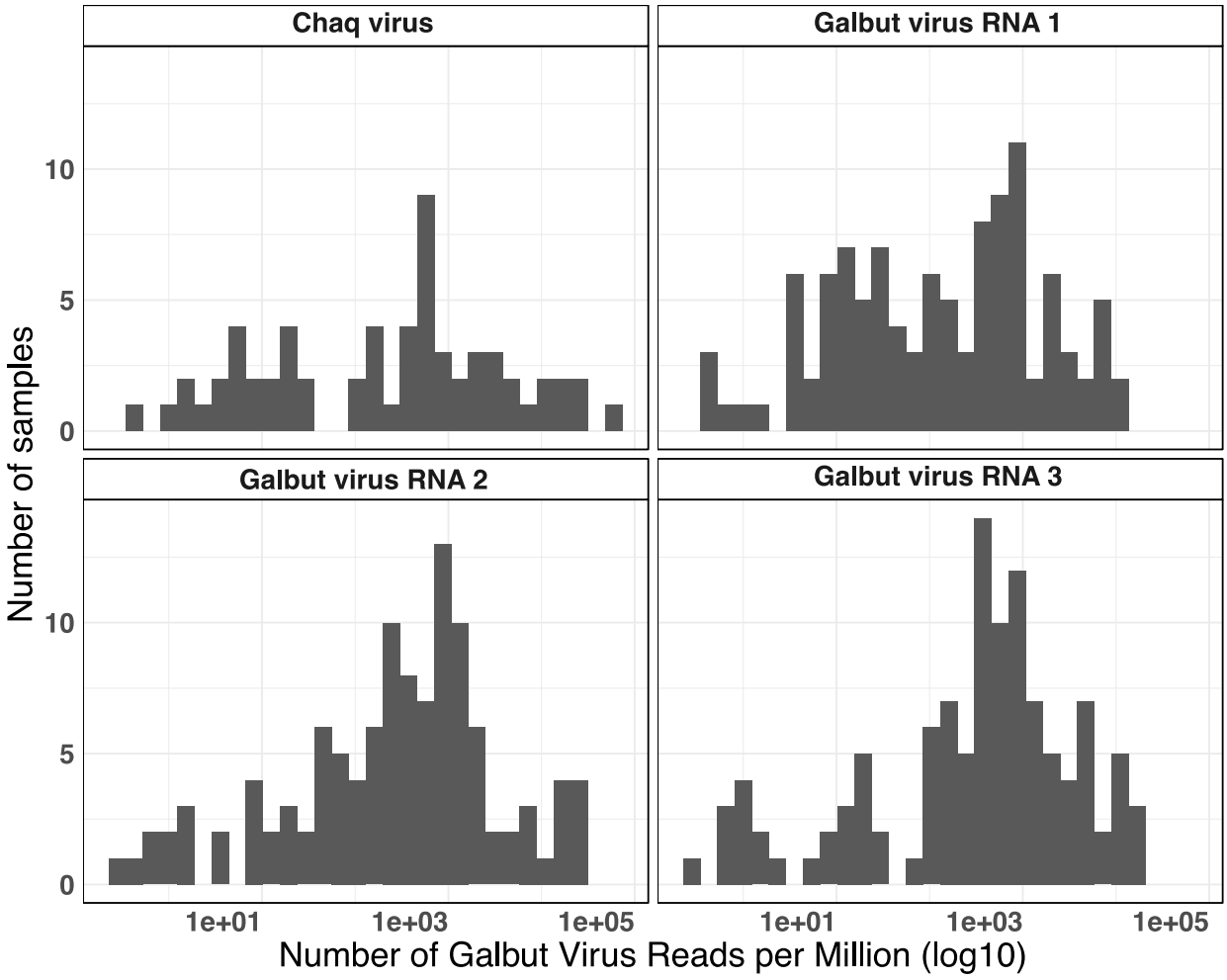
Supplemental Figure 2: Midpoint rooted maximum likelihood tree of galbut virus RNA 2 nucleotide sequences extended data. Data as shown in Figure 5 but with no clades collapsed.



Supplemental Figure 3: Midpoint rooted maximum likelihood tree of galbut virus RNA 3 nucleotide sequences extended data. Data as shown in Figure 5 but with no clades collapsed.



Supplemental Figure 4: Midpoint rooted maximum likelihood tree of chaq virus nucleotide sequences extended data. Data as shown in Figure 5 but with no clades collapsed.



Supplemental Figure 5: There is a bimodal infection pattern in galbut virus RNAs 1 and 2 and chaq virus but a trimodal infection pattern in galbut virus RNA 3. The number of galbut virus and chaq virus reads (per million) are shown as a histogram of the number of samples in each bin. Plots are faceted by segment.

Appendix D: Chapter 4 Supplemental Tables

Supplemental Table 1: Dunn's test pairwise comparisons of prevalence between sampling locations

Comparison	Z	P.unadj	P.adj
Adobe - Briarwood	1.98	4.79E-02	1.00E+00
Adobe - Elm	2.84	4.56E-03	3.01E-01
Briarwood - Elm	-0.49	6.23E-01	1.00E+00
Adobe - James	1.23	2.21E-01	1.00E+00
Briarwood - James	-0.85	3.97E-01	1.00E+00
Elm - James	-0.68	4.94E-01	1.00E+00
Adobe - JFK Parkway	-0.61	5.40E-01	1.00E+00
Briarwood - JFK Parkway	-2.57	1.02E-02	6.71E-01
Elm - JFK Parkway	-6.04	1.52E-09	1.00E-07
James - JFK Parkway	-1.89	5.86E-02	1.00E+00
Adobe - Linden	1.70	8.89E-02	1.00E+00
Briarwood - Linden	-1.26	2.07E-01	1.00E+00
Elm - Linden	-2.48	1.31E-02	8.64E-01
James - Linden	-0.25	8.06E-01	1.00E+00
JFK Parkway - Linden	4.88	1.04E-06	6.85E-05
Adobe - Maine	0.89	3.74E-01	1.00E+00
Briarwood - Maine	-1.54	1.23E-01	1.00E+00
Elm - Maine	-2.41	1.60E-02	1.00E+00
James - Maine	-0.65	5.13E-01	1.00E+00
JFK Parkway - Maine	2.14	3.22E-02	1.00E+00
Linden - Maine	-0.89	3.72E-01	1.00E+00
Adobe - Myrtle	3.35	8.14E-04	5.37E-02
Briarwood - Myrtle	0.14	8.92E-01	1.00E+00
Elm - Myrtle	1.25	2.13E-01	1.00E+00
James - Myrtle	1.35	1.78E-01	1.00E+00
JFK Parkway - Myrtle	5.46	4.89E-08	3.23E-06
Linden - Myrtle	2.96	3.10E-03	2.04E-01
Maine - Myrtle	2.99	2.75E-03	1.81E-01
Adobe - Ohio	1.34	1.80E-01	1.00E+00
Briarwood - Ohio	-0.23	8.15E-01	1.00E+00
Elm - Ohio	0.10	9.23E-01	1.00E+00
James - Ohio	0.46	6.45E-01	1.00E+00
JFK Parkway - Ohio	1.72	8.54E-02	1.00E+00
Linden - Ohio	0.69	4.91E-01	1.00E+00
Maine - Ohio	0.95	3.44E-01	1.00E+00
Myrtle - Ohio	-0.39	6.98E-01	1.00E+00
Adobe - Pennsylvania	-2.64	8.23E-03	5.43E-01
Briarwood - Pennsylvania	-3.79	1.51E-04	9.94E-03
Elm - Pennsylvania	-7.47	7.73E-14	5.10E-12
James - Pennsylvania	-3.39	7.02E-04	4.63E-02
JFK Parkway - Pennsylvania	-3.15	1.65E-03	1.09E-01
Linden - Pennsylvania	-6.57	4.87E-11	3.22E-09
Maine - Pennsylvania	-4.32	1.53E-05	1.01E-03
Myrtle - Pennsylvania	-7.00	2.60E-12	1.72E-10
Ohio - Pennsylvania	-2.74	6.22E-03	4.11E-01
Adobe - Rampart	3.19	1.43E-03	9.41E-02
Briarwood - Rampart	-0.30	7.68E-01	1.00E+00
Elm - Rampart	0.55	5.82E-01	1.00E+00
James - Rampart	0.93	3.51E-01	1.00E+00
JFK Parkway - Rampart	6.78	1.22E-11	8.06E-10
Linden - Rampart	3.24	1.21E-03	8.01E-02
Maine - Rampart	2.87	4.07E-03	2.69E-01
Myrtle - Rampart	-0.87	3.85E-01	1.00E+00
Ohio - Rampart	0.06	9.53E-01	1.00E+00
Pennsylvania - Rampart	8.00	1.22E-15	8.05E-14
Adobe - Wabash	1.59	1.11E-01	1.00E+00
Briarwood - Wabash	-1.35	1.77E-01	1.00E+00
Elm - Wabash	-2.93	3.42E-03	2.26E-01
James - Wabash	-0.35	7.27E-01	1.00E+00
JFK Parkway - Wabash	5.00	5.85E-07	3.86E-05
Linden - Wabash	-0.41	6.79E-01	1.00E+00
Maine - Wabash	0.73	4.66E-01	1.00E+00
Myrtle - Wabash	-3.21	1.34E-03	8.85E-02
Ohio - Wabash	-0.75	4.51E-01	1.00E+00
Pennsylvania - Wabash	6.61	3.92E-11	2.59E-09
Rampart - Wabash	-3.76	1.71E-04	1.13E-02

Supplemental Table 2: Sites experiencing purifying selection.

Segment	Codon	Alpha	Beta	Alpha=Beta	LRT	p-value	Total Branch Length
Chaq	58	10.04	0.00	5.84	4.83	0.03	0.05
Chaq	61	6.56	0.00	6.49	2.93	0.09	0.06
Chaq	84	5.44	0.00	2.99	2.96	0.09	0.03
Chaq	120	5.03	0.00	2.89	3.14	0.08	0.02
Chaq	127	4.99	0.00	2.85	3.35	0.07	0.02
Chaq	144	7.89	0.00	4.47	4.29	0.04	0.04
Chaq	149	9.06	0.00	5.38	4.89	0.03	0.05
Chaq	166	5.66	0.00	3.57	3.39	0.07	0.03
Chaq	175	11.36	0.00	6.11	4.74	0.03	0.05
Chaq	179	7.46	0.00	4.37	4.22	0.04	0.04
Chaq	196	6.52	0.00	3.64	3.74	0.05	0.03
Chaq	206	10.78	0.00	6.17	5.44	0.02	0.05
Chaq	210	5.05	0.00	2.38	2.98	0.08	0.02
Chaq	218	6.57	0.00	2.94	3.63	0.06	0.03
Chaq	221	5.18	0.00	3.29	3.37	0.07	0.03
Chaq	232	6.45	0.00	3.86	3.19	0.07	0.04
Chaq	233	6.21	0.00	3.08	3.15	0.08	0.03
Chaq	261	5.00	0.00	2.53	2.82	0.09	0.02
Chaq	263	7.14	0.00	3.30	3.70	0.05	0.03
Chaq	291	8.05	0.00	4.11	3.61	0.06	0.06
Chaq	309	6.99	0.00	4.83	3.35	0.07	0.04
RNA1	7	2.54	0.00	1.34	3.53	0.06	0.11
RNA1	8	6.78	0.00	5.15	6.00	0.01	0.36
RNA1	15	2.64	0.00	1.45	3.25	0.07	0.10
RNA1	24	4.76	0.00	2.55	6.75	0.01	0.18
RNA1	53	1.78	0.00	1.06	3.08	0.08	0.08
RNA1	57	4.82	0.00	2.49	6.76	0.01	0.18
RNA1	58	2.82	0.00	1.46	4.00	0.05	0.10
RNA1	60	2.37	0.00	1.48	3.77	0.05	0.11
RNA1	62	1.95	0.00	1.11	3.28	0.07	0.08
RNA1	65	8.48	0.00	4.08	10.10	0.00	0.29
RNA1	75	2.16	0.00	0.99	3.09	0.08	0.07
RNA1	79	4.49	0.00	2.81	5.41	0.02	0.21
RNA1	92	1.93	0.00	0.91	2.99	0.08	0.06
RNA1	97	3.01	0.00	1.85	3.86	0.05	0.13
RNA1	99	1.88	0.00	0.91	2.81	0.09	0.06
RNA1	101	1.79	0.00	1.07	3.12	0.08	0.08
RNA1	111	4.45	0.00	2.37	6.39	0.01	0.17
RNA1	119	5.14	0.00	2.60	6.87	0.01	0.18
RNA1	128	2.46	0.00	1.47	3.82	0.05	0.10
RNA1	140	2.67	0.00	1.39	4.20	0.04	0.10
RNA1	141	3.55	0.00	1.99	5.50	0.02	0.14
RNA1	146	3.02	0.00	1.79	5.01	0.03	0.13
RNA1	147	2.10	0.00	1.30	2.83	0.09	0.09
RNA1	148	3.56	0.00	1.48	5.56	0.02	0.11
RNA1	149	2.24	0.00	1.22	3.59	0.06	0.09
RNA1	150	46.09	0.00	0.00	6.17	0.01	0.01
RNA1	161	2.84	0.00	1.34	4.08	0.04	0.10
RNA1	162	2.15	0.00	1.21	3.40	0.07	0.09
RNA1	170	2.05	0.00	1.26	2.76	0.10	0.09
RNA1	176	2.06	0.00	1.15	3.34	0.07	0.08
RNA1	180	2.87	0.00	1.86	4.54	0.03	0.13
RNA1	182	1.64	0.00	1.01	2.93	0.09	0.07
RNA1	186	3.23	0.00	1.41	4.47	0.03	0.10
RNA1	188	1.77	0.00	1.04	3.07	0.08	0.07
RNA1	210	4.17	0.00	2.52	5.64	0.02	0.18
RNA1	212	5.38	0.00	3.06	7.14	0.01	0.22
RNA1	213	1.93	0.00	1.16	3.09	0.08	0.08
RNA1	214	2.67	0.00	1.30	4.01	0.05	0.09
RNA1	215	1.74	0.00	1.03	3.03	0.08	0.07
RNA1	218	3.04	0.00	2.02	3.74	0.05	0.14
RNA1	219	4.39	0.00	2.37	5.88	0.02	0.17
RNA1	224	2.46	0.00	1.50	3.96	0.05	0.11
RNA1	225	2.64	0.00	1.88	2.86	0.09	0.15
RNA1	227	3.95	0.00	3.05	4.01	0.05	0.22
RNA1	231	6.63	0.00	2.79	7.29	0.01	0.20
RNA1	233	2.28	0.00	1.35	4.19	0.04	0.10
RNA1	235	7.38	0.00	4.59	9.22	0.00	0.32
RNA1	236	1.67	0.00	1.03	2.97	0.09	0.07

RNA1	241	3.09	0.00	1.70	4.51	0.03	0.12
RNA1	245	1.88	0.00	1.06	3.34	0.07	0.07
RNA1	246	1.90	0.00	1.07	3.34	0.07	0.08
RNA1	254	1.93	0.00	1.13	3.05	0.08	0.08
RNA1	255	1.85	0.00	0.90	2.77	0.10	0.06
RNA1	257	1.86	0.00	1.08	3.20	0.07	0.08
RNA1	258	3.22	0.00	1.62	4.49	0.03	0.11
RNA1	259	2.21	0.00	1.32	2.90	0.09	0.09
RNA1	261	1.69	0.00	0.85	2.76	0.10	0.06
RNA1	265	3.04	0.00	1.53	4.29	0.04	0.11
RNA1	271	5.24	0.00	2.89	5.73	0.02	0.20
RNA1	273	4.06	0.00	2.41	6.34	0.01	0.17
RNA1	286	5.21	0.00	2.38	6.90	0.01	0.17
RNA1	288	3.24	0.00	1.62	4.50	0.03	0.11
RNA1	289	3.18	0.00	1.79	4.84	0.03	0.13
RNA1	290	1.88	0.00	1.15	2.92	0.09	0.09
RNA1	294	4.52	0.00	2.24	5.65	0.02	0.16
RNA1	296	6.04	0.00	4.39	6.31	0.01	0.32
RNA1	297	1.68	0.00	0.86	2.74	0.10	0.06
RNA1	298	3.27	0.00	2.60	3.23	0.07	0.18
RNA1	301	2.65	0.00	1.52	4.73	0.03	0.11
RNA1	304	4.58	0.00	2.58	4.49	0.03	0.19
RNA1	312	5.63	0.00	3.08	6.03	0.01	0.22
RNA1	313	5.19	0.00	2.37	6.11	0.01	0.17
RNA1	314	8.93	0.00	5.84	10.91	0.00	0.41
RNA1	315	3.92	0.00	2.08	5.60	0.02	0.15
RNA1	319	3.68	0.00	2.16	6.35	0.01	0.15
RNA1	320	2.16	0.00	1.20	3.54	0.06	0.08
RNA1	326	3.42	0.00	2.03	5.10	0.02	0.14
RNA1	327	1.73	0.00	1.03	3.09	0.08	0.07
RNA1	338	3.00	0.00	1.65	4.29	0.04	0.12
RNA1	341	4.28	0.00	2.45	6.20	0.01	0.17
RNA1	354	2.03	0.00	0.88	3.38	0.07	0.06
RNA1	390	4.00	0.00	2.30	6.20	0.01	0.16
RNA1	394	3.94	0.00	2.23	4.67	0.03	0.16
RNA1	395	2.74	0.00	1.46	3.32	0.07	0.10
RNA1	396	4.27	0.00	2.05	5.56	0.02	0.14
RNA1	400	4.20	0.00	2.05	6.04	0.01	0.14
RNA1	404	3.37	0.00	1.57	4.54	0.03	0.11
RNA1	408	4.83	0.00	2.59	6.85	0.01	0.18
RNA1	410	2.58	0.00	1.62	3.77	0.05	0.11
RNA1	411	2.18	0.00	1.28	3.42	0.06	0.09
RNA1	416	4.58	0.00	2.00	5.66	0.02	0.14
RNA1	418	2.48	0.00	1.44	4.38	0.04	0.10
RNA1	420	4.64	0.00	2.73	5.36	0.02	0.20
RNA1	421	2.19	0.00	1.21	3.56	0.06	0.09
RNA1	424	2.78	0.00	1.57	4.85	0.03	0.11
RNA1	427	2.04	0.00	1.19	3.06	0.08	0.08
RNA1	430	3.46	0.00	2.02	5.49	0.02	0.14
RNA1	431	2.73	0.00	1.61	3.92	0.05	0.11
RNA1	435	3.05	0.00	1.59	4.63	0.03	0.11
RNA1	438	5.33	0.00	3.37	7.50	0.01	0.24
RNA1	441	3.61	0.00	2.24	3.72	0.05	0.16
RNA1	442	2.84	0.00	1.51	3.43	0.06	0.11
RNA1	448	2.05	0.00	1.12	3.56	0.06	0.08
RNA1	449	8.79	0.00	5.35	8.37	0.00	0.38
RNA1	453	2.20	0.00	1.52	2.99	0.08	0.11
RNA1	454	2.55	0.00	1.46	4.19	0.04	0.10
RNA1	456	3.97	0.00	2.31	6.66	0.01	0.16
RNA1	461	1.84	0.00	1.08	3.18	0.07	0.08
RNA1	463	2.10	0.00	1.09	3.64	0.06	0.08
RNA1	469	3.49	0.00	1.64	4.65	0.03	0.12
RNA1	470	2.41	0.00	1.87	2.71	0.10	0.13
RNA1	471	3.33	0.00	1.76	4.90	0.03	0.12
RNA1	475	3.74	0.00	2.95	3.69	0.05	0.21
RNA1	479	1.96	0.00	1.10	3.29	0.07	0.08
RNA1	480	2.04	0.00	1.20	3.12	0.08	0.09
RNA1	481	4.77	0.00	2.44	6.61	0.01	0.17
RNA1	484	3.21	0.00	2.00	4.88	0.03	0.14
RNA1	487	1.72	0.00	0.87	2.74	0.10	0.06

RNA1	488	12.62	0.00	7.16	12.15	5e-04	0.50
RNA1	492	2.53	0.00	1.94	2.72	0.10	0.14
RNA1	494	5.18	0.00	3.06	7.50	0.01	0.22
RNA1	503	3.17	0.00	1.88	5.18	0.02	0.13
RNA1	504	4.07	0.00	2.48	4.92	0.03	0.17
RNA1	507	1.79	0.00	1.01	3.27	0.07	0.07
RNA1	510	4.23	0.00	2.47	6.65	0.01	0.17
RNA1	512	3.22	0.00	1.51	4.65	0.03	0.11
RNA1	514	2.80	0.00	1.47	4.02	0.05	0.10
RNA1	520	2.65	0.00	1.66	3.48	0.06	0.12
RNA1	521	4.98	0.00	2.37	5.23	0.02	0.17
RNA1	522	3.66	0.00	1.55	4.99	0.03	0.11
RNA1	526	1.77	0.00	1.06	3.07	0.08	0.08
RNA1	528	1.52	0.00	0.75	2.91	0.09	0.05
RNA1	529	1.94	0.00	0.88	2.98	0.08	0.06
RNA1	531	2.56	0.00	1.48	4.21	0.04	0.11
RNA1	540	3.27	0.00	1.99	4.91	0.03	0.14
RNA2	25	9.58	0.00	5.82	5.23	0.02	0.13
RNA2	34	6.87	0.00	3.32	4.90	0.03	0.06
RNA2	58	3.97	0.00	2.36	3.05	0.08	0.06
RNA2	60	6.85	0.00	3.74	4.81	0.03	0.06
RNA2	70	4.97	0.00	2.87	2.97	0.08	0.05
RNA2	89	9.26	0.00	4.50	6.89	0.01	0.08
RNA2	106	4.76	0.00	2.38	3.93	0.05	0.04
RNA2	118	3.55	0.00	2.39	3.13	0.08	0.04
RNA2	129	4.34	0.00	2.40	3.50	0.06	0.04
RNA2	139	16.61	0.00	6.31	12.60	4e-04	0.11
RNA2	145	6.64	0.00	4.08	5.30	0.02	0.07
RNA2	146	5.76	0.00	4.35	3.64	0.06	0.07
RNA2	155	4.19	0.00	2.50	3.79	0.05	0.04
RNA2	159	26.24	0.00	9.96	17.32	0.00	0.17
RNA2	170	5.05	0.00	3.26	3.92	0.05	0.05
RNA2	173	5.58	0.00	2.99	3.23	0.07	0.05
RNA2	200	11.27	0.00	6.33	5.79	0.02	0.11
RNA2	204	6.06	0.00	3.58	5.01	0.03	0.08
RNA2	223	7.42	0.00	3.03	6.06	0.01	0.05
RNA2	226	9.99	0.00	4.21	7.70	0.01	0.07
RNA2	231	4.35	0.00	2.77	3.47	0.06	0.05
RNA2	244	3.92	0.00	2.50	3.33	0.07	0.04
RNA2	245	3.15	0.00	2.04	2.74	0.10	0.03
RNA2	262	10.33	0.00	5.31	7.28	0.01	0.09
RNA2	269	6.07	0.00	2.65	5.78	0.02	0.04
RNA2	276	17.24	0.00	8.10	9.91	0.00	0.14
RNA2	302	4.25	0.00	2.19	3.85	0.05	0.04
RNA2	321	7.17	0.00	3.53	5.59	0.02	0.06
RNA2	334	5.05	0.00	2.53	4.05	0.04	0.04
RNA2	335	6.64	0.00	3.45	5.47	0.02	0.07
RNA2	336	3.19	0.00	1.90	2.90	0.09	0.03
RNA2	337	5.66	0.00	2.68	5.66	0.02	0.06
RNA2	343	5.08	0.00	2.57	3.85	0.05	0.04
RNA2	347	5.74	0.00	2.48	4.59	0.03	0.04
RNA2	350	7.18	0.00	3.46	6.06	0.01	0.06
RNA2	351	4.89	0.00	2.89	4.24	0.04	0.07
RNA2	358	6.62	0.00	2.61	5.82	0.02	0.04
RNA2	366	2.57	0.00	1.27	2.76	0.10	0.02
RNA2	375	4.24	0.00	2.68	3.54	0.06	0.07
RNA2	379	14.41	0.00	5.22	11.05	9e-04	0.09
RNA2	380	8.97	0.00	4.78	6.43	0.01	0.08
RNA2	382	4.32	0.00	2.16	3.93	0.05	0.04
RNA2	388	8.05	0.00	3.63	6.53	0.01	0.06
RNA2	395	6.24	0.00	4.59	4.31	0.04	0.08
RNA2	398	4.72	0.00	2.69	4.12	0.04	0.05
RNA2	399	3.41	0.00	1.45	3.30	0.07	0.02
RNA2	401	2.70	0.00	1.35	2.71	0.10	0.02
RNA2	410	4.02	0.00	2.52	3.66	0.06	0.04
RNA2	411	3.17	0.00	1.56	2.79	0.09	0.03
RNA2	420	4.85	0.00	3.60	3.45	0.06	0.06
RNA2	421	3.45	0.00	1.68	2.77	0.10	0.03
RNA2	422	19.48	0.00	10.59	12.66	4e-04	0.18
RNA2	445	6.76	0.00	4.61	5.17	0.02	0.08

RNA2	446	8.71	0.00	3.51	6.43	0.01	0.06
RNA2	463	9.04	0.00	4.22	6.72	0.01	0.07
RNA2	464	5.11	0.00	2.54	4.01	0.05	0.04
RNA2	473	7.24	0.00	3.67	5.09	0.02	0.06
RNA2	496	4.34	0.00	2.64	3.04	0.08	0.09
RNA3	48	2.54	0.00	1.76	3.05	0.08	0.01
RNA3	51	3.71	0.00	2.49	3.61	0.06	0.02
RNA3	121	2.58	0.00	1.97	3.03	0.08	0.01
RNA3	129	2.54	0.00	1.86	2.77	0.10	0.01
RNA3	137	6.89	0.00	5.01	6.05	0.01	0.01
RNA3	146	4.84	0.00	3.27	5.04	0.02	0.01
RNA3	150	3.10	0.00	1.75	3.57	0.06	0.01
RNA3	156	3.10	0.00	1.75	3.34	0.07	0.01
RNA3	157	7.32	0.00	5.41	6.24	0.01	0.01
RNA3	171	5.96	0.00	4.60	5.84	0.02	0.01
RNA3	222	2.40	0.00	1.77	2.81	0.09	0.01
RNA3	231	2.90	0.00	2.02	3.31	0.07	0.01
RNA3	244	4.01	0.00	2.79	4.14	0.04	0.01
RNA3	264	3.43	0.00	2.16	3.70	0.05	0.01
RNA3	270	4.80	0.00	3.92	3.43	0.06	0.02
RNA3	271	10.97	0.00	7.88	7.22	0.01	0.03
RNA3	295	5.63	0.00	3.91	5.32	0.02	0.01
RNA3	298	5.93	0.00	4.21	5.37	0.02	0.01
RNA3	308	3.01	0.00	1.79	3.21	0.07	0.01
RNA3	309	3.12	0.00	2.66	2.71	0.10	0.01
RNA3	311	8.29	0.00	6.16	6.71	0.01	0.02
RNA3	312	5.25	0.00	3.28	6.21	0.01	0.01
RNA3	318	2.53	0.00	1.50	3.12	0.08	0.00
RNA3	320	4.11	0.00	3.10	3.25	0.07	0.01
RNA3	323	3.38	0.00	2.88	2.87	0.09	0.01
RNA3	324	3.05	0.00	2.18	3.55	0.06	0.01
RNA3	330	2.89	0.00	2.02	3.64	0.06	0.01
RNA3	336	3.14	0.00	2.34	3.45	0.06	0.01
RNA3	344	2.33	0.00	1.36	3.29	0.07	0.00
RNA3	346	5.02	0.00	4.02	4.91	0.03	0.01
RNA3	356	2.31	0.00	1.81	2.76	0.10	0.01
RNA3	359	6.11	0.00	4.43	5.74	0.02	0.01
RNA3	366	2.53	0.00	1.50	3.12	0.08	0.00
RNA3	367	2.78	0.00	1.60	3.31	0.07	0.00
RNA3	369	2.94	0.00	2.14	3.72	0.05	0.01
RNA3	371	3.65	0.00	2.17	4.62	0.03	0.01
RNA3	376	5.38	0.00	4.25	5.44	0.02	0.01
RNA3	378	3.73	0.00	2.67	3.82	0.05	0.01
RNA3	379	4.39	0.00	3.52	3.53	0.06	0.01
RNA3	382	3.43	0.00	1.92	3.50	0.06	0.01
RNA3	386	2.76	0.00	1.59	3.30	0.07	0.00
RNA3	387	4.89	0.00	3.42	4.59	0.03	0.01
RNA3	391	3.90	0.00	2.70	3.98	0.05	0.01
RNA3	396	12.39	0.00	9.38	7.51	0.01	0.02
RNA3	413	2.46	0.00	1.75	2.83	0.09	0.01
RNA3	417	10.43	0.00	7.63	8.60	0.00	0.02
RNA3	418	6.66	0.00	4.81	6.22	0.01	0.01
RNA3	429	4.14	0.00	2.56	4.83	0.03	0.01
RNA3	61	25.09	0.00	11.51	2.80	0.09	0.43
RNA3	129	31.26	0.00	17.89	2.99	0.08	0.66
RNA3	271	32.41	0.00	17.76	3.18	0.07	0.66
(1) Synonymous substitution rate at a site							
(2) Non-synonymous substitution rate at a site							
(3) The rate estimate under the neutral model							
(4) Likelihood ratio test statistic for beta = alpha, versus beta /= alpha							
(5) Asymptotic p-value for evidence of selection							
(6) The total length of branches contributing to inference at this site, used to scale dN-dS							