Using bioinformatic tools to mitigate vector-borne disease

by

Edward John Bird

B.S., Oklahoma State University, 2019

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Interdepartmental Genetics
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2025

Abstract

Vector-borne diseases pose significant threats to both livestock and human health, yet our understanding of vector-pathogen interactions remains limited in certain systems due to inadequate genomic resources and surveillance methodologies. This dissertation addresses these gaps through integrated genomic, transcriptomic, and metagenomic approaches applied to veterinary and medically important arthropod vectors. We present a chromosome-scale genome assembly of *Culicoides sonorensis*, a North American vector of several livestock diseases, including bluetongue virus. Using third-generation sequencing technologies, we assembled three chromosomes representing 99.4% of the nuclear genome, with comprehensive gene annotations based on RNA-seq data from each life stage. Comparative genomics revealed that *Culicoides* represent a unique clade of hematophagous vectors and identified lineage-specific gene expansions. We then leveraged this genomic resource to examine transcriptomic responses of *C. sonorensis* to bluetongue virus serotype 17 (BTV-17) and vesicular stomatitis New Jersey virus (VSNJV) infections. Differential expression analysis revealed distinct virus-specific responses, with BTV-17 eliciting a stronger transcriptional response than VSNJV. Pathway analysis showed that genes associated with abnormal behavior were differentially expressed in response to BTV-17 infection, as well as components of the innate immune system, in response to both BTV-17 and VSNJV. This analysis also identified several genes that may play crucial roles in vector competence. Lastly, we utilized these genomic analysis technologies to develop a xenosurveillance framework using house flies (*Musca domestica*) as bioindicators for environmental antimicrobial resistance and pathogen monitoring, demonstrating location-specific resistome profiles between urban and rural environments. Together, these advances provide essential genomic foundations for vector biology research, reveal novel insights into vector-

pathogen interactions, and establish scalable surveillance methodologies for monitoring emerging threats in diverse environmental contexts.

Using bioinformatic tools to mitigate vector-borne disease

by

Edward John Bird

B.S., Oklahoma State University, 2019

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Interdepartmental Genetics
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2025

Approved by:

Co-Major Professor
Dana Nayduch

Approved by:

Co-Major Professor
Kristopher Silver

Copyright

Abstract

Vector-borne diseases pose significant threats to both livestock and human health, yet our understanding of vector-pathogen interactions remains limited in certain systems due to inadequate genomic resources and surveillance methodologies. This dissertation addresses these gaps through integrated genomic, transcriptomic, and metagenomic approaches applied to veterinary and medically important arthropod vectors. We present a chromosome-scale genome assembly of *Culicoides sonorensis*, a North American vector of several livestock diseases, including bluetongue virus. Using third-generation sequencing technologies, we assembled three chromosomes representing 99.4% of the nuclear genome, with comprehensive gene annotations based on RNA-seq data from each life stage. Comparative genomics revealed that *Culicoides* represent a unique clade of hematophagous vectors and identified lineage-specific gene expansions. We then leveraged this genomic resource to examine transcriptomic responses of *C. sonorensis* to bluetongue virus serotype 17 (BTV-17) and vesicular stomatitis New Jersey virus (VSNJV) infections. Differential expression analysis revealed distinct virus-specific responses, with BTV-17 eliciting a stronger transcriptional response than VSNJV. Pathway analysis showed that genes associated with abnormal behavior were differentially expressed in response to BTV-17 infection, as well as components of the innate immune system, in response to both BTV-17 and VSNJV. This analysis also identified several genes that may play crucial roles in vector competence. Lastly, we utilized these genomic analysis technologies to develop a xenosurveillance framework using house flies (*Musca domestica*) as bioindicators for environmental antimicrobial resistance and pathogen monitoring, demonstrating location-specific resistome profiles between urban and rural environments. Together, these advances provide essential genomic foundations for vector biology research, reveal novel insights into vector-

pathogen interactions, and establish scalable surveillance methodologies for monitoring emerging threats in diverse environmental contexts.

Table of Contents

List of Figures

Acknowledgements

I would like to acknowledge my past and present advisors who have guided me through my academic journey, especially Dr. Kristopher Silver, Dr. Dana Nayduch, and Dr. Astri Wayadande. I would also like to thank my committee members and all the colleagues I have worked with at both KSU and the USDA. I would also like to thank the K-State Entomology department for being a great host to my cross-disciplinary program. Lastly, I would like to thank my family and friends for supporting me through this time, especially my wife Victoria Pickens.

This work was funded by collocative research with the Arthropod-Borne Animal Diseases Research Unit, ARS Project numbers 3020-32000-018-00D, 3020-32000-020-00D, and 3020- 32000-019-00D. This research also used computational resources provided by the SCINet project of the USDA Agricultural Research Service, ARS project numbers 0201-88888-003-00D and 0201-88888-002-00D. Additionally, some of the computing for this project was performed on the Beocat Research Cluster at Kansas State University, which is funded in part by NSF grants CNS-1006860, EPS-1006860, EPS-0919443, ACI-1440548, CHE-1726332, and NIH P20GM113109.

Chapter 1 - Application of One Health and FAIR Principles to Understudied Vectors

In preparation as: Edward Bird, Dana Nayduch, and Kristopher Silver. Application of One Health and FAIR Principles to Understudied Vectors

Abstract

Vector-borne diseases (VBDs) pose an enormous global health challenge, accounting for an estimated 17% of all human infections worldwide. VBD is a multifaceted problem involving interactions between vectors, pathogens, hosts, and environmental factors, making it challenging to address. One Health recognizes the interconnectedness of human health to animals and the environment we share and encourages communication and collaboration between researchers. This approach naturally leads to interdisciplinary research and questions about how these different factors combine to affect pathogen transmission and disease spread by vectors. Researchers often pursue omics data (e.g., genomics, transcriptomics, proteomics, metabolomics, etc.) to better understand the underlying molecular processes that contribute to the biology of vectors and generate new hypotheses. Omics data often requires extensive processing to gain relevant insights, but analytical workflows used to process the data often produce results that are not always reproducible. While these datasets are still useful, they are often not curated according to Findable, Accessible, Interoperable, and Reusable (FAIR) principles, thus limiting their utility to other researchers and future research initiatives. In this review, we discuss how the use of omics and other big data types has already furthered the knowledge of VBD, and the future potential of these analyses in two understudied vector systems; *Culicoides* biting midges and house flies. In addition, we discuss how the implementation of One Health approaches and

reproducible FAIR data analytic and stewardship practices will enhance the reliability and reusability of datasets to further our understanding of VBD in these systems.

Introduction

Vector-borne diseases (VBDs) represent one of the most complex and pressing challenges in public health, affecting human and animal populations globally. VBD accounts for approximately 17% of all human infections worldwide, resulting in an estimated 700,000 deaths annually (Manikandan et al., 2022; Aditi Kulkarni et al., 2023). In addition, vector-borne pathogens caused 28.8% of all newly identified emerging infectious diseases (EIDs) between 1998 and 2008 (Jones et al., 2008). These EIDs include significant human health threats, such as dengue virus, West Nile virus, Lyme disease, and many others (CDC, 2025; Petersen & Roehrig, 2001; Rico-Hesse, 2007; Stone et al., 2017). Beyond their immediate health impacts, these diseases impose severe economic burdens, causing over $100 billion USD in losses annually to livestock systems alone (Harvesting Agriculture's "Natural" Insect Farms: ARSX 2021 Winning Project : USDA ARS, 2021).

Aside from these classical mosquito and tick vectored diseases, filth flies (house flies, blow flies, and flesh flies), play key roles in the emergence of antimicrobial resistant (AMR) bacteria (Bogri et al., 2024; Nayduch et al., 2023). AMR is one of the top global threats to human health, being responsible for millions of deaths per year (Antimicrobial Resistance, 2023). It is currently estimated that AMR will cause more than 10 million deaths in the year 2050 (GBD 2021 Antimicrobial Resistance Collaborators, 2024). Insects contribute to the spread and dispersal of AMR bacteria as mechanical vectors that can move long distances where resistance subsequently can be transferred to pathogenic bacteria. Insects, like house flies, also contribute more directly to the emergence of AMR, facilitating horizontal gene transfer (HGT) in

their gut (Akhtar et al., 2009; Fukuda et al., 2016). This is just one example of the complicated nature of the evolution of AMR and the role insects such as house flies might play.

The complexity of infectious disease patterns is fundamentally linked to the environments we share with other animals, plants, and broader ecological systems. This interconnectedness of factors affecting transmission and the prevalence of VBD is the bedrock principle of One Health approaches to mitigate their impact on human and animal health. Consequently, One Health strategies necessitate multidisciplinary approaches that integrate diverse data types including genomic sequences, transcriptomic profiles, climate data, clinical case records, and field observations. Additionally, integration of these diverse data sets increasingly relies on computational approaches and bioinformatic analyses to understand pathogen-vector interactions, model disease emergence, and develop control strategies. Whereas these analyses hold great promise for helping to solve complex issues surrounding VBD, much of the data generated in these studies remains inaccessible or unusable after publication. This is due to research progress being hampered by the lack of Findable, Accessible, Interoperable, and Reusable (FAIR) data practices or reproducible bioinformatic analyses in vector-borne disease research. For example, analytical workflows are poorly documented or standardized, limiting the reproducibility and reusability of research findings. This problem is of particular importance in understudied vector systems where limited resources compound the challenges of data sharing and analytical standardization.

This review emphasizes the critical need for standardized bioinformatic pipelines, accessible data repositories, and reproducible analytical frameworks to accelerate our understanding of vector-pathogen interactions and improve disease surveillance and control strategies. We use two understudied vector-borne disease systems to highlight both the potential

and the challenges of modern computational approaches in vector biology. *Culicoides* midges, despite their critical role in transmitting economically important diseases like bluetongue virus (BTV) and epizootic hemorrhagic disease virus (EHDV), suffer from limited genomic resources that hinder comprehensive understanding of mechanisms of vector competence. Similarly, house flies represent both an emerging concern as mechanical vectors of AMR bacteria, as well as an opportunity to utilize their biology to study the distribution of AMR pathogens in the environment. We illustrate how FAIR data practices and reproducible analytical workflows can accelerate scientific progress in vector-borne disease research in these understudied systems.

Vector-Borne Disease

The One Health Approach

VBD is affected by numerous interconnected determinants including human and animal host populations, vector distribution patterns, climate, and more. Acknowledgement of the interconnectedness among environmental, livestock, and human health systems has assumed increasing significance in recent years, with the formulation and implementation of One Health strategies (Pitt and Gunn 2024). The One Health framework promotes research to recognize that patterns of infectious disease are linked with the environments we share as well as our interactions with other animals, plants, and wider environmental systems. These foundational principles enable the integrated approaches necessary to address complex health challenges like VBD.

Communication, coordination, collaboration, and capacity building are the core principles in One Health approaches. Communication could be considered the foundation of One Health approaches and necessitates the creation of effective channels of information exchange between human health practitioners, veterinarians, environmental scientists, and other actors (One Health

High-Level Expert Panel (OHHLEP) et al., 2022). Effective communication guarantees that surveillance data, research results, and control measures are transmitted smoothly across disciplines and sectors. Coordination entails the structured arrangement of activities across various sectors to prevent duplication of research and ensure effective use of resources. This includes coordinating surveillance systems, research agendas, and response mechanisms to enhance overall impact. Collaboration builds upon coordination by promoting active partnerships and collective problem solving across historically distinct spheres. Successful collaboration involves assembling teams of diverse experts, including epidemiologists, microbiologists, entomologists, veterinarians, climatologists, and policy makers to collectively solve common goals. Lastly, capacity building aims at creating the competencies, knowledge, and infrastructure required to support One Health approaches in the long term, such as training/education, technology transfer, and institutional strengthening. Relevant to this review, capacity building may also include developing standardized bioinformatic tools and genomic resources that can be shared among research groups.

Embracing One-Health approaches to enhance research of VBD naturally leads to multidisciplinary research that relies strongly on diverse data sources and types. The ability of researchers to identify and recycle available scientific data is a key facilitator of One Health research. Studies trying to explain intricate phenomena, such as the geographical expansion of dengue fever, have used a wide array of data sources, including genomic data, transcriptomic data, climate data, and clinical human case data in combination with laboratory and field observation studies (Morin et al. 2013, Sim and Hibberd 2016, Barcellos et al. 2024, Xu et al. 2024). Much of the data used in such integrative research is not generated *de novo* but rather repurposed from past studies. Some datasets are too complex to fully explore in a single study, as

seen with population data sets, or are combined with other datasets to gain deeper insights (Kraemer et al., 2015, 2019; Uelmen et al., 2023). Often there is a balance between generating *de novo* data and reusing previously generated data. For example, studies that rely on genomic references, like analysis of gene expression, are reliant on existing genome assemblies and gene annotations to analyze expression patterns of the newly generated data. Some domains, particularly in the realm of human medical research, have adopted standards and practices to enhance data reuse (Margolis et al. 2014), but many scientific disciplines still have their own types of data and different standards for analysis, despite initiatives aimed at enhancing data stewardship practices.

The Importance of omics Datasets in One Health

Arguably the most valuable and recyclable data in One Health research are omics datasets. Researchers have generated many omics datasets from vector species to answer specific questions and generate new hypotheses. For example, metagenomic studies have been used to reveal the incredible diversity of pathogens within tick populations, discovering the presence of *Rickettsia tarasevichiae* in *Dermacentor nuttalli* in China for the first time (Jiao et al., 2021). In addition, metabolomic investigations have revealed changes in utilization of metabolic pathways during feeding in ticks infected with *Borrelia* spp. (Hoxmeier et al., 2017). These and other omics studies contribute to vector biology knowledge, which is central to the understanding of pathogen transmission mechanisms and the development of successful management strategies. These data have been instrumental in enhancing our knowledge of the interactions between vectors and pathogens; however, a deeper understanding can only be achieved through the combination of multiple datasets and perspectives.

The combination of several types of omics techniques can reveal biological processes that may otherwise go unnoticed when examined individually. An example of this is research which explored the effects of climate change on the vector capacity of *Aedes albopictus* to transmit Chikungunya virus (CHIKV) (Bellone et al. 2023). This study applied transcriptomics, microbiome profiling of the gut, and viral genomics to investigate viral evolution and climate effects on viral replication in the vector. They used available genomic tools including the *Aedes albopictus* genome and annotation information for differential gene expression analysis. Increased temperatures led to significant modulation of the microbiome of the mosquito carcass (body without gut) and increased rates of viral evolution in mosquito cells associated with the increased temperature. Changes in the mosquito microbiome, particularly increased *Serratia spp.* abundance, were expected to enhance vector competence through disruption of the gut membrane, whereas increased rates of CHIKV evolution could be a sign of adaptation to higher temperature conditions. The findings showed relationships between climate data, vector competence, and viral evolution that would not have been obvious without the combination of multiple data sources. Multi-omics analyses like this show great potential for furthering the understanding of vector-pathogen interactions, and consequently human health, but their success relies on the availability of previously generated, easily accessible datasets, highlighting how important proper data stewardship is to vector-borne research.

FAIR Practices

Large datasets such as climate data and omics data have become more commonplace in vector-borne disease research over the past decade due to reduced production costs (Pećina-Šlaus & Pećina, 2015). The utilization of such data offers promise for One Health strategies but also enhances the urgent need for standardized practices in data stewardship and sharing (Scott et al.,

2023). The exponential increase in data production renders the adoption of FAIR data principles not only advantageous but necessary to optimize the scientific return on these datasets. FAIR practices offer guidelines to help ensure that the data being produced can adequately serve the collaborative, multidisciplinary research that One Health requires. When properly employed, these practices lead to the sharing of not just scientific knowledge but also the scientific data used to inform this knowledge. Often in biological research, omics data (genomic, transcriptomic, proteomic, etc.) can be utilized to test more than a single hypothesis or can be supplemented with new data to gain deeper insights. This concept is visualized in Figure 1.1, with four research groups (1-4) not only building upon previous scientific knowledge but also integrating available scientific data. While researchers that do not follow these guidelines still contribute their findings to the general scientific knowledge in the field, their scientific data is not available for secondary exploration by other groups, reducing the efficiency of our scientific inquiry and increasing the risk of data loss (Figure 1.1, research group 4). In VBD, particularly in understudied systems, every dataset represents a wealth of information critical to progress in the research field that can be lost without proper data stewardship.

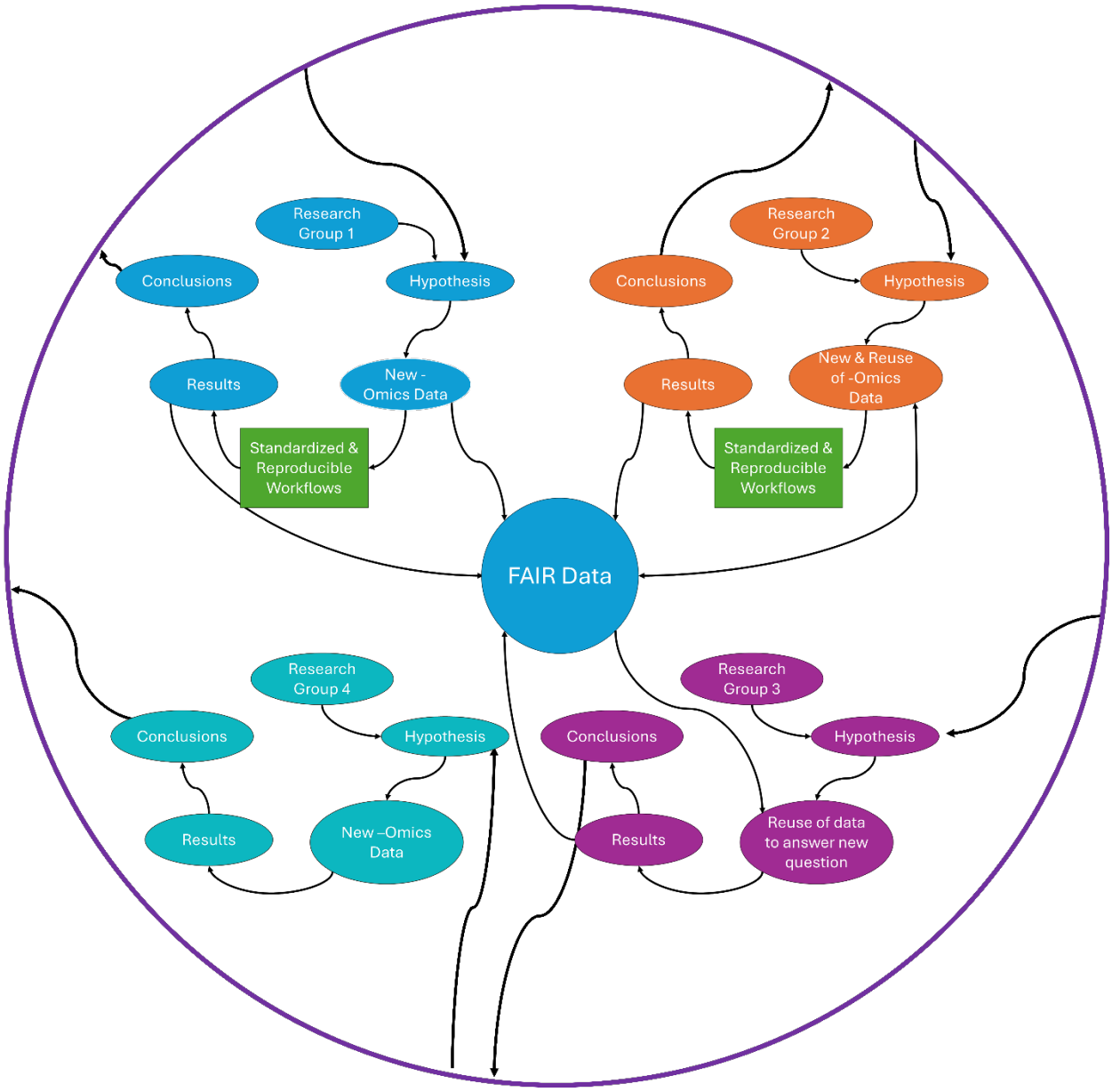


Figure 1.1 One Health approach to FAIR data management in vector biology research.

The purple circle represents the pool of scientific knowledge about VBD. Research Group 1 illustrates researchers that generate new data and practice FAIR data practices. Research Group 2 & 3 represent different ways researchers can utilize and contribute to FAIR data to further VBD research. Group 4 represents how scientific data can be isolated and/or lost if not managed properly.

The FAIR principles constitute a holistic data stewardship approach that tackles the root issues associated with data intensive research. Scientific data needs to be findable and accessible,

with unique and persistent identifiers, and registered in searchable databases, so that researchers can locate and retrieve relevant datasets (Mark D. Wilkinson et al., 2016, 2019). Once data is found, it must also be interoperable and reusable. Data need to be described with rich metadata, with common language that is machine readable. Data must also be reusable, and in a standardized format and units so it is ready for reuse without intensive pre-processing. These practices are essential for One Health research that crosses human, animal, and environmental health domains using large data sets. In the absence of standardized data organization, annotation, and sharing practices, these datasets become less useful to One Health research. As the size and complexity of omics datasets continue to grow, the application of FAIR practices becomes even more urgent.

Application of One Health and FAIR Principles to Understudied Vectors

While some vector systems, such as *Aedes* and *Culex* mosquitoes, have been extensively studied, others like *Culicoides* biting midges and house flies (*Musca domestica*) remain comparatively underexplored, despite their global distribution and importance in pathogen transmission. The following sections identify the gaps in our knowledge of these understudied vector systems and highlight how integrated One Health approaches may lead to new knowledge and strategies for mitigating the risks of VBD posed by these insects.

Culicoides Midges

Culicoides biting midges (Diptera: Ceratopogonidae) represent a diverse group of hematophagous insects that are globally distributed (Mullens et al. 2015). More than 1,347 *Culicoides* species have been described across the globe, but just 30 have been incriminated as vectors of bluetongue virus (BTV) (Borkent and Grogan, Jr. 2009; Meiswinkel et al. 2004). Similarly, in the United States, there are 141 recognized *Culicoides* species; however, only two

species, *Culicoides sonorensis* and *Culicoides insignis*, have been confirmed as BTV vectors (Jones et al. 1977, Tanya et al. 1992), implying that specific genetic or physiological characteristics dictate vectorial capacity. Of these two species, the majority of research has focused on *C. sonorensis*, as it is the only *Culicoides* species for which a stable lab colony and cell line are available (McHolland and Mecham 2003, Nayduch et al. 2014, Erram and Burkett-Cadena 2021). Whereas *C. sonorensis* is the primary model organism for biting midge research in the United States, our knowledge of vector competence or the effects of arboviral infection on vector biology is lacking.

Disease Transmission

C. sonorensis is a confirmed vector of three arboviruses in the US: Epizootic hemorrhagic disease virus (EHDV), BTV, and vesicular stomatitis virus (VSV) (Jones et al. 1977, Ruder et al. 2017, Rozo-Lopez et al. 2018) and therefore plays a crucial role in animal health. BTV is an orbivirus that is typically transmitted only by *Culicoides* spp.**;** however, certain strains of BTV present outside the U.S. are occasionally transmitted from animal to animal through vertical transmission or direct contact (van der Sluijs et al. 2016, Thabet and Lajnef 2024). BTV affects a wide range of domestic and wild ruminants; primarily sheep, cattle, and deer. Although often subclinical in cattle, clinical disease in sheep and deer includes tongue and facial swelling, lameness, abortion, fever, and mortality (MacLachlan 1994). BTV has caused billions (USD) of economic losses worldwide, with a single outbreak of BTV-8 in Europe causing an estimated three billion dollars in economic loss (Häsler et al. 2012, Rushton et al. 2015, Alkhamis et al. 2020, Gethmann et al. 2020). EHDV, a related orbivirus, also infects domestic and wild ruminants; primarily cattle and deer. While disease is typically mild in cattle, severe clinical disease is seen in deer with respiratory problems, lameness, swelling, sores of the

mouth, fever, weakness, and death (Newcomer et al. 2021). EHDV is particularly devastating when introduced into naïve populations of white-tailed deer, causing significant mortality (Stallknecht and Howerth 2004).

VSV is a zoonotic virus in the family *Rhabdoviridae* that can be transmitted by a range of hematophagous arthropods, including *Culicoides* midges, mosquitoes, black flies, and sand flies (Rozo-Lopez et al. 2018). Though not considered endemic to the United States, the virus periodically enters from endemic areas in Central and South America, resulting in sporadic outbreaks. These outbreaks often result in quarantines and movement restrictions on affected livestock operations to limit spread (Pelzel-McCluskey et al. 2021). VSV infects several domesticated species, including horses, cattle, pigs, sheep, and goats. Clinically, VSV can resemble foot-and-mouth disease in cattle and pigs, as it causes vesicular lesions in and around the mouth, lameness, anorexia, and weight loss, thereby complicating diagnosis and response efforts (Alderink 1984; Hayek et al. 1998). The economic and regulatory effects of VSV outbreaks can be substantial due to the similarities with other high-consequence diseases and the disruption of animal movement and trade.

Pathogen-Vector Interactions and Behavioral Changes

Arboviral infections in *Culicoides* midges result in complex pathogen-vector interactions that affect multiple physiological systems. Previous work has shown that many tissues are invaded by virus during infection with EHDV and BTV, including the cephalic ganglia, thoracic ganglia, and ommatidia (Fu et al. 1999a, McDermott et al. 2015, Mills et al. 2017). VSV similarly causes systemic infection patterns in *C. sonorensis* tissues, with the virus infecting most tissues, including the salivary glands and ovaries (Drolet et al. 2005) and can be maintained in adult midge populations by venereal transmission (Rozo-Lopez et al. 2020). Systemic

infection of midge hosts by these arboviruses could lead to physiological and/or behavioral changes; however, these effects are still poorly understood in these *Culicoides*.

Several lines of evidence suggest that infection of insect hosts with arboviruses cause changes in vector behavior (Gaburro, Bhatti, et al. 2018, Gaburro, Paradkar, et al. 2018, Tallon et al. 2020, Wei Xiang et al. 2022). Infection of *Aedes aegypti* by Zika virus and Dengue virus 2 increases locomotor and blood feeding activity levels (Gao et al. 2025). In *Culicoides* midges, McDermott et al. observed that UV light-based traps captured significantly fewer BTV-infected midges than CO_2-baited traps, hypothesizing that infection of the ommatidia changed phototactic responses to the UV light. Subsequent transcriptomic profiling of EHDV infected *C. sonorensis* also revealed 122 unigenes linked to vision, olfaction, sensory, brain function, behavior, memory, and learning were downregulated in response virus infection (Nayduch et al. 2019), suggesting that EHDV altered neuro-sensory organs. Changes to the visual and behavioral habits of vectors could alter host seeking or biting preferences thereby affecting dynamics of vector competence. Changes could also diminish the effectiveness of trapping methods, thereby affecting surveillance and disease risk assessments (Hoffmann et al. 2009, Goffredo et al. 2013).

Vector Competence and Genetic Factors

Previous work analyzed differential expression of transcripts to determine key genetic factors affecting vector competence in *C. sonorensis*; however, there are still many gaps in our knowledge. Early EHDV infection (36 h) was accompanied by upregulation of the insect innate immune system such as components of the Toll, Imd, and JAK/STAT, and ProPO pathways in midges, and including effectors like antimicrobial peptides (Nayduch et al. 2019). In contrast, during infection with VSV, only a transient upregulation of some immune components was observed in the *C. sonorensis* cell line at one hour post-infection (Scroggs et al. 2023). These

results suggest that *Culicoides* responses to infection may vary with the arbovirus that is infecting them, although it is important to note that the cell line may not fully replicate the complex, systemic immune responses that occur in live midges.

Other studies have provided insights into the mechanisms of vector competence in *C. sonorensis*. One gene, *glutathione S-transferase* 1 (*gst*-1), has long been implicated in the vector competence for BTV in *C. sonorensis* (Jones and Foster 1974, Fu et al. 1999b, Abdallah et al. 2000). Expression of *gst*-1 differed between BTV susceptible and refractory lines of *C. sonorensis* (Morales-Hojas et al. 2018), and individuals with high levels of *gst-1* expression were more refractory to BTV infection, suggesting the role of this gene product in vector competence. In addition to *gst-1*, the upregulation of the antiviral helicase, *ski2,* was also associated with refractoriness to BTV infection in midges. These two genes represent the largest insights thus far into the molecular mechanisms of vector competence in *C. sonorensis*, with no larger genetic elements, like gene clusters, being associated. However, neither gene has been determined to definitively play a role in midge vector competence for BTV in experimental studies.

Genomic Resource Limitations

One major limitation to vector competence studies in *C. sonorensis* is the lack of high-quality genomic resources. The currently available reference genome for *C. sonorensis* was generated using short-read sequencing and consequently is highly fragmented (Morales-Hojas et al. 2018), which limits its utility in many ways. Fragmentation affects gene prediction, resulting in predicted gene models often being truncated, split, or incomplete (Denton et al. 2014). Analyses that depend on gene positioning, such as quantitative trait locus (QTL) mapping, are also not possible with the current assembly. Overall, the fragmented genome greatly limits the experiments that can be performed, excluding broad-scale genomic analyses necessary to

establish the molecular mechanisms of vector competence in this disease vector of economic importance.

Recent advances in long-read and scaffolding technologies have made the generation of chromosome-level assemblies from individual insect specimens increasingly feasible. Other small insects, like *A. aegypti* (Fisher et al., 2022), *Phlebotomus papatasi,* and *Lutzomyia longipalpis* (Huang et al., 2024) have had chromosome-level assemblies produced from single individuals. With a high-quality, chromosome-level genome for *C. sonorensis*, researchers could begin to characterize genomic features associated with vector competence, such as structural variants, gene family expansions, or regulatory elements with greater resolution. Such a resource would not only improve our understanding of the genetic basis of vector competence but would also enable investigation of other biological processes including anautogeny, reproductive biology, olfactory mechanisms, and host-seeking behaviors, all of which could inform the development of targeted control strategies, surveillance tools, and management approaches for *Culicoides* midges. Many other sequence datasets are likely to be available to compare with *C. sonorensis* but remain difficult to discover or access by researchers, due to inconsistent data sharing practices and lack of standardized formats. This limited access represents a barrier to VBD research in this understudied organism where the loss of any dataset represents a setback in research progress.

Need for One Health and FAIR Standards in Culicoides *Research*

The future of *Culicoides* research can dramatically improve through the adoption of One Health principles and FAIR data practices, both of which are essential for holistically tackling the complex interactions between vectors, pathogens, host populations and environmental factors such as climate variability (Borycz et al., 2020; Delpy et al., 2024; Emmanuel Ajibola Olagunju

et al., 2023; White & Hughes, 2019). The adoption of FAIR data practices will ensure the curation and preservation of important datasets, such as genome and protein annotations, transcriptomic profiles, climate data, field population surveillance records, etc., and ensure access for future research. This will allow for more integrated approaches, incorporating previous data to understand the mechanisms that control vector competence and behavior. This collaborative, multidisciplinary approach will help characterize genetic determinants underlying vector competence, inform the development of targeted control measures, and ultimately enhance our ability to predict and mitigate the economic impacts of *Culicoides*-transmitted diseases like BTV and EHDV.

Antimicrobial Resistance and House Flies

Antibiotics play a crucial role in treating bacterial infections and have drastically reduced mortality in both human and animal health since their introduction. However, the widespread use, misuse and overuse of antibiotics has also led to resistance to these drugs. Shortly after the mass introduction of antibiotics in the 1940s, antimicrobial resistance (AMR) was observed, and genes associated with resistance to antibiotics were also observed being transferred horizontally between bacteria, enhancing the spread of resistance phenotypes (Watanabe 1963, Davies and Davies 2010). Evolution and emergence of AMR has only increased over time due to overuse and the stagnation of novel antibiotic discovery (Spellberg and Gilbert 2014, Skender 2024). The rise of AMR is largely the result of the overuse and misuse of antibiotics in human medicine, veterinary medicine, and agricultural systems (Granados-Chinchilla and Rodríguez 2017, Akram et al. 2023). As a result, human health is currently faced with a rising tide of bacteria that are resistant to multiple drugs, including some of the most reserved antibiotics (Shah et al. 2007, Painuli et al. 2023).

The problem of AMR has become an urgent global health crisis (Ventola 2015, Salam et al. 2023), with AMR bacteria causing more than one million deaths in 2021, and these bacteria being linked to an additional four million deaths in the same year. Current estimates indicate that AMR will lead to more than 90 million fatalities from 2025 to 2050 if present trends persist (GBD 2021 Antimicrobial Resistance Collaborators 2024). This crisis threatens a return to a time when simple bacterial infections could be fatal (CDC Washington Testimony March 25, 2015). AMR bacteria have been identified in a wide range of settings outside of the human health sphere, including agricultural and environmental settings. Contamination of antibiotics from sources like human and animal waste, antibiotic disposal, and pollution from antibiotic production all play key roles in AMR evolution and maintenance in the environment (Larsson et al. 2007, Anwar et al. 2020, Wang et al. 2020) and serve as the major driver of environmental AMR (Allen et al. 2010). The environment also provides an increasingly important reservoir for the spread of AMR, as several ecological niches allow resistance genes to be conserved and spread. Additionally, the ability of ARGs in the environment to find their way back into the human health sphere through horizontal transfer, direct human/animal contact, or through other sources like raw vegetables (Larsson and Flach 2022) highlights the seriousness of this issue.

Though overuse and misuse of antibiotics has contributed significantly to the widespread development of AMR, it has been theorized that limiting antibiotic use is no longer sufficient to halt the emergence of new resistance genes (McEwen and Collignon 2018). ARGs are often encoded on mobile genetic elements (MGEs) that can easily undergo HGT from one organism to another. In some cases, HGT can occur between different species of bacteria, leading to the appearance of novel modes of resistance across taxa. Due to these phenomena and the high prevalence of ARGs on MGEs, it has been hypothesized that the primary driver of the AMR

crisis we face today is the HGT of ARGs carried on MGEs (Collignon et al. 2018). As a result of the complexity of the movement, evolution, and maintenance of ARGs, AMR has been recognized as requiring a One Health approach to successfully understand and manage (White and Hughes 2019, Velazquez-Meza et al. 2022). With the rapid spread and development of AMR in the 21st century, surveillance of AMR and ARGs by monitoring programs such as the National Antimicrobial Resistance Monitoring System (NARMS) and the European Antimicrobial Resistance Surveillance Network (EARS-Net) are critical to remaining ahead of both current and emerging microbial threats. These programs enable researchers and public health workers to identify emergence of new resistance determinants early, trace transmission routes between human, animal, and environmental reservoirs, review the efficacy of intervention efforts, and employ preventative measures before new resistance patterns become commonplace (Delpy et al. 2024, Franklin et al. 2024, Karp et al. 2017, Simonsen 2018). Whereas surveillance of AMR and ARGs from a multitude of environments is crucial to the success of these programs, sampling is often limited in scope due to the diversity of sampling required. The development of standardized methods to comprehensively survey AMR bacteria will be crucial to understand and address the crisis we face.

House Flies as AMR Vectors and Potential Tools for AMR and Pathogen Surveillance

Filth-associated insects like flies and cockroaches represent a threat to human and animal health by harboring pathogenic and AMR bacteria (Bogri et al. 2024, Thomson et al. 2017, Nayduch et al. 2023, Neupane et al. 2020, Neupane et al. 2023). House flies are particularly important due to their global distribution across and proclivity for occupying microbe-rich niches in both agricultural and urban environments (Geden et al. 2021, Nayduch et al. 2023). House flies actively seek filth for both food and reproduction, bringing them into contact with a plethora

of microbial taxa. House fly larvae require bacteria for successful development, and adult house flies opportunistically utilize these same sources to acquire protein necessary for egg development (Moon, 2002; Zurek et al., 2000). When interacting with these microbe-rich substrates, they become burdened with various bacterial taxa through direct contact, ingestion, or grooming behaviors (Nayduch & Burrus, 2017). After acquiring these bacteria, house flies can act as mechanical vectors by spreading them to other environments (Ghosh & Zurek, 2015; Yap et al., 2008).

House flies represent a snapshot of the sources they visit, picking up a wide range of bacterial taxa (Neupane & Nayduch, 2022; Neupane et al., 2023) due to their ability to carry diverse and representative samples of microbial taxa from visited sources. Research has demonstrated that flies not only acquire microbes from substrates like manure but also carry taxa directly associated with the livestock, highlighting how they capture microbial diversity from multiple sources simultaneously (Neupane et al. 2024). House flies must visit microbe-rich substrates for survival, and females are particularly attracted to protein-rich sources including animal excretions, wounds, eyes, and mucous membranes (Nayduch et al. 2023). This attraction extends to sick animals, further enhancing their potential as sentinels for pathogenic and resistant bacteria. These qualities make house flies potentially ideal sentinels to survey the environment for AMR. Initial explorations of this concept have been performed through metagenomic analyses of house flies collected from human environments. Investigation revealed a broad diversity of ARGs (Yang et al., 2024), but several underlying questions remain about the utility of house flies for AMR surveillance. It is unclear to what degree ARGs can be linked to specific bacterial taxa for the detection of health risks or if ARGs are located on MGEs facilitating HGT

and resistance spread. Examining these questions is important for determining if house flies can be reliable indicators for environmental AMR risk and gene mobility assessment.

The use of house flies as xenosurveillance tools for antimicrobial resistance monitoring represents a promising but data-intensive approach that requires robust computational infrastructure and standardized methodologies. This methodology depends heavily on developing reproducible bioinformatic pipelines capable of accurately identifying ARGs to link them to specific bacterial taxa and determine their association with MGEs. These analyses are important for assessing both health risks and AMR evolution but are unavailable without the development of additional bioinformatic tools and resources. The massive volumes of sequence data and complex secondary annotations generated from these metagenomes necessitate the implementation of FAIR data practices to ensure these datasets remain findable, accessible, interoperable, and reusable for future study and questions. Only through computational standardization and data stewardship can house fly xenosurveillance evolve from a promising concept into a reliable, scalable tool for One Health AMR monitoring and early warning systems.

Conclusions

VBD is an escalating threat to human and animal health systems. From the spread of viral diseases like DENV in humans and BTV in animals by hematophagous vectors, to the spread of AMR bacteria by filth-associated insects like house flies, scientific data plays a key role in studying and understanding these complex systems. Multidisciplinary approaches that incorporate One Health principles are key to addressing VBD problems, as is the ability of researchers to reliably and reproducibly use and share data among groups. Given the increasing use of omics data for studying the biology of vectors as well as the pathogens they transmit, standardizing the analysis and stewardship of these data according to FAIR practices will be

invaluable. This is particularly true in understudied species, like *Culicoides* midges and house flies, where omics datasets are more limited, more difficult to find, and the loss of any dataset would be more damaging to the field. Accordingly, adoption of data storage practices using standardized formats described with rich metadata, and the processing of omics data with reproducible bioinformatics best practices pipelines utilizing workflow managers and containerized environments, will significantly improve the efficiency and productivity of VBD research endeavors, particularly in understudied vector species.

References

Abdallah, M. A., R. S. Pollenz, F. N. Droog, et al. 2000. Isolation and characterization of a cDNA clone coding for a glutathione S-transferase class delta enzyme from the biting midge *Culicoides variipennis sonorensis* Wirth and Jones. Biochem Genet. 38. 11-12. 177-390

Akram, F., M. Imtiaz, and I. U. Haq. 2023. Emergent crisis of antibiotic resistance: A silent pandemic threat to 21st century. Microb. Pathog. 174: 105923.

Akhtar, M., H. Hirt, and L. Zurek. 2009. Horizontal transfer of the tetracycline resistance gene tetM mediated by pCF10 among Enterococcus faecalis in the house fly (*Musca domestica* L.) alimentary canal. Microb. Ecol. 58: 509–518.

Alderink, F. J. 1984. Vesicular stomatitis epidemic in Colorado: clinical observations and financial losses reported by dairymen. Prev. Vet. Med. 3: 29–44.

Alkhamis, M. A., C. Aguilar-Vega, N. M. Fountain-Jones, et al. 2020. Global emergence and evolutionary dynamics of bluetongue virus. Sci. Rep. 10: 21677.

Allen, H. K., J. Donato, H. H. Wang, et al. 2010. Call of the wild: antibiotic resistance genes in natural environments. Nat. Rev. Microbiol. 8: 251–259.

Anwar, M., Q. Iqbal, and F. Saleem. 2020. Improper disposal of unused antibiotics: an often overlooked driver of antimicrobial resistance. Expert Rev. Anti Infect. Ther. 18: 697–699.

Antimicrobial resistance. 2023. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance

Baichoo, S., Y. Souilmi, S. Panji, et al. 2018. Developing reproducible bioinformatics analysis workflows for heterogeneous computing environments to support African genomics. BMC Bioinformatics 19: 457.

Barcellos, C., V. Matos, R. M. Lana, et al. 2024. Climate change, thermal anomalies, and the recent progression of dengue in Brazil. Sci. Rep. 14: 5948.

Bellone, R., P. Lechat, L. Mousson, et al. 2023. Climate change and vector-borne diseases: a multi-omics approach of temperature-induced changes in the mosquito. J. Travel Med. 30: taad062.

Bogri, A., E. E. B. Jensen, A. V. Borchert, et al. 2024. Transmission of antimicrobial resistance in the gut microbiome of gregarious cockroaches: the importance of interaction between antibiotic exposed and non-exposed populations. mSystems 9: e0101823.

Borkent, A., and W. L. Grogan, Jr. 2009. Catalog of the New World biting midges north of Mexico (Diptera: Ceratopogonidae). Zootaxa 2273: 1.

Borycz, J., J. Borycz, B. C. Carroll, and B. C. Carroll. 2020. Implementing FAIR data for people and machines: impacts and implications—results of a research data community workshop. Inf. Serv. Use 40: 71–85.

Bradley, A., B. A. Mullens, E. G. McDermott, et al. 2015. Progress and knowledge gaps in *Culicoides* ecology and control. Vet. Ital. 51: 313–323.

CDC. 2015. Washington testimony March 25, 2015. https://archive.cdc.gov/www_cdc_gov/washington/testimony/2015/t20150325.htm

CDC. 2025. Lyme disease surveillance and data. https://www.cdc.gov/lyme/data-research/facts-stats/index.html

Chen, G., J. C. Ramírez, N. Deng, et al. 2019. Restructured GEO: restructuring Gene Expression Omnibus metadata for genome dynamics analysis. Database 2019: bay145.

Chifanzwa, R., and D. Nayduch. 2018. Dose-dependent effects on replication and persistence of *Salmonella enterica* serovar typhimurium in house flies (Diptera: Muscidae). J. Med. Entomol. 55: 225–229.

Clough, E., T. Barrett, S. E. Wilhite, et al. 2024. NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. Nucleic Acids Res. 52: D138–D144.

Collignon, P., J. J. Beggs, T. R. Walsh, et al. 2018. Anthropological and socioeconomic factors contributing to global antimicrobial resistance: a univariate and multivariable analysis. Lancet Planet. Health 2: e398–e405.

Davies, J., and D. Davies. 2010. Origins and evolution of antibiotic resistance. Microbiol. Mol. Biol. Rev. 74: 417–433.

Delpy, L., C. C. Astbury, C. Aenishaenslin, et al. 2024. Integrated surveillance systems for antibiotic resistance in a One Health context: a scoping review. BMC Public Health 24: 1.

Denton, J. F., J. Lugo-Martinez, A. E. Tucker, et al. 2014. Extensive error in the number of genes inferred from draft genome assemblies. PLoS Comput. Biol. 10: e1003998.

Di Tommaso, P., M. Chatzou, E. W. Floden, et al. 2017. Nextflow enables reproducible computational workflows. Nat. Biotechnol. 35: 316–319.

Erram, D., and N. Burkett-Cadena. 2021. Oviposition of *Culicoides insignis* (Diptera: Ceratopogonidae) under laboratory conditions with notes on the developmental life history traits of its immature stages. Parasit. Vectors 14: 522.

Ewels, P. A., A. Peltzer, S. Fillinger, et al. 2020. The nf-core framework for community-curated bioinformatics pipelines. Nat. Biotechnol. 38: 276–278.

Fisher, C. R., M. Wilson, and J. G. Scott. 2022. A chromosome-level assembly of the widely used Rockefeller strain of *Aedes aegypti*, the yellow fever mosquito. G3 Genes|Genomes|Genetics 12: 11.

Flies, cockroaches and AMR. 2021. Nat. Microbiol. 6: 1213–1214.

Franklin, A. M., D. L. Weller, L. M. Durso, et al. 2024. A one health approach for monitoring antimicrobial resistance: developing a national freshwater pilot effort. Front. Water 6: 1359109.

Fu, H., C. J. Leake, P. P. Mertens, et al. 1999. The barriers to bluetongue virus infection, dissemination and transmission in the vector, *Culicoides variipennis* (Diptera: Ceratopogonidae). Arch. Virol. 144: 747–761.

Fukuda, A., M. Usui, T. Okubo, and Y. Tamura. 2016. Horizontal transfer of plasmid-mediated cephalosporin resistance genes in the intestine of houseflies (*Musca domestica*). Microb. Drug Resist. 22: 336–341.

GBD 2021 Antimicrobial Resistance Collaborators. 2024. Global burden of bacterial antimicrobial resistance 1990-2021: a systematic analysis with forecasts to 2050. Lancet 404: 1199–1226.

Geden CJ, Nayduch D, Scott JG, et al. 2021. House fly (Diptera: Muscidae): biology, pest status, current management prospects, and research needs. Carley D, editor. J. Integr. Pest Manag. 12(1):39.

Gethmann, J., C. Probst, and F. J. Conraths. 2020. Economic impact of a bluetongue serotype 8 epidemic in Germany. Front. Vet. Sci. 7: 65.

Ghosh, A., and L. Zurek. 2015. Fresh steam-flaked corn in cattle feedlots is an important site for fecal coliform contamination by house flies. J. Food Prot. 78: 567–572.

Goffredo, M., F. Monaco, G. Capelli, et al. 2013. Schmallenberg virus in Italy: a retrospective survey in *Culicoides* stored during the bluetongue Italian surveillance program. Prev. Vet. Med. 111: 230–236.

Granados-Chinchilla, F., and C. Rodríguez. 2017. Tetracyclines in food and feedingstuffs: from regulation to analytical methods, bacterial resistance, and environmental and health implications. J. Anal. Methods Chem. 2017: 1315497.

Grüning, B., J. Chilton, J. Köster, et al. 2018. Practical computational reproducibility in the life sciences. Cell Syst. 6: 631–635.

Harvesting agriculture's "natural" insect farms: ARSX 2021 winning project. 2021. USDA ARS. https://www.ars.usda.gov/office-of-international-research-engagement-and-cooperation/overseas-biological-control-obcl-highlights/harvesting-agriculture-s-natural-insect-farms-arsx-2021-winning-project/

Häsler, B., K. S. Howe, E. Di Labio, et al. 2012. Economic evaluation of the surveillance and intervention programme for bluetongue virus serotype 8 in Switzerland. Prev. Vet. Med. 103: 93–111.

Hayek, A. M., B. J. McCluskey, G. T. Chavez, and M. D. Salman. 1998. Financial impact of the 1995 outbreak of vesicular stomatitis on 16 beef ranches in Colorado. J. Am. Vet. Med. Assoc. 212: 820–823.

Hoffmann, B., B. Bauer, C. Bauer, et al. 2009. Monitoring of putative vectors of bluetongue virus serotype 8, Germany. Emerg. Infect. Dis. 15: 1481–1484.

Hoxmeier, J. C., A. C. Fleshman, C. D. Broeckling, et al. 2017. Metabolomics of the tick-*Borrelia* interaction during the nymphal tick blood meal. Sci. Rep. 7: 44394.

Huang, M., S. Kingan, D. Shoue, et al. 2024. Improved high quality sand fly assemblies enabled by ultra-low input long read sequencing. Sci. Data 11: 1.

Jiao, J., Z. Lu, Y. Yu, et al. 2021. Identification of tick-borne pathogens by metagenomic next-generation sequencing in *Dermacentor nuttalli* and *Ixodes persulcatus* in Inner Mongolia, China. Parasit. Vectors 14: 1.

Jones, K. E., N. G. Patel, M. A. Levy, et al. 2008. Global trends in emerging infectious diseases. Nature 451: 990–993.

Jones, R. H., and N. M. Foster. 1974. Oral infection of *Culicoides variipennis* with bluetongue virus: development of susceptible and resistant lines from a colony population. J. Med. Entomol. 11: 316–323.

Jones, R. H., R. D. Roughton, N. M. Foster, et al. 1977. *Culicoides*, the vector of epizootic hemorrhagic disease in white-tailed deer in Kentucky in 1971. J. Wildl. Dis. 13: 2–8.

Joyner, C., M. K. Mills, and D. Nayduch. 2013. *Pseudomonas aeruginosa* in *Musca domestica* L.: temporospatial examination of bacteria population dynamics and house fly antimicrobial responses. PLoS One 8: e79224.

Kadri, S., A. Sboner, A. Sigaras, et al. 2022. Containers in bioinformatics. J. Mol. Diagn. 24: 442–454.

Karp BE, Tate H, Plumblee JR, et al. 2017. National antimicrobial resistance monitoring system: Two decades of advancing public health through integrated surveillance of antimicrobial resistance. Foodborne Pathog. Dis. 14(10):545–557.

Köster, J., and S. Rahmann. 2012. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 28: 2520–2522.

Kraemer, M. U. G., R. C. Reiner, O. J. Brady, et al. 2019. Past and future spread of the arbovirus vectors *Aedes aegypti* and *Aedes albopictus*. Nat. Microbiol. 4: 854–863.

Kraemer, M. U. G., M. E. Sinka, K. A. Duda, et al. 2015. The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. albopictus*. eLife 4: e08347.

Kulkarni, A., F. Delgadillo, S. Gayathrinathan, et al. 2023. Current status of omics studies elucidating the features of reproductive biology in blood-feeding insects. Insects 14: 802.

Kurtzer, G. M., V. Sochat, and M. W. Bauer. 2017. Singularity: scientific containers for mobility of compute. PLoS One 12: e0177459.

Larsson, D. G. J., and C.-F. Flach. 2022. Antibiotic resistance in the environment. Nat. Rev. Microbiol. 20: 257–269.

Larsson, D. G. J., C. de Pedro, and N. Paxeus. 2007. Effluent from drug manufactures contains extremely high levels of pharmaceuticals. J. Hazard. Mater. 148: 751–755.

Lim, N., S. Tesar, M. Belmadani, et al. 2021. Curation of over 10 000 transcriptomic studies to enable data reuse. Database 2021: baab006.

Manikandan, S., S. Manikandan, A. Mathivanan, et al. 2022. A review on vector borne disease transmission: current strategies of mosquito vector control. Indian J. Entomol. 2022: 1–11.

MacLachlan NJ. 1994. The pathogenesis and immunology of bluetongue virus infection of ruminants. Comp. Immunol. Microbiol. Infect. Dis. 17(3–4):197–206.

Margolis, R., L. Derr, M. Dunn, et al. 2014. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. J. Am. Med. Inform. Assoc. 21: 957–958.

McDermott, E. G., C. E. Mayo, A. C. Gerry, et al. 2015. Bluetongue virus infection creates light averse *Culicoides vectors* and serious errors in transmission risk estimates. Parasit. Vectors 8: 460.

McEwen, S. A., and P. J. Collignon. 2018. Antimicrobial resistance: a One Health perspective. In: S. Schwarz, L. M. Cavaco, and J. Shen, editors. Antimicrobial resistance in bacteria from livestock and companion animals. ASM Press, Washington, DC. p. 521–547.

McHolland, L. E., and J. O. Mecham. 2003. Characterization of cell lines developed from field populations of *Culicoides sonorensis* (Diptera: Ceratopogonidae). J. Med. Entomol. 40: 348–351.

Meiswinkel, R., L. M. Gomulski, J.-C. Delécolle, et al. 2004. The taxonomy of *Culicoides* vector complexes - unfinished business. Vet. Ital. 40: 151–159.

Mills, M. K., M. G. Ruder, D. Nayduch, et al. 2017. Dynamics of epizootic hemorrhagic disease virus infection within the vector, *Culicoides sonorensis* (Diptera: Ceratopogonidae). PLoS One 12: e0188865.

Moon, R. D. 2002. Muscid flies (Muscidae). In: Medical and veterinary entomology. Elsevier. p. 279–301.

Morales-Hojas, R., M. Hinsley, I. M. Armean, et al. 2018. The genome of the biting midge *Culicoides sonorensis* and gene expression analyses of vector competence for bluetongue virus. BMC Genomics 19: 624.

Morin, C. W., A. C. Comrie, and K. Ernst. 2013. Climate and dengue transmission: evidence and implications. Environ. Health Perspect. 121: 1264–1272.

Nayduch, D., and R. G. Burrus. 2017. Flourishing in filth: house fly–microbe interactions across life history. Ann. Entomol. Soc. Am. 110: 6–18.

Nayduch, D., L. W. Cohnstaedt, G. Sarai, et al. 2014. Studying *Culicoides* vectors of BTV in the post-genomic era: resources, bottlenecks to progress and future directions. Virus Res. 182: 43–49.

Nayduch, D., S. Neupane, V. Pickens, et al. 2023. House flies are underappreciated yet important reservoirs and vectors of microbial threats to animal and human health. Microorganisms 11: 583.

Nayduch, D., V. Shankar, M. K. Mills, et al. 2019. Transcriptome response of female *Culicoides sonorensis* biting midges (Diptera: Ceratopogonidae) to early infection with epizootic hemorrhagic disease virus (EHDV-2). Viruses 11: 473.

Neupane, S., and D. Nayduch. 2022. Effects of habitat and sampling time on bacterial community composition and diversity in the gut of the female house fly, *Musca domestica* Linnaeus (Diptera: Muscidae). Med. Vet. Entomol. 36: 435–443.

Neupane, S., J. L. Talley, D. B. Taylor, and D. Nayduch. 2023. Bacterial communities and prevalence of antibiotic resistance genes carried within house flies (Diptera: Muscidae) associated with beef and dairy cattle farms. J. Med. Entomol. 60: 1388–1397.

Neupane, S., K. White, J. L. Thomson, et al. 2020. Environmental and sex effects on bacterial carriage by adult house flies (*Musca domestica* L.). Insects 11: 401.

Newcomer BW, Cebra C, Chamorro MF, et al. 2021. Diseases of the hematologic, immunologic, and lymphatic systems (multisystem diseases). In: Pugh DG, Baird AN, Edmondson MA, et al., editors. Sheep, Goat, and Cervid Medicine (Third Edition). St. Louis: Elsevier. p. 405–438

Olagunju, E. A., A. S. Olagunju, and J. O. Teibo. 2023. The need to implement One Health approach in controlling vector-borne diseases in Nigeria. One Health Risk Manag. 4: 20–26.

One Health High-Level Expert Panel (OHHLEP), W. B. Adisasmito, S. Almuhairi, et al. 2022. One Health: a new definition for a sustainable and healthy future. PLoS Pathog. 18: e1010537.

Painuli, S., P. Semwal, R. Sharma, et al. 2023. Superbugs or multidrug resistant microbes: a new threat to the society. Health Sci. Rep. 6: e1480.

Patel, H., J. Manning, P. Ewels, et al. 2025. nf-core/rnaseq: nf-core/rnaseq v3.19.0 - Tungsten Turtle. Computer software.

Pećina-Šlaus, N., and M. Pećina. 2015. Only one health, and so many omics. Cancer Cell Int. 15: 64.

Petersen, L. R., and J. T. Roehrig. 2001. West Nile virus: a reemerging global pathogen. Emerg. Infect. Dis. 7: 611–614.

Pitt, S., and A. Gunn. 2024. The One Health concept. Br. J. Biomed. Sci. 81: 12366.

Rico-Hesse, R. 2007. Dengue virus evolution and virulence models. Clin. Infect. Dis. 44: 1462–1466.

Rozo-Lopez, P., B. S. Drolet, and B. Londoño-Renteria. 2018. Vesicular stomatitis virus transmission: a comparison of incriminated vectors. Insects 9: 190.

Rozo-Lopez, P., B. Londono-Renteria, and B. S. Drolet. 2020. Venereal transmission of vesicular stomatitis virus by *Culicoides sonorensis* midges. Pathogens 9: 316.

Ruder, M. G., D. Johnson, E. Ostlund, et al. 2017. The first 10 years (2006-15) of epizootic hemorrhagic disease virus serotype 6 in the USA. J. Wildl. Dis. 53: 901–905.

Rushton, J., F. Macchioni, L. Chelucci, et al. 2015. Economic impact of bluetongue: a review of the effects on production. Vet. Ital. 51: 401–406.

Salam, M. A., M. Y. Al-Amin, M. T. Salam, et al. 2023. Antimicrobial resistance: a growing serious threat for global public health. Healthcare 11: 1946.

Scott, P., T. Adedeji, H. Nakkas, and E. Andrikopoulou. 2023. One Health in a digital world: technology, data, information and knowledge. Yearb. Med. Inform. 32: 10–18.

Scroggs, S. L. P., E. J. Bird, D. C. Molik, et al. 2023. Vesicular stomatitis virus elicits early transcriptome response in *Culicoides sonorensis* cells. Viruses 15: 2108.

Shah, N. S., A. Wright, G.-H. Bai, et al. 2007. Worldwide emergence of extensively drug-resistant tuberculosis. Emerg. Infect. Dis. 13: 380–387.

Sim, S., and M. L. Hibberd. 2016. Genomic approaches for understanding dengue: insights from the virus, vector, and host. Genome Biol. 17: 38.

Simonsen GS. 2018. Antimicrobial resistance surveillance in Europe and beyond. Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull. 23(42):1800560.

Skender, B. 2024. The demise of the antibiotic pipeline: the Bayer case. Humanit. Soc. Sci. Commun. 11: 1069.

Spellberg, B., and D. N. Gilbert. 2014. The future of antibiotics and resistance: a tribute to a career of leadership by John Bartlett. Clin. Infect. Dis. 59: S71-75.

Stallknecht, D. E., and E. W. Howerth. 2004. Epidemiology of bluetongue and epizootic haemorrhagic disease in wildlife: surveillance methods. Vet. Ital. 40: 203–207.

Stone, B. L., Y. Tourand, and C. A. Brissette. 2017. Brave new worlds: the expanding universe of Lyme disease. Vector Borne Zoonotic Dis. 17: 619–629.

Tanya, V. N., E. C. Greiner, and E. P. J. Gibbs. 1992. Evaluation of *Culicoides insignis* (Diptera: Ceratopogonidae) as a vector of bluetongue virus. Vet. Microbiol. 32: 1–14.

Thabet, S., and R. Lajnef. 2024. Potential mechanisms underlying bluetongue virus emergence and spread. Front. Virol. 4: 1448192.

Thomson, J. L., K. M. Yeater, L. Zurek, and D. Nayduch. 2017. Abundance and accumulation of *Escherichia coli* and *Salmonella typhimurium* procured by male and female house flies (Diptera: Muscidae) exposed to cattle manure. Ann. Entomol. Soc. Am. 110: 37–44.

Tong, L., P.-Y. Wu, J. H. Phan, et al. 2020. Impact of RNA-seq data analysis algorithms on gene expression estimation and downstream prediction. Sci. Rep. 10: 17925.

Uelmen, J. A., C. D. Mapes, A. Prasauskas, et al. 2023. A habitat model for disease vector *Aedes aegypti* in the Tampa Bay area, Florida. J. Am. Mosq. Control Assoc. 39: 96–107.

van der Sluijs, M. T. W., A. J. de Smit, and R. J. M. Moormann. 2016. Vector independent transmission of the vector-borne bluetongue virus. Crit. Rev. Microbiol. 42: 57–64.

Velazquez-Meza, M. E., M. Galarde-López, B. Carrillo-Quiróz, et al. 2022. Antimicrobial resistance: One Health approach. Vet. World 15: 743–749.

Ventola, C. L. 2015. The antibiotic resistance crisis: part 1: causes and threats. P T 40: 277–283.

Wang, J., L. Chu, L. Wojnárovits, et al. 2020. Occurrence and fate of antibiotics, antibiotic resistant genes (ARGs) and antibiotic resistant bacteria (ARB) in municipal wastewater treatment plant: an overview. Sci. Total Environ. 744: 140997.

Watanabe, T. 1963. Infective heredity of multiple drug resistance in bacteria. Bacteriol. Rev. 27: 87–115.

White, A., and J. M. Hughes. 2019. Critical importance of a One Health approach to antimicrobial resistance. EcoHealth 16: 404–409.

Wilkinson, M. D., M. Dumontier, Ij. J. Aalbersberg, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. Sci. Data 3: 160018.

Wilkinson, M. D., M. Dumontier, Ij. J. Aalbersberg, et al. 2019. Addendum: the FAIR guiding principles for scientific data management and stewardship. Sci. Data 6: 6.

Wilkinson, S. R., G. Eisenhauer, A. J. Kapadia, et al. 2022. F*** workflows: when parts of FAIR are missing. In: 2022 IEEE 18th International Conference on e-Science (e-Science). IEEE, Salt Lake City, UT. p. 507–512.

Wratten, L., A. Wilm, and J. Göke. 2021. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nat. Methods 18: 1161–1168.

Xu, C., J. Xu, and L. Wang. 2024. Long-term effects of climate factors on dengue fever over a 40-year period. BMC Public Health 24: 1451.

Yang, Y., P. Xu, W. He, and F. Tao. 2024. Metagenomic analysis reveals houseflies as indicators for monitoring environmental antibiotic resistance genes. Environ. Microbiol. Rep. 16: e70032.

Yap, K. L., M. Kalpana, and H. L. Lee. 2008. Wings of the common house fly (*Musca domestica* L.): importance in mechanical transmission of *Vibrio cholerae*. Trop. Biomed. 25: 1–8.

Ziemann, M., P. Poulain, and A. Bora. 2023. The five pillars of computational reproducibility: bioinformatics and beyond. Brief. Bioinform. 24: bbad375.

Zurek, L., C. Schal, and D. W. Watson. 2000. Diversity and contribution of the intestinal bacterial community to the development of *Musca domestica* (Diptera: Muscidae) larvae. J. Med. Entomol. 37: 924–928.

Chapter 2 - A chromosome-scale genome of *Culicoides sonorensis* shows novel gene expansion and structural variation across Diptera

In preparation as: Edward Bird, Brandon Hall, David Molik, Scott M Geib, Tyler J Simmonds, Dana Nayduch, Kristopher Silver, Phillip Shults. A chromosome scale genome for *Culicoides sonorensis* shows novel gene expansion and structural variation across Diptera

Abstract

Culicoides biting midges are globally distributed hematophagous flies and primary vectors of significant livestock arboviruses of economic importance, including bluetongue virus, epizootic hemorrhagic disease virus, and African horse sickness virus. *Culicoides sonorensis* is the only confirmed North American vector species that is maintained in stable laboratory colonies and therefore serves as the model organism for understanding the biology of this important vector group. Studies of vector-virus interactions in *C. sonorensis* have been hampered by the limited genomic resources available in this model species. The existing genome for this species is a fragmented draft assembly that is not suitable for many types of analyses. Here, we provide a chromosome-scale genome of *C. sonorensis* assembled using PacBio HiFi long-read sequencing along with Hi-C scaffolding. The assembled genome spans 136.5 Mb, scaffolded into three chromosomes that make up 99.4% of the nuclear genome, in addition to a complete circular mitochondrial genome. Gene annotation, informed by 109 RNA-seq datasets from all life stages, produced two comprehensive gene models. Repeat analysis revealed 21.6% of the genome consists of repetitive elements, and comparative genomics with 30 dipteran species illustrated the phylogenetic divergence between *Culicoides* and other hematophagous vectors. Notably, we identified a lineage-specific gene expansion of *cytochrome P450 4d1-like* genes with a possible basis for enhanced detoxification capacity and insecticide resistance. This chromosome-level

assembly represents a significant enhancement to the previous iteration and establishes a strong basis for genomics-based studies in *C. sonorensis*. The evolutionary divergence of this species from other vector species underscores the potential of this system to examine the molecular foundations of vector competence and to guide the development of new vector control measures.

Introduction

Arthropod-borne diseases are among the most significant threats to both veterinary and human health globally. Arboviruses, or viruses transmitted by arthropods, pose a particularly serious threat to livestock and are responsible for substantial economic losses. Hematophagous insects play a central role in arbovirus transmission; however, much of the existing genomic and molecular research has focused on mosquito vectors. *Culicoides* biting midges (Diptera: Ceratopogonidae) are a diverse, globally distributed group of hematophagous insects (Bradley A. Mullens et al., 2015). Several *Culicoides* species are vectors of arboviruses, including African horse sickness, bluetongue virus (BTV), epizootic hemorrhagic disease virus (EHDV), vesicular stomatitis virus (VSV), Oropouche virus, Akabane virus, and Schmallenberg virus (McGregor et al., 2022; Purse et al., 2015).

Among the 151 recognized *Culicoides* species in the United States (Borkent & Grogan, Jr., 2009) only two, *Culicoides sonorensis* and *Culicoides insignis*, are confirmed vectors of Orbiviruses (Jones et al., 1977; Tabachnick, 1996; Tanya et al., 1992). While attempts have been made to colonize *C. insignis* (Erram & Burkett-Cadena, 2021) and several other species of *Culicoides* (Nayduch et al., 2014), *C. sonorensis* remains the only North American *Culicoides* species maintained in a stable laboratory colony. In the USA, *C. sonorensis* transmits BTV, EHDV, and VSV, making it highly relevant to the livestock industry. Due to its established role in arbovirus transmission and the availability of laboratory colonies, *C. sonorensis* remains

poised as the model *Culicoides* species for understanding and studying vector-virus dynamics and developing vector control strategies.

A draft genome of *C. sonorensis* was previously assembled using Illumina short-read sequencing (Morales-Hojas et al., 2018). While this assembly has acted as the basis for many genomics studies, its highly fragmented nature limits its utility for many types of analyses. A complete, chromosome-level genome assembly is essential for accurate gene prediction, reliable estimation of gene copy numbers, positional information for genes, and a comprehensive view of genome architecture. For example, chromosome-level genomes have been used to identify candidates for novel control strategies, such as temperature-based selection in *Anopheles albopictus* (Diniz et al., 2022; John H. Boyle et al., 2021), as well as gene drives for sex-based control in multiple *Anopheles* mosquitoes (Alekos Simoni et al., 2020; Xu et al., 2025). Additionally, a chromosome scale genome allows for comparative genomics, allowing for the integration of information from well-studied vectors and model insects.

Due to the relatively small size of *Culicoides* midges (less than 3 mm) it was not previously possible to produce whole genome sequencing from a single individual. Recent advances in long read library generation techniques, in combination with the high-fidelity (HiFi) long read data produced by PacBio makes single specimen whole genome sequencing a possibility (Jessica Castellanos-Labarcena et al., 2025). The combination of scaffolding quality assemblies with high-throughput chromatin confirmation capture (Hi-C) makes it feasible to reliably assemble a chromosome-scale genome from single insects.

This study presents a chromosome-level assembly for a female *Culicoides sonorensis* that includes three chromosomes, representing 99.4% of the genome. This new assembly was compared to other available chromosome-level genomes to assess its place in Diptera. This

resource enables new opportunities for comparative genomics, particularly among hematophagous vectors. As *Culicoides* midges are phylogenetically distant from other insects with available genomes like sand flies (Diptera: Psychodidae), and mosquitoes (Diptera: Culicidae), this genome provides an important foundation for exploring the evolutionary history of vector competence. With many questions remaining about the diversity and vector potential of *Culicoides* species in North America, this genome also lays the groundwork for studies aimed at understanding virus–vector interactions in this understudied group.

Methods

Sample Generation & Sequencing

Two adult female *C. sonorensis* obtained from the KS-colony, established in 2022 and maintained at the USDA in Manhattan, KS were used for genome sequencing; one for HiFi and one for HiC. High molecular-weight (HMW) DNA was extracted from one specimen using a Qiagen MagAttract HMW DNA kit (Qiagen, Hilden, Germany). The DNA quality was assessed using a Denovix DS-11 fluorometer/spectrophotometer (Denovix Inc Wilmington, DE), and the concentration was measured using a Qubit Fluorometer (Thermo Fisher Scientific, Waltham, MA) with the dsDNA BR Assay Kit. The extracted DNA was sheared to 15–20 kb using a Megaruptor 3 (Diagenode, Denville, NJ), and a single SMRTbell library was prepared using the SMRTbell 3.0 Library Preparation Kit following the manufacturer's low-input protocol (Pacific Biosciences, Menlo Park, CA). The final library was size selected with Ampure PB beads (Beckman Coulter, Brea, CA) to remove fragments <3 kb in length. Sequencing was performed on a Sequel II System using Binding Kit v2.0, Sequencing kit v2.0, and SMRT Cell 8 M. For Hi-C sequencing, the second female was used to prepare a proximity-ligation library from crosslinked tissue following standard protocols (adapted from Lafontaine et al. 2021), using the

restriction enzymes DdeI and DpnII. The sample was then prepared using an Elevate Enzymatic Library Prep Kit and sequenced with 150 bp paired-end reads on an Element Biosciences AVITI system (Element Biosciences, San Diego, CA).

Life History RNAseq Panel

Pools of *C. sonorensis* individuals from the AK colony (established 1973, USDA ARS, Manhattan, KS) were collected from three independent rearing pans. Pools were collected at each developmental stage, including stages where sex could not be determined (egg; larval instars L1–L4) and those where sex could be distinguished (male/female pupae and adults). The number of individuals per pool varied by life stage: egg = 1,000; L1 = 100; L2 = 50; L3 = 50; L4 = 50; pupae = 25; adults = 25. Pools were homogenized in 200 µL of TRIzol reagent (Zymo Research, Irvine, CA) and centrifuged (4 °C, 10,000 x g, 5 minutes) to pellet debris. A total of 60 µL of bromochloropropane was added to the supernatant, mixed by gentle inversion, and centrifuged at 4 °C at 10,000 x g, for 15 minutes. The resulting aqueous phase was recovered, and RNA was purified using the Direct-zol RNA Miniprep Kit (Zymo Research, city, state), following the manufacturer's instructions. Stranded mRNA libraries were prepared by Novogene using poly(A) enrichment and sequenced on an Illumina NovaSeq platform, generating 300 bp paired-end reads, with a target depth of 20 million reads per sample.

Mitochondrial Genome Assembly and Removal

All software, unless explicitly mentioned, was utilized though Apptainer v1.3.1 (Gregory M. Kurtzer et al., 2017). Specific commands used to run the analysis can be found in the GitHub repository (https://github.com/edwardbirdlab/CSON_GENOME). PacBio HiFi reads were first processed with pbadapterfilt v3 (https://github.com/sheinasim-USDA/HiFiAdapterFilt/tree/master) to remove sequencing adapters. To prevent mitochondrial

reads from interfering with nuclear genome assembly, the mitochondrial genome was assembled separately using MitoHiFi v3.2.2 (Uliano-Silva et al., 2023). MitoHiFi identified mitochondrial reads, assembled them with HiFi-ASM, and selected a single circularized assembly. The *C. sonorensis* reference mitochondrial genome (BK065013.1) was provided as a guide to aid in identifying mitochondrial reads. Following mitochondrial genome assembly, Minimap2 v2.26 (Li, 2018) was used to align all adapter-filtered HiFi reads against the assembled mitochondrial genome and reads that mapped to the mitochondrial genome were removed. The remaining unmapped reads, representing nuclear genomic sequences, were extracted using Samtools v1.17 (Danecek et al., 2021) and subsequently used for nuclear genome assembly.

Nuclear Genome Assembly

Mitochondrial-free HiFi reads were assembled using Hifiasm v0.24.0 (Cheng et al., 2021), integrated with Hi-C data for haplotype phasing. The primary assembled contigs were extracted from the Hifiasm output for downstream analysis. To identify potential contamination in the draft assembly, adapter-filtered HiFi reads were mapped back to the draft assembly using Minimap2 v2.26 and converted to BAM format with Samtools v1.17. To generate taxonomic classifications, protein homology searches were conducted using Diamond v2.0.15 (Buchfink et al., 2021) against the UniProt/SwissProt database (Accessed: 2/21/2025) with an e-value threshold of 1e-25. These alignment and homology results were imported into Blobtools v1.1 (Laetsch & Blaxter, 2017) for visualization. Based on GC content, taxonomic assignment, and read coverage, a list of scaffolds identified as putative contaminants was generated and the assembly was filtered accordingly.

Chromosome-Scale Scaffolding

Chromosome-level scaffolding was performed using the Hi-C–based 3D-DNA pipeline v190716 (Dudchenko et al., 2017) in combination with Juicer v2.0.1 (Durand, Shamim, et al., 2016). The filtered haplotype genome was indexed using BWA v0.7.19 (Li & Durbin, 2010), and DpnII cut sites were identified with the generate_site_positions.py script from Juicer. A FASTA index of the draft genome was generated using SAMtools v1.17, and a corresponding sequence size file was produced. Juicer was then used to create a non-redundant Hi-C contact matrix. This matrix served as input for the 3D-DNA pipeline to perform automated, chromosome-level scaffolding.

Manual Curation of Hi-C Scaffolding

The draft chromosome assembly was manually curated using Juicebox v2.15 (Durand, Robinson, et al., 2016) following the guidelines in the Genome Assembly Cookbook from the Aiden Lab (https://aidenlab.org/assembly/manual_180322.pdf). The output assembly file from 3D-DNA and the Hi-C contact map were loaded into Juicebox for manual inspection and scaffold adjustment. Manual curation primarily involved minor adjustments to the automatically generated assembly, which already showed clear chromosome-level organization with relatively few unplaced scaffolds. Some ambiguous unplaced scaffolds showing contact with chromosomes were left unplaced due to uncertain positioning. The manually curated assembly was exported from Juicebox and the final gapped chromosome assembly was generated using the 3D-DNA post-review pipeline.

Debris Cleanup & Contamination Screening

The final assembly was subjected to a final contamination screening using NCBI's Foreign Contamination Screen (FCS-GX v3.0.4, database from January 19, 2023) with the

taxonomic ID 179676 (*C. sonorensis*). No contamination was detected by FCS-GX, although several debris pieces had no identifiable hits in the database. To identify potential duplicate regions in unplaced scaffolds, the assembly was split into chromosome and debris sequences and MUMmer v4.0 (Marçais et al., 2018) was used to compare the debris sequences to the chromosome assembly. Lastly, a preliminary annotation was generated using NCBI's Eukaryotic Genome Annotation Pipeline – External (eGAPx v0.3.2-alpha) to determine if debris contained any coding/non-coding genes.

A rule set was developed to define acceptable evidence for removal of debris from the assembly. Unplaced scaffolds were systematically evaluated for removal based on four distinct criteria (Types I-IV). Type I removal included haplotype contamination as observed by Hi-C mapping. These scaffolds displayed strong Hi-C signals within the debris and weak or absent signals in the corresponding chromosome region. Type II removal included haplotype contamination as indicated by Hi-C & MUMmer outputs. Unplaced scaffolds with little to no Hi-C contacts, but high query coverage (>90%) when compared to chromosomal sequences using MUMmer, were removed as potential haplotigs. Type III removal focused on small scaffolds with no annotations. These included elimination of small scaffolds (<10 kb) with no predicted annotations from the eGAPx preliminary annotation. Type IV removal included unidentified sequences with no annotations which included sequences that FCS was unable to taxonomically classify and that received no annotations from eGAPx. Based on these criteria, several scaffolds were removed in two phases. Before preliminary eGAPx annotation (Phase 1) and after (Phase 2). Phase 1 removals included three scaffolds that met one or more Type I or Type II conditions. Phase 2 removals included six scaffolds that met either Type III or Type IV conditions.

Chromosomes were named in order of size as chr1, chr2, and chr3. Unplaced scaffolds were named according to their relationship with chromosomes: Unplaced scaffolds with Hi-C contact evidence, with a specific chromosome but ambiguous positioning, were labeled as chr#unplaced# (numbered by size). Unplaced scaffolds without clear Hi-C contact evidence with any chromosome were labeled as unplaced_# (numbered by size).

Repeat Annotation

Repetitive elements, including transposable elements, satellite DNA, and low-complexity repeats, were annotated and masked using Earl Grey v5.1.1 (Baril et al., 2024). The pipeline integrates both homology-based and *de novo* approaches to improve repeat detection and classification. Initial masking was performed with RepeatMasker v4.1.6 using the Dfam v3.7 database (Storer et al., 2021), and Earl Grey subsequently refined consensus sequences through sequence alignment, clustering, and classification steps, producing a non-redundant repeat library that was used for final repeat masking. For comparative purposes, the *C. brevitarsis* genome (GCA_036172545.2) was also analyzed with Earl Grey using the same parameters.

Transcriptome Data Collection and Processing

Raw data from the life history RNA-seq was processed using the 'sr_qc_only' subworkflow of the BALROG-MSR pipeline (https://github.com/edwardbirdlab/BALROG-MSR), implemented in Nextflow. Initial quality assessment was performed using FastQC v0.12.1 (Andrews, 2010), followed by adapter trimming and quality filtering with fastp v0.20.1 (Chen, 2023), applying a minimum read length threshold of 50 bp. Summary metrics were compiled and visualized using MultiQC v1.22.2 (Ewels et al., 2016).

Publicly available RNA-seq datasets corresponding to *C. sonorensis* (NCBI Taxonomic ID: 179676) were downloaded from the NCBI SRA using SRA-FETCH

(https://github.com/edwardbirdlab/SRA-FETCH) and processed identically with FastQC, FastP, and MultiQC. All samples, with the exception of SRR9091274, passed quality thresholds and were retained for downstream genome annotation. In total, 109 RNA-seq samples were included in the final analysis.

To generate both genome browser tracks and inputs for downstream genome annotation, RNA-seq data was processed using the genome_prep pipeline (https://github.com/edwardbirdlab/genome_prep), also implemented in Nextflow. Trimmed reads were aligned to the final scaffolded genome using STAR v2.7.11b (Dobin et al., 2013). Samtools v1.17 was used to convert alignments to BAM format, followed by sorting and indexing. Python deepTools v3.5.5 (Ramírez et al., 2014) was then used to generate both raw and normalized bigWig tracks. A unified RNA-seq track was produced using University of California, Santa Cruz's bigWigMerge utility (Perez et al., 2025). In addition, all sorted BAM files were merged using Samtools v1.17.1 to produce a single alignment file for input to BRAKER3.

Genome Annotation

Evidence-based gene annotation was performed using the eGAPx v0.3.2-alpha, the publicly accessible version of the NCBI Eukaryotic Genome Annotation Pipeline. eGAPx integrates transcriptomic and protein evidence to guide high-confidence structural gene prediction. The pipeline uses STAR for RNA-seq alignment and miniprot for protein sequence alignment, with both evidence types passed to Gnomon for gene model construction and refinement. In total, 109 high-quality RNA-seq samples were provided as transcriptomic evidence, and the Dipteran orthologous protein database from NCBI was used for protein-based support. The resulting annotation includes *ab initio* predictions supplemented by evidence-based models and functional annotations.

Gene prediction was also performed using BRAKER3, an automated pipeline that integrates extrinsic evidence from RNA-seq and protein homology to predict high-confidence gene models. BRAKER3 (Gabriel et al., 2024) combines GeneMark-ETP (Brůna et al., 2024) and AUGUSTUS (Stanke et al., 2006) to generate gene predictions supported by RNA-seq alignments and homologous protein sequences. For this annotation, RNA-seq alignments from 109 samples and protein homology evidence from the Arthropoda lineage in OrthoDB (accessed April 18, 2025) were provided as input. The resulting gene set includes only those gene models with strong support, including *ab initio* gene predictions.

Comparative Genomics

A curated set of Dipteran genomes and annotations was compiled for downstream synteny and orthology analyses. Chromosome-scale assemblies with RefSeq annotations were collected from NCBI, with a maximum of one representative genome per genus to avoid taxonomic bias. In addition to these, the chromosome-scale genome of *Polypedilum vanderplanki* (GCA_018290095.1) was included, along with its user-submitted annotation. For *Culicoides stellifer*, a fragmented genome assembly (GCA_040583785.1) was annotated using BRAKER2 with OrthoDB v12 as protein evidence. The eGAPx annotation of *C. sonorensis*, generated in this study, was used in place of a RefSeq annotation, as it closely mirrors NCBI's standard pipeline output. A complete list of all genome assemblies and corresponding annotations used in this section is provided in Supplementary Table 2.1.

All genomes were processed using the edwardbirdlab/nf_multi_synteny (github.com/edwardbirdlab/nf_multi_synteny) pipeline. This pipeline performed QUAST v5.2.0 (Gurevich et al., 2013) on each genome to calculate assembly statistics, and protein sequences were extracted from the longest isoform per gene using AGAT v1.4.2 (Jacques Dainat et al.,

2024). BUSCO v5.8.2 (Manni et al., 2021) was run on these protein sets to assess genome and annotation completeness. To identify reciprocal best hits, Diamond BlastP v2.1.11 searches were performed between all pairwise pairs of species. These results were used as input for OrthoFinder v3.0.1 (Emms and Kelly 2015, 2017, 2018, 2019), which inferred orthologs, gene trees, and additional comparative statistics across the 30 genomes. The same BLAST results, along with gene position information, were also used as input to MCScanX v0.1 (Wang et al., 2024) to identify pairwise collinear gene blocks across species.

To construct a gene tree for *cytochrome P450 4d1-like*, orthologs were retrieved from OrthoDB (version 12) and incorporated into the orthologous gene sets identified by OrthoFinder. Protein sequences were aligned using MAFFT v7.5 (Katoh & Standley, 2013) with automatic model selection, which selected the FFT-NS-i algorithm as the optimal alignment strategy. The resulting alignment was trimmed using TrimAl v1.5 to remove erroneous gaps and low-quality sequences. A maximum-likelihood gene tree was then constructed with FastTree v2.1.11 (Price et al., 2010), using the WAG substitution model with gamma-distributed rate variation among sites. The final tree was midpoint-rooted for visualization.

Results

Genome Assembly & Scaffolding

MitoHiFi assembled a circular mitochondrial genome that was 16,228 bp. This mitogenome had a minimum coverage of ~600x and contained all the expected 37 mitochondrial genes. The annotation and position of these genes is displayed in Appendix Figure B1. The final Hi-C–scaffolded genome consisted of three chromosome-scale scaffolds and 29 unplaced scaffolds. The Hi-C contact map was visualized using Juicebox and is shown in Appendix Figure B2. Of the 29 unplaced scaffolds, nine were removed for violating one of the five criteria defined

in the cleanup section of the method, resulting in a final set of three chromosomes and 20 unplaced scaffolds. Among these, six scaffolds showed ambiguous Hi-C contacts with a single chromosome and were named accordingly. The final assembly totaled 136,531,319 bp, with chromosome 1 measuring 48,373,389 bp, chromosome 2 measuring 46,011,774 bp, and chromosome 3 measuring 41,322,406 bp. The unplaced scaffolds accounted for 823,750 bp, with the largest being 192,232 bp and showing no contact with any specific chromosome. Using the Dipteran BUSCO database (version 10), all but one BUSCO in genome mode were identified on one of the three chromosome scaffolds. The representation of the three chromosomes and the density of BUSCOs, as well as overall BUSCO completeness, are shown in Figure 2.1A and 2.1F, respectively.

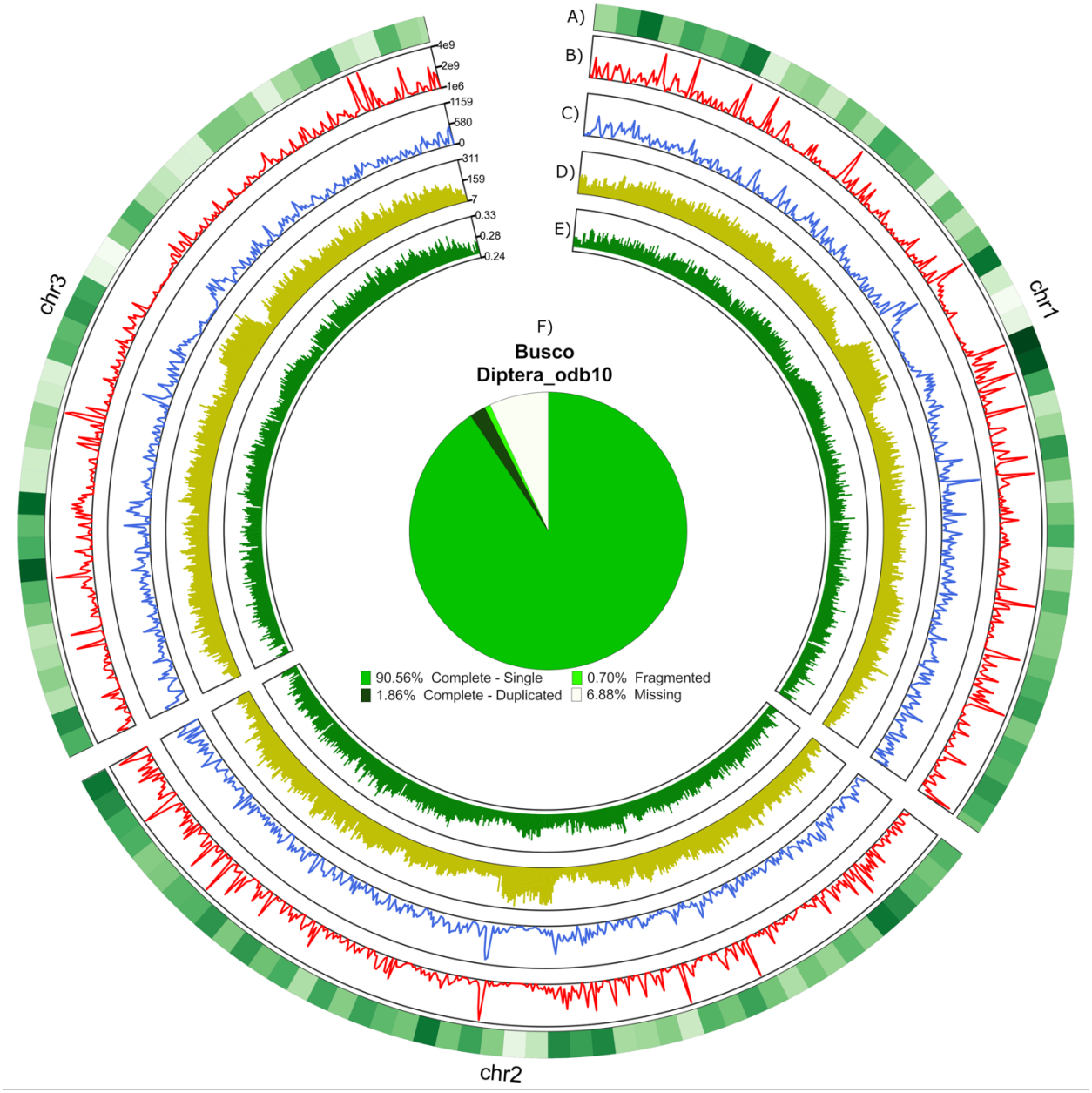


Figure 2.1 Circular representation of the *C. sonorensis* genome assembly and annotation features.

A) BUSCO gene density, with darker regions indicating higher density; B) RNA-seq read depth based on all 109 samples, scaled to the 99th percentile; C) gene density based on Braker3 annotation; D) repeat density as identified by Earl Grey; and E) GC content. F) BUSCO completeness summary (ODB10 dataset), showing proportions of complete, duplicated, fragmented, and missing BUSCOs.

Genome Repeat Annotation

Repeat annotation of the *C. sonorensis* genome using Earl Grey identified 29,437,377 bp of repetitive elements, comprising 21.56% of the genome. Of these, 15,120,091 bp (11.07% of the genome) could not be classified by Earl Grey. The remaining repeats were annotated as DNA transposons, rolling-circle elements, Penelope elements, long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), long terminal repeat (LTR), simple repeats, microsatellites, or RNA elements. The *C. brevitarsis* genome was annotated using the same Earl Grey pipeline and was found to contain 19,505,668 bp of repetitive elements, representing 15.06% of the genome. Of these, 7,271,903 bp (5.61% of the genome) were unclassified. All repeat types identified in *C. sonorensis* were also detected in *C. brevitarsis*, except for SINEs and LTRs. While only 10 SINEs were detected in *C. sonorensis*, 1,009 LTR elements were identified, accounting for 0.3% of the genome. The distribution of repetitive elements and G/C content across the three *C. sonorensis* chromosomes is visualized in Figure 2.1D and 2.1E, respectively.

Genome Annotation

eGAPx predicted 14,754 genes in the *C. sonorensis* genome. Of these, 11,367 were predicted to be protein-coding genes, and 3,833 were predicted as long non-coding RNAs. Among the protein-coding genes, 3,193 were predicted to have variants, resulting in 18,776 predicted mRNAs. In comparison, Braker3 identified significantly more genes than eGAPx, with 13,759 predicted protein-coding genes corresponding to 17,373 transcripts. However, Braker3 predicted significantly fewer isoforms per gene, with an average of 1.3 isoforms, compared to eGAPx's average of 1.7. Additionally, eGAPx predicted 1,171 overlapping genes, while Braker3 identified only 114. Gene density based on the Braker3 annotation, and RNA-seq read density

from all 109 samples are visualized in Figure 2.1B and 2.1C, respectively. The *C. stellifer* genome was also annotated using Braker2, resulting in 16,130 predicted protein-coding genes and 17,335 transcripts, with an average of 1.1 isoforms per gene.

Comparative Genomics

From the 30 species analyzed, OrthoFinder3 identified 19,541 orthogroups (Supplementary Table 2.2). Of these, 11,974 orthogroups contained at least four genes and were used to generate gene trees using FastTree with the GTR+gamma models. Among these, 2,521 orthogroups were considered high quality by Species Tree from All Genes (STAG) (minimum of 96.7% of species having single-copy genes) and were used for assembling the all-genes tree. All nodes in the resulting tree received 100% support from the constituent ortholog gene trees. Additionally, the phylogeny was supported by 1,259 observed duplication events, with only six duplications lacking support. This all-genes tree was used to compare genome quality, completeness, and gene count across all 30 species, as shown in Figure 2.2. Among the dipteran genomes, *P. vanderplanki* had the smallest genome (119 Mb), with *Culicoides* species among the next smallest, ranging from 119 Mb (*C. stellifer*) to 137 Mb (*C. sonorensis*). Notably, all three *Culicoides* genomes were missing 491–507 (9.8–10%) single-copy orthologs (SCOs) from the BUSCO Diptera v12 database. Among these missing SCOs, 360 were consistently absent across all three *Culicoides* genomes.

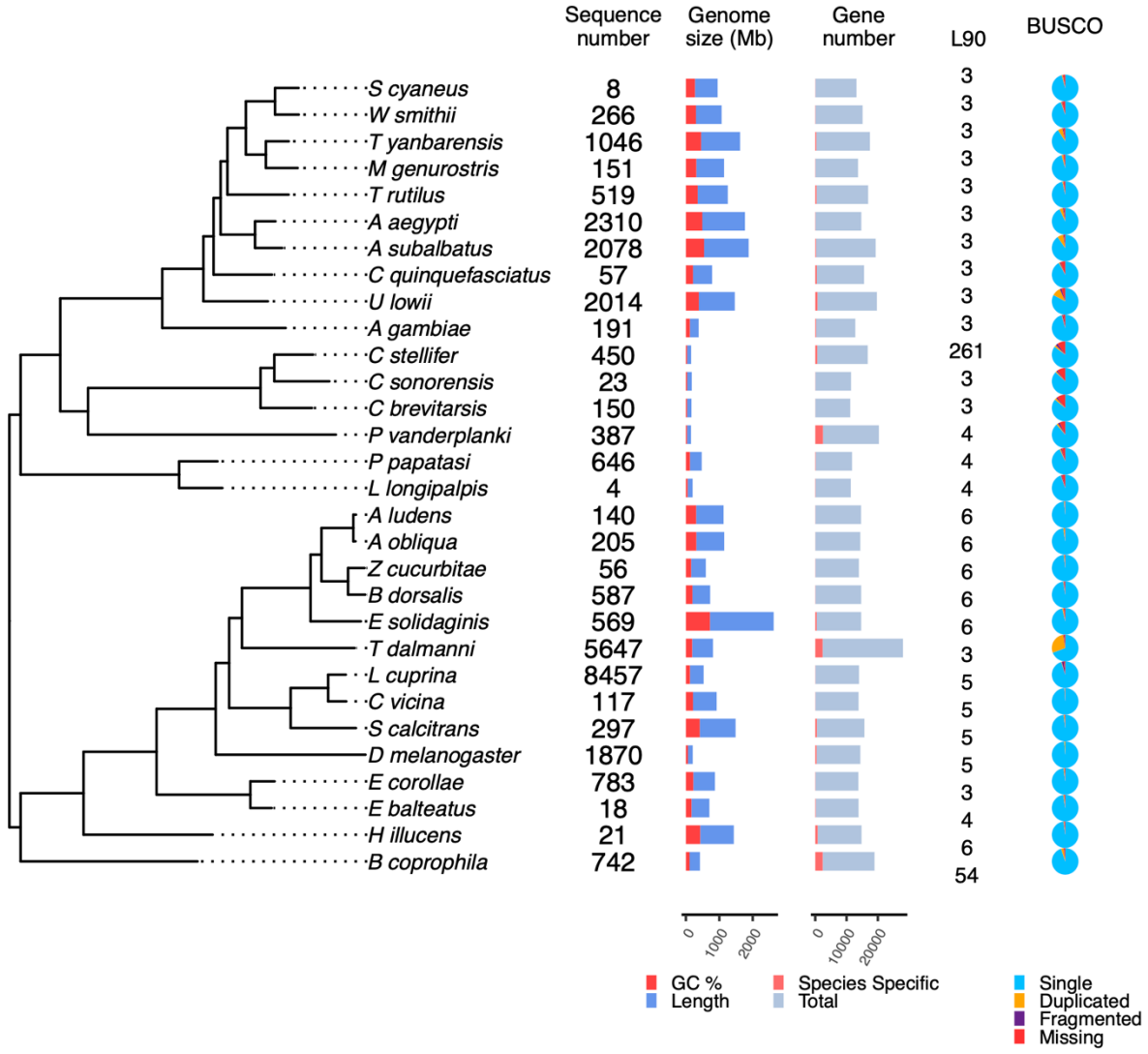


Figure 2.2 Species tree and genome characteristics of included dipterans.

Species tree generated by OrthoFinder3 with STAG using 2,521 orthogroups, with rooting performed by STRIDE, placing higher Diptera as the outgroup. All nodes are supported by 100% of the gene trees used in the consensus. For each species (Supplementary Table 2.1), genome statistics are displayed alongside the corresponding tip: total number of sequences in the assembly, genome size and GC content, total number of genes and number of genes in species-specific orthogroups, L90, and BUSCO completeness (Diptera OrthoDB v12).

Of the 19,541 orthogroups identified by OrthoFinder3, 96.8% of all input genes (total genes = 447,007) were assigned to an orthogroup. Among these, 4,884 orthogroups contained at least one protein from all 30 species, and 1,040 orthogroups contained exactly one gene from each species. *P. vanderplanki* was the most isolated species in the analysis, with 9.5% of its proteins unassigned to any orthogroup. The number of orthogroups shared between species is visualized in Figure 2.3 (Supplementary Table 2.3). *C. stellifer* shared the greatest number of orthogroups with *C. sonorensis* (n = 8,947), followed by *C. brevitarsis*, which shared 8,573 orthogroups. Outside of the Ceratopogonidae family, *Bradysia coprophila* shared the most orthogroups with *C. sonorensis* (n = 8,058). The next most similar group was the clade Culicidae, which included *Armigeres subalbatus*, shared 7,980 orthogroups. The species with the fewest shared orthogroups was *Stomoxys calcitrans*, with 7,530 overlapping orthogroups.

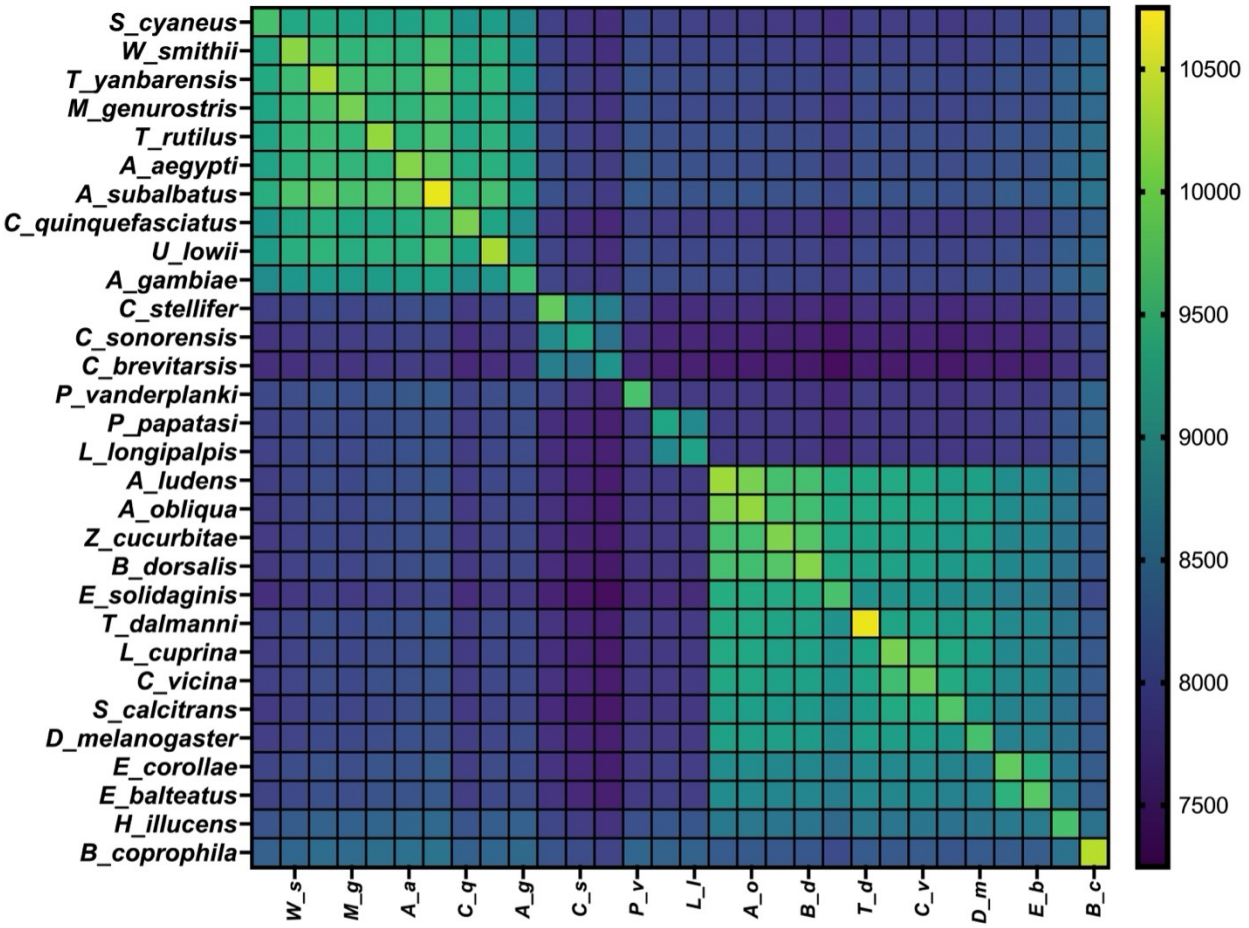


Figure 2.3 Shared orthogroups between dipteran species

Heatmap showing the number of orthogroups shared between each pairwise combination of the 30 species included in this study. Values represent the number of orthogroups identified by OrthoFinder3 that contain at least one protein from each species in the pair.

Several identified orthogroups were *Culicoides*-specific, meaning they contained only proteins from *Culicoides* species, with no representatives from the other 27 species analyzed. One orthogroup of particular interest, OG0002693, contained 34 *Culicoides*-specific proteins, 12 from *C. brevitarsis*, 12 from *C. sonorensis*, and 10 from *C. stellifer*. The proteins in this orthogroup were identified as Cytochrome P450 4d1-like (Cyp4d1) and were compared against the existing Cyp4d1 orthogroup from OrthoDB (7227_0:002854). The resulting gene tree revealed a clear separation between *Culicoides* Cyp4d1-like proteins and those from OrthoDB

(Appendix Figure B.3). Furthermore, a significantly higher copy number of Cyp4d1-like proteins was observed in *Culicoides* (n = 3 genomes), with an average of 11.3 copies per genome, compared to an average of 3.14 copies per genome in Culicidae (n = 28 genomes).

Synteny

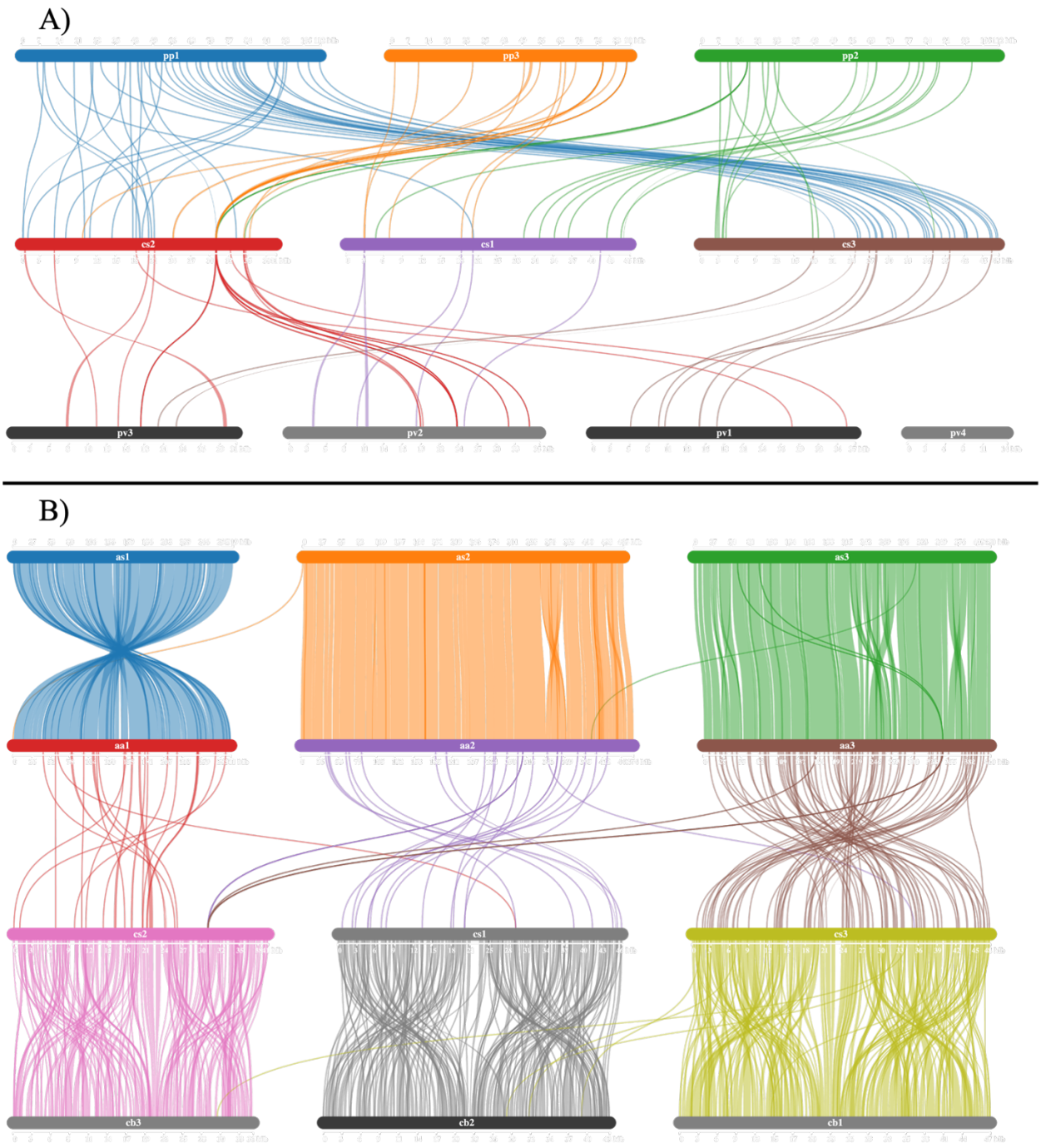


Figure 2.4 . Synteny between *C. sonorensis* genome and those of select Nematocera species.

Colinear blocks of conserved gene order between species pairs, identified using MCScanX with consistent parameters and visualized using SynVisio. (A) Synteny between *Phlebotomus papatasi* (top, pp), *C. sonorensis* (middle, cs), and *P. vanderplanki* (bottom, pv). (B) Synteny between two mosquito species from different genera, *A. subalbatus* (top; as) and *Aedes aegypti*

(middle; aa), and two biting midges from the genera *Culicoides*, *C. sonorensis* (cs) and *C. brevitarsis* (cb).

The two chromosome-scale *Culicoides* genomes (*C. sonorensis* and *C. brevitarsis*) showed that 30.6% of their genes were in colinear blocks. In comparison, two mosquito species from different genera (*Armigeres* and *Aedes*) exhibited 28.1% of their genes in syntenic blocks. Additionally, as shown in Figure 2.4B, the syntenic blocks in mosquitoes are much longer and display fewer spatial rearrangements than those between *C. sonorensis* and *C. brevitarsis*. Chromosome 1 in *C. sonorensis* shared significantly more syntenic blocks with *Aedes* chromosome 2 than with any other chromosome. In contrast, when comparing gene structure between *C. sonorensis* and its closest available non-Ceratopogonidae relative, *P. vanderplanki* (Figure 2.4A), only 0.7% of genes were found in colinear blocks, indicating minimal synteny. Similarly, comparisons with *P. papatasi* (Figure 2.4B) and *A. aegypti* (Figure 2.4A) revealed limited but detectable conserved gene order, with 1.85% and 2.83% of genes, respectively, found in colinear blocks.

Discussion

The chromosome-level assembly presented in this study provides a valuable resource for *Culicoides* research, enabling genetic investigations that were previously inaccessible. Over 99% of the genome was scaffolded into chromosomes, representing a major improvement over the previous assembly, which had an N50 of 109.2 KB (GCA_900258525.3). This chromosomal resolution is essential for advanced genomic applications, such as the design of genome editing experiments, mapping long-range regulatory interactions, and resolving gene duplications and gene order across species. It will also enable more accurate interpretation of structural variation and facilitate future studies aimed at linking genotype to phenotype. Additionally, *Culicoides*

spp. represent a clade of blood-feeding vectors that is likely phylogenetically separated by at least 200 million years (Borkent & Craig, 2004; Szadziewski, 2016) from the nearest group with chromosome-scale genomic resources, which are Culicidae mosquitoes. This divergence presents a unique opportunity to explore the evolution of diverse biological processes like hematophagy, anautogeny, vector competence and other traits across long evolutionary timescales. Additionally, this genome helps to fill a gap in an underrepresented portion of dipteran phylogeny, broadening the genomic resources available.

Two different gene model annotations are provided with this genome assembly, both incorporating RNAseq evidence from 109 samples from a diverse set of conditions and life stages (excluding only the first larval stage). The first annotation was produced with eGAPx and predicts fewer total genes, but with substantially more transcript isoforms per gene (1.7 per gene). The eGAPx pipeline closely mirrors the internal RefSeq annotation pipeline and may be particularly useful for comparative genomics with RefSeq-annotated genomes. The second annotation was generated with BRAKER3 and predicts a substantially higher number of protein-coding genes, but with fewer transcript variants on average (1.3 per gene). This annotation may capture additional genes that lack strong RNA-seq support or were not predicted by eGAPx. While a large amount of RNAseq evidence was utilized in this study, the average number of transcript isoforms per gene (1.7 from eGAPx) is somewhat lower than in some Culicidae species predicted by RefSeq, such as *Aedes aegypti* (1.9 isoforms; NCBI accession) and *Anopheles gambiae* (2.4 isoforms; NCBI accession), suggesting the potential for missed splice forms in the *C. sonorensis* annotation. We also assessed genome completeness using BUSCO (Diptera ODBv12), which found that 496 genes (9.8%) were missing. Notably, *C. stellifer* and *C. brevitarsis* showed similar proportions of missing genes (9.7% and 10%, respectively). Among

these, the same 360 genes (7.1%) were consistently absent from all three *Culicoides* genomes, suggesting that current BUSCO expectations may not fully reflect the gene content of this lineage.

This new genome was compared to 29 other dipteran genomes from NCBI, including at least one representative from each genus with a chromosome-level assembly and a RefSeq annotation. Phylogenetic reconstruction using OrthoFinder3 produced a species tree consistent with current understandings of dipteran evolution (Wiegmann et al., 2011). The *Culicoides* species exhibited substantial phylogenetic divergence from other analyzed species, forming a distinct lineage with considerable branch length separation (Figure 22.), sharing fewer orthogroups with other genera compared to both higher dipterans and mosquitoes (Figure 2.3). OrthoFinder3 assigned 96.8% of the 447,007 genes to 19,54 orthogroups, showing that the selected dipteran species did not represent large evolutionary gaps in the phylogeny. Interestingly, *C. sonorensis* did not share the most orthologs with its closest known non-*Culicoides* relative, *Polypedilum vanderplanki* but rather with a fungus gnat (*B. coprophila*) and a mosquito (*A. subalbatus*). This discrepancy may be partially due to differences in annotation methods. *P. vanderplanki* lacks a RefSeq annotation and was instead annotated with BRAKER3. This may introduce bias in ortholog inference. However, *C. stellifer* was also annotated with Breaker3, and shows the highest number of orthologs with *C. sonorensis*. These results highlight the sparse genomic resources among *Culicoides* and related genera which will require additional resources to fill.

Syntenic analysis was also conducted across all chromosome-scale genomes using uniform parameters, to compare the conservation of gene order. *C. sonorensis* shared few syntenic blocks with other dipterans outside of its own family. Only 288 genes (0.7%) were part

of syntenic blocks with *P. vanderplanki*, the closest available relative for comparison outside of Ceratopogonidae. In contrast, greater synteny was observed with more distantly related species such as *A. aegypti* (n = 1,558; 2.83%) and *P. papatasi* (n = 812; 1.85%). These results may reflect differences in lifestyle and the ecological niches that they fill. Whereas *P. vanderplanki* is the most recent common ancestor to Ceratopogonidae in this study, its non-blood feeding, plant-associated lifestyle may have led to relatively more rapid divergence or rearrangement of genes as compared to the other blood feeders who may have conserved or convergent selection pressures associated with hematophagy. In contrast, *B. coprophila*, also a plant feeder, shares a higher number of orthologs with *C. sonorensis*, but its current RefSeq annotated genome is fragmented, and consequently, it is not possible to make an informative analysis of synteny. Future annotation and use of the available chromosome level *B. coprophila* genome will permit clearer analysis of synteny between these two species.

The *C. sonorensis* genome was also compared to two other *Culicoides* genomes, *C. stellifer* and *C. brevitarsis*. Repeat content varied substantially among the three *Culicoides* species, with *C. sonorensis* harboring the highest proportion followed by *C. brevitarsis* (15.1%) and *C. stellifer* (11%). Notably, *C. brevitarsis* lacked detectable LTR elements, which made up 0.3% and 0.4% of the *C. sonorensis* and *C. stellifer* genomes, respectively. Repeats can significantly influence genome evolution and function by driving mutations, promoting rearrangements, and modulating gene expression (Bourque et al., 2018). Furthermore, specific repeat types, such as endogenous retroviruses (ERVs), which have been previously identified in *C. stellifer* (Jessica Castellanos-Labarcena et al., 2025), have been implicated in innate immune regulation (Chuong et al., 2016). Future research into the classification of unidentified repeats

and the identification of ERVs may be key for understanding the evolutionary history and vector competence of different *Culicoides* species.

The phylogenetic relationships among the three *Culicoides* species were also examined. *C. sonorensis* and *C. stellifer* showed closer phylogenetic relationships to each other compared to *C. brevitarsis*, which is consistent with current taxonomic understanding of the genus. Syntenic analysis between the *Culicoides* species revealed less conserved gene order in comparison to two different genera of mosquitoes. Whereas the proportion of genes in syntenic blocks was similar between *Culicoides* and mosquitoes, the *Culicoides* genes were organized into smaller, more numerous blocks and showed a greater degree of spatial rearrangement. In addition to chromosomal rearrangements, gene family expansions also contribute to genomic and functional divergence across species. Several orthogroups were found to be *Culicoides*-specific, suggesting lineage-restricted expansions. Most notably, orthogroup OG0002693 encoded 34 cytochrome P450 4d1-like proteins exclusively from *Culicoides* genomes, with an average of 11.3 copies per genome compared to only 3.14 in Culicidae. Given the role of cytochrome P450 genes in detoxification and insecticide resistance, this *Culicoides*-specific expansion may contribute to important ecological adaptations or resistance mechanisms crucial for midge survival. Together, these genomic differences in repeat architecture, synteny, and gene families highlight the remarkable diversity within the *Culicoides*.

We present a high-quality, chromosome-scale genome and two comprehensive gene annotations for *C. sonorensis*, a significant agricultural pest and confirmed vector of livestock arboviruses in the United States. This assembly represents a major improvement over the previous short-read based genome assembly and provides a much-needed genomic resource for an underrepresented section of dipteran genomes. Through comparative analyses across Diptera,

we examined shared orthologs and syntenic gene order, revealing the deep evolutionary divergence of *Culicoides* from other hematophagous vectors and highlighting the unique opportunities this presents for studying vector biology. Importantly, this genome provides a foundation for future work on vector competence, immune function, and the development of monitoring and control strategies in *Culicoides*, a group that remains understudied despite their global significance. While *C. sonorensis* is one of only two confirmed vector species in the United States, expanding genomic resources to include both vector and non-vector *Culicoides* species will be essential for understanding the genetic basis of vector competence and for guiding the development of novel strategies to limit the transmission of arboviruses.

References

Andrews, S. 2010. FastQC. Computer software.

Baril, T., J. Galbraith, and A. Hayward. 2024. Earl Grey: A fully automated user-friendly transposable element annotation and analysis pipeline. Mol. Biol. Evol. 41: msae068.

Borkent, A., and D. A. Craig. 2004. *Austroconops* Wirth and Lee, a lower Cretaceous genus of biting midges yet living in western Australia: A new species, first description of the immatures and discussion of their biology and phylogeny (Diptera: Ceratopogonidae). Am. Mus. Novit. 3449: 1-67.

Borkent, A., and W. L. Grogan, Jr. 2009. Catalog of the New World biting midges north of Mexico (Diptera: Ceratopogonidae). Zootaxa 2273: 1-48.

Bourque, G., K. H. Burns, M. Gehring, V. Gorbunova, A. Seluanov, M. Hammell, M. Imbeault, Z. Izsvák, H. L. Levin, T. S. Macfarlan, D. L. Mager, and C. Feschotte. 2018. Ten things you should know about transposable elements. Genome Biol. 19: 199.

Boyle, J. H., P. Rastas, X. Huang, A. G. Garner, I. Vythilingam, and P. A. Armbruster. 2021. A linkage-based genome assembly for the mosquito *Aedes albopictus* and identification of chromosomal regions affecting diapause. Insects 12: 167.

Brůna, T., A. Lomsadze, and M. Borodovsky. 2024. GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. Genome Res. 34: 757-768.

Buchfink, B., K. Reuter, and H.-G. Drost. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat. Methods 18: 366-368.

Castellanos-Labarcena, J., Y. Milián-García, T. A. Elliott, D. Steinke, R. Hanner, and S. J. Adamowicz. 2025. Single specimen genome assembly of *Culicoides stellifer* shows evidence of a non-retroviral endogenous viral element. BMC Genomics 26: 1-15.

Chen, S. 2023. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta 2: e107.

Cheng, H., G. T. Concepcion, X. Feng, H. Zhang, and H. Li. 2021. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. Nat. Methods 18: 170-175.

Chuong, E. B., N. C. Elde, and C. Feschotte. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science 351: 1083-1087.

Dainat, J., D. Hereñú, K. D. Murray, E. Davis, I. Ugrin, K. Crouch, LucileSol, N. Agostinho, pascal-git, Z. Zollman, and tayyrov. 2024. NBISweden/AGAT: AGAT-v1.4.1. Computer software.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. 2021. Twelve years of SAMtools and BCFtools. GigaScience 10: giab008.

Diniz, D. F. A., T. P. Romão, E. Helvécio, D. De Carvalho-Leandro, M. D. N. Xavier, C. A. Peixoto, O. P. De Melo Neto, M. A. V. D. Melo-Santos, and C. F. J. Ayres. 2022. A comparative analysis of *Aedes albopictus* and *Aedes aegypti* subjected to diapause-inducing conditions reveals conserved and divergent aspects associated with diapause, as well as novel genes associated with its onset. Curr. Res. Insect Sci. 2: 100047.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. 2013. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29: 15-21.

Dudchenko, O., S. S. Batra, A. D. Omer, S. K. Nyquist, M. Hoeger, N. C. Durand, M. S. Shamim, I. Machol, E. S. Lander, A. P. Aiden, and E. L. Aiden. 2017. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 356: 92-95.

Durand, N. C., J. T. Robinson, M. S. Shamim, I. Machol, J. P. Mesirov, E. S. Lander, and E. L. Aiden. 2016. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 3: 99-101.

Durand, N. C., M. S. Shamim, I. Machol, S. S. P. Rao, M. H. Huntley, E. S. Lander, and E. L. Aiden. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 3: 95-98.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16(1):157.

Emms DM, Kelly S. 2017. STRIDE: Species Tree Root Inference from Gene Duplication Events. Mol. Biol. Evol. 34(12):3267–3278.

Emms DM, Kelly S. 2018. STAG: Species Tree Inference from All Genes. Biorxiv.

Emms, D. M., and S. Kelly. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biol. 20: 238.

Erram, D., and N. Burkett-Cadena. 2021. Oviposition of *Culicoides insignis* (Diptera: Ceratopogonidae) under laboratory conditions with notes on the developmental life history traits of its immature stages. Parasites Vectors 14: 522.

Ewels, P., M. Magnusson, S. Lundin, and M. Käller. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32: 3047-3048.

Gabriel, L., T. Brůna, K. J. Hoff, M. Ebel, A. Lomsadze, M. Borodovsky, and M. Stanke. 2024. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. bioRxiv 2023.06.10.544449.

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler. 2013. QUAST: Quality assessment tool for genome assemblies. Bioinformatics 29: 1072-1075.

Jones, R. H., R. D. Roughton, N. M. Foster, and B. M. Bando. 1977. *Culicoides*, The vector of epizootic hemorrhagic disease in white-tailed deer in Kentucky in 1971. J. Wildl. Dis. 13: 2-8.

Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30: 772-780.

Kurtzer, G. M., V. Sochat, and M. W. Bauer. 2017. Singularity: Scientific containers for mobility of compute. PLoS ONE 12: e0177459.

Laetsch, D. R., and M. L. Blaxter. 2017. BlobTools: Interrogation of genome assemblies. F1000Research 6: 1287.

Li, H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094-3100.

Li, H., and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-595.

Manni, M., M. R. Berkeley, M. Seppey, F. A. Simão, and E. M. Zdobnov. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol. Biol. Evol. 38: 4647-4654.

Marçais, G., A. L. Delcher, A. M. Phillippy, R. Coston, S. L. Salzberg, and A. Zimin. 2018. MUMmer4: A fast and versatile genome alignment system. PLoS Comput. Biol. 14: e1005944.

McGregor, B. L., P. T. Shults, and E. G. McDermott. 2022. A review of the vector status of North American *Culicoides* (Diptera: Ceratopogonidae) for bluetongue virus, epizootic hemorrhagic disease virus, and other arboviruses of concern. Curr. Trop. Med. Rep. 9: 130-139.

Morales-Hojas, R., M. Hinsley, I. M. Armean, R. Silk, L. E. Harrup, A. Gonzalez-Uriarte, E. Veronesi, L. Campbell, D. Nayduch, C. Saski, W. J. Tabachnick, P. Kersey, S. Carpenter, and M. Fife. 2018. The genome of the biting midge *Culicoides sonorensis* and gene expression analyses of vector competence for bluetongue virus. BMC Genomics 19: 624.

Mullens, B. A., E. G. McDermott, and A. C. Gerry. 2015. Progress and knowledge gaps in *Culicoides* ecology and control. Vet. Ital. 51: 313-323.

Nayduch, D., L. W. Cohnstaedt, C. A. Saski, D. J. Lawson, P. J. Kersey, M. Fife, and S. Carpenter. 2014. Studying *Culicoides* vectors of BTV in the post-genomic era: Resources, bottlenecks to progress and future directions. Virus Res. 182: 43-49.

Perez, G., G. P. Barber, A. Benet-Pages, J. Casper, H. Clawson, M. Diekhans, C. Fischer, J. N. Gonzalez, A. S. Hinrichs, C. M. Lee, L. R. Nassar, B. J. Raney, M. L. Speir, M. J. van Baren, C. J. Vaske, D. Haussler, W. J. Kent, and M. Haeussler. 2025. The UCSC Genome Browser database: 2025 update. Nucleic Acids Res. 53: D1243-D1249.

Price, M. N., P. S. Dehal, and A. P. Arkin. 2010. FastTree 2 – Approximately maximum-likelihood trees for large alignments. PLoS ONE 5: e9490.

Purse, B. V., S. Carpenter, G. J. Venter, G. Bellis, and B. A. Mullens. 2015. Bionomics of temperate and tropical *Culicoides* midges: Knowledge gaps and consequences for transmission of *Culicoides*-borne viruses. Annu. Rev. Entomol. 60: 373-392.

Ramírez, F., F. Dündar, S. Diehl, B. A. Grüning, and T. Manke. 2014. deepTools: A flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 42: W187-W191.

Simoni, A., A. Hammond, A. Beaghton, R. Galizi, C. Taxiarchi, K. Kyrou, D. Meacci, M. O. Gribble, G. Morselli, and A. Crisanti. 2020. A male-biased sex-distorter gene drive for the human malaria vector *Anopheles gambiae*. Nat. Biotechnol. 38: 1054-1060.

Stanke, M., O. Keller, I. Gunduz, A. Hayes, S. Waack, and B. Morgenstern. 2006. AUGUSTUS: *Ab initio* prediction of alternative transcripts. Nucleic Acids Res. 34: W435-W439.

Storer, J., R. Hubley, J. Rosen, T. J. Wheeler, and A. F. Smit. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob. DNA 12: 2.

Szadziewski, R. 2016. Biting midges (Diptera: Ceratopogonidae) as indicators of biostratigraphy, ecological reconstructions and identification of amber deposits. Earth Environ. Sci. Trans. R. Soc. Edinb. 107: 219-230.

Tabachnick, W. J. 1996. *Culicoides variipennis* and bluetongue virus epidemiology in the United States. Annu. Rev. Entomol. 41: 23-43.

Tanya, V. N., E. C. Greiner, and E. P. J. Gibbs. 1992. Evaluation of *Culicoides insignis* (Diptera: Ceratopogonidae) as a vector of bluetongue virus. Vet. Microbiol. 32: 1-14.

Uliano-Silva, M., J. G. R. N. Ferreira, K. Krasheninnikova, Darwin Tree of Life Consortium, M. Blaxter, N. Mieszkowska, N. Hall, P. Holland, R. Durbin, T. Richards, P. Kersey, P. Hollingsworth, W. Wilson, A. Twyford, E. Gaya, M. Lawniczak, O. Lewis, G. Broad, F. Martin, and S. A. McCarthy. 2023. MitoHiFi: A python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. BMC Bioinformatics 24: 288.

Wang, Y., H. Tang, X. Wang, Y. Sun, P. V. Joseph, and A. H. Paterson. 2024. Detection of colinear blocks and synteny and evolutionary analyses based on utilization of MCScanX. Nat. Protoc. 19: 2206-2229.

Wiegmann, B. M., M. D. Trautwein, I. S. Winkler, N. B. Barr, J.-W. Kim, C. Lambkin, M. A. Bertone, B. K. Cassel, K. M. Bayless, A. M. Heimberg, B. M. Wheeler, K. J. Peterson, T. Pape, B. J. Sinclair, J. H. Skevington, V. Blagoderov, J. Caravas, S. N. Kutty, U. Schmidt-Ott, and D. K. Yeates. 2011. Episodic radiations in the fly tree of life. Proc. Natl. Acad. Sci. USA 108: 5690-5695.

Xu, X., J. Chen, Y. Wang, Y. Liu, Y. Zhang, J. Yang, X. Yang, B. Chen, Z. He, and J. Champer. 2025. Gene drive-based population suppression in the malaria vector *Anopheles stephensi*. Nat. Commun. 16: 1007.

Chapter 3 - Transcriptional response to arbovirus infection in *Culicoides sonorensis* biting midge

In preparation as: Edward Bird, Brandon Hall, Kristopher Silver, Dana Nayduch. Common transcriptomic responses of *Culicoides sonorensis* biting midges to infection with different arboviruses

Abstract

Culicoides biting midges are vectors of economically important arboviruses that affect the livestock industry, yet the molecular mechanisms of vector competence remain relatively unknown. In this study, we examined the transcriptomic response of *Culicoides sonorensis* following infection with bluetongue virus serotype 17 (BTV-17) and vesicular stomatitis New Jersey virus (VSNJV) using RNA sequencing and the newly assembled chromosome-scale genome of *C. sonorensis*. Female midges were injected with purified virus or PBS controls and whole-body transcriptomes were analyzed five days post-infection. BTV infection elicited a much broader transcriptional response than VSV, with 862 differentially expressed genes (DEGs) for BTV versus 96 DEGs for VSV. Both infections induced expression of components of innate immune pathways, such as prophenoloxidase cascade, Toll signaling, endosomal sorting complex required for transport machinery (ESCRT) genes. Of note, 55 DEGs were shared among both viral infection treatments, suggesting a core antiviral response. Functional enrichment revealed that BTV infection significantly upregulated immune-related pathways but suppressed expression of genes associated with neurological functions, including synaptic transmission and behavioral regulation. A large proportion of the DEGs were uncharacterized proteins or long non-coding RNAs, reflecting the unique biology of *Culicoides* vectors and suggesting clade-specific antiviral responses. These findings provide novel insights into

Culicoides-virus interactions and outline future targets for vector control strategies, while underscoring the need for greater functional characterization of these important, understudied vectors.

Introduction

Culicoides biting midges (Diptera: Ceratopogonidae) are hematophagous biting flies that cause significant economic damage to both farmed and wild ruminants both through pestiferous biting activities and by transmitting arboviruses that cause morbidity or mortality. Over 1,347 *Culicoides* species have been described globally, yet only a small fraction are known as competent disease vectors. In the United States, only two of the 141 recognized species have been confirmed as vectors of arboviruses, *Culicoides sonorensis* and *Culicoides insignis* (Jones et al., 1977; Tabachnick, 1996; Tanya et al., 1992). *C. sonorensis* has become the primary model organism for biting midge research due to its established laboratory colony and cell line, as well as serving as a competent vector for three economically important arboviruses: bluetongue virus (BTV), epizootic hemorrhagic disease virus (EHDV), and vesicular stomatitis virus (VSV). Whereas prior research assessed the response of *C. sonorensis* to infection with EHDV (Nayduch et al., 2019), limited work has been done to characterize *C. sonorensis* responses to infection with BTV and VSV.

VSV is a zoonotic rhabdovirus that can be transmitted by multiple hematophagous vectors, including *Culicoides* biting midges (Diptera: Ceratopogonidae), mosquitoes, black flies, and sand flies (Rozo-Lopez et al., 2018). While not endemic to the U.S., periodic incursions from central and southern endemic regions of Mexico trigger outbreaks that necessitate quarantine and movement restrictions of U.S. livestock (Pelzel-McCluskey et al., 2021). Although VSV primarily affects horses, it can also infect cattle, swine, sheep, goats, and other

livestock. Clinically, VSV presents symptoms similar to foot-and-mouth disease, including oral lesions, lameness, and weight loss, making differential diagnosis challenging without further testing (Alderink, 1984; Hayek et al., 1998).

In contrast, BTV is a non-zoonotic orbivirus transmitted exclusively by biting midges of the genus *Culicoides*. This virus is endemic to the U.S. where it causes periodic outbreaks in livestock and wild ruminants, with the latter serving as viral reservoirs (Drolet et al., 1990; Gaydos et al., 2002). BTV primarily affects sheep and cattle but can infect a wide range of ruminants, causing fever, facial and tongue swelling, lameness, reproductive failure, and mortality, particularly in naïve populations. Globally, BTV outbreaks have resulted in billions of dollars in economic losses through the costs of vaccination, reduced livestock productivity, and restricted trade of infected animals or products (Alkhamis et al., 2020; Gethmann et al., 2020; Häsler et al., 2012; Rushton et al., 2015). Despite their economic importance, the molecular interactions between arboviruses and *Culicoides* spp. remain poorly understood compared to well-characterized mosquito-arbovirus systems (Cheng et al., 2016), yet this field of study offers opportunities to uncover novel transmission-blocking targets.

Both VSV and BTV establish persistent, non-lytic infections in *C. sonorensis*, allowing infected midges to complete their full life cycle while harboring and transmitting the virus (Labadie et al., 2020). This non-lethal persistence is particularly significant because female midges require multiple blood meals for egg development, typically feeding up to four times over their lifetime (Mullens & Schmidtmann, 1982), greatly enhancing their capacity as vectors. This is in contrast with the acute, lytic infections these same viruses cause in animals. The molecular mechanisms underlying host-specific differences in viral pathology remain unknown,

yet these differences in pathology may be central to understanding the complex interactions between *Culicoides* midges and the arboviruses that they transmit.

Differential expression studies, such as via transcriptomic analyses, have previously identified genes crucial to viral infection and transmission in mosquito-virus systems (Bellone et al., 2023; David W. Severson et al., 2016), and have also provided some insight in *Culicoides* midges. Previous transcriptomic studies have provided insights into host-pathogen interactions in *C. sonorensis*. Infection with EHDV, another a closely related orbivirus, triggers upregulation of several innate immune pathways in *C. sonorensis*, consistent with patterns observed in arbovirus-infected mosquitoes (Nayduch et al., 2019). Additionally, EHDV infection caused significant differential expression of vision-related genes, consistent with the heavy ocular viral load observed with BTV and EHDV infected *C. sonorensis* (McDermott et al. 2015, Mills et al. 2017). Supporting this finding, UV-light traps were less effective at capturing BTV-infected midges compared to uninfected individuals, suggesting that infection may alter visual function or associated phototaxis behavior (McDermott et al., 2015). Conversely, VSV infection of the *C. sonorensis* W8 cell line produced only a transient transcriptomic response limited to early infection stages (Scroggs et al., 2023). Notably, vision-associated genes were not differentially expressed during VSV infection in the W8 cell line, indicating these effects may be virus-specific or the result of differences in gene regulation between *in vivo* and *in vitro* systems (e.g., differentiated cells in whole insects vs. nondifferentiated cell lines). Additionally, due to the fragmentation of the previously available genome (Morales-Hojas et al., 2018), both studies relied on assembly of *de novo* transcriptomes for read mapping, which can limit the power of differential expression analysis to identify transcript levels altered by viral infection.

Here, we characterized the transcriptomic responses of *C. sonorensis* to infection with BTV and VSV using the newly assembled chromosome-scale *C. sonorensis* genome (Chapter 2). We analyzed virus-specific gene expression patterns as well as shared responses across both infections to identify functional pathways associated with innate immune responses to viral infection. This work provides insight into differential immune responses by *C. sonorensis* midges to infections with different viruses, mechanisms that may cause behavioral alterations following infection with BTV, and potential targets for future study of virus-vector interactions in *C. sonorensis*.

Methods

Virus Stock Generation

Stocks of purified virus were generated in midge cell lines prior to infection treatments. *C. sonorensis* W8 cells were maintained at 27 °C in Schneider's insect medium (MilliporeSigma, St. Louis, MO, USA) supplemented with sodium bicarbonate, L-glutamine, reduced glutathione, L-asparagine, bovine insulin, and 5% fetal bovine serum (FBS). Cells were infected with either BTV-17 or VSV (VSV-New Jersey) at a multiplicity of infection of 0.1 and incubated for 5 days. Cells were lysed by two freeze-thaw cycles, and cell debris was removed by centrifugation at 1,500 × g for 30 min at 4 °C. VSV and BTV supernatants were further purified by ultracentrifugation at 100,000 × g through a 30% sucrose cushion to pellet virus. Pelleted virus was resuspended in phosphate-buffered saline (PBS) and stored at −80 °C. Final viral titers were determined via plaque assay on Vero MARU cells (Middle America Research Unit, Panama), which were maintained at 37 °C with 5% CO_2 in 199E medium supplemented with 2% FBS and standard antibiotics. The final titers used in experiments were 3.5×10^9 PFU/mL for VSNJV and 9.14×10^6 PFU/mL for BTV.

Midge Treatments and Sample Generation

The AK colony of *C. sonorensis* (Wirth & Jones) biting midges were maintained at the USDA-ARS in Manhattan, KS. Pupae were collected and sexed. Female pupae (n = 50) were placed into six separate cages, placed on moistened substrate, and kept in a growth chamber with a 13:11 h light:dark cycle until adults emerged. Four days after emergence, midges were anesthetized with CO_2 and each cage was inspected for male contamination. Cages without males were pooled into a single group which was then randomly divided into three treatment groups: VSV-injected, BTV-injected, and PBS-injected (control). Injection of 49 nL of purified VSV, purified BTV, or PBS were performed through the dorsal thorax using a Nanoject II (Drummond Scientific, Broomall, PA). Injected midges were placed in cages by treatment and maintained under the same growth conditions. Five days post-injection, individual midges were anesthetized with CO_2, euthanized in 100% ethanol, and transferred to 1.5 mL microcentrifuge tubes containing ceramic maceration beads (Revvity, Kennesaw, GA) and 100 µL of DNA/RNA Shield (Zymo Research, Irvine, CA).

RNA extraction

Samples were homogenized immediately after collection using an Omni Bead Ruptor Elite (Omni, Kennesaw, GA) at 3.1 m/s for 2 min. Total RNA was extracted from individual midges using a protocol modified from the Zymo Quick-RNA Microprep Kit (Zymo Research, Irvine, CA). To each macerated sample, 100 µL of RNA lysis buffer was added, vortexed, and 200 µL of absolute ethanol was added. The mixture was vortexed again, briefly centrifuged (10,000 × g, 60 sec), and passed through a Zymo-Spin III-F Filter (Zymo Research) into a clean 1.5 mL microcentrifuge tube. The filtered homogenate was then purified with a spin column according to the manufacturer's instructions (Zymo Research). Due to the low RNA yield from

single midges, 2 µL of each sample was diluted 1:10 for quality control, while the remaining RNA was stored at −80 °C until used for library preparation. RNA concentrations were quantified using a Qubit Flex fluorometer (Thermo Fisher Scientific, Waltham, MA) with the Broad Range RNA Quantification Kit (Thermo Fisher Scientific). RNA purity was assessed using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific), and RNA integrity was evaluated using a TapeStation 4150 with the high sensitivity RNA ScreenTape (Agilent, Santa Clara, CA).

Viral Infection Confirmation

All samples were tested for viral infection prior to library preparation using reverse transcription quantitative PCR (RT-qPCR). For VSNJV, RT-qPCR was performed as previously described (Rozo-Lopez et al., 2020). Primer and probe sequences are provided in Supplementary Table 3.1. TaqMan™ Fast Virus 1-Step Master Mix (Thermo Fisher Scientific) was used in a 20 µL reaction containing 0.8 µL of forward primer, 1.8 µL of reverse primer, 0.5 µL of TaqMan probe, 8.9 µL of nuclease-free water, and 3 µL of RNA template. Reactions were run on a Bio-Rad CFX Opus 96 thermocycler under the following conditions: reverse transcription at 50 °C for 5 min; initial denaturation at 95 °C for 20 s; followed by 40 cycles of 95 °C for 15 s and 60 °C for 60 s. For BTV, segment 10 was targeted for quantification as previously described (Drolet et al., 2013). RT-qPCR was carried out in a 25 µL reaction using the AgPath-ID™ One-Step RT-PCR Kit (Thermo Fisher Scientific), consisting of 12.5 µL of 2× buffer, 1.05 µL of 10 µM primer mix, 0.2 µL of 10 µM probe, 1 µL of 25× enzyme mix, 5.25 µL of DEPC-treated water, and 5 µL of RNA template. Thermal cycling was performed on the CFX Opus 96 with the following protocol: reverse transcription at 48 °C for 10 min; initial denaturation at 95 °C for 10 min; followed by 40 cycles of 95 °C for 15 s and 60 °C for 1 min. All reactions were performed

in duplicate, with viral RNA positive controls and no-template negative controls included in each run.

Library Preparation and Sequencing

Stranded mRNA-enriched libraries were prepared using the Watchmaker mRNA Library Prep Kit (Watchmaker Genomics, Boulder, CO) following the manufacturer's instructions. Samples prepared from the three treatments, BTV-, VSV-, or PBS-injected were processed individually (N = 8 per treatment; total N = 24). Briefly, mRNA was captured with mRNA capture beads, fragmented, and reverse transcribed to generate cDNA. Stubby adapters (Element Biosciences, San Diego, CA) were ligated onto cDNA fragments, which were then purified using Ampure XP beads (Beckman Coulter, Brea, CA). Indexes were incorporated during library amplification via 10 PCR cycles, with two additional cycles to accommodate the targeted insert size of 290–310 base pairs. Final libraries were evaluated for size distribution using an Agilent 4150 TapeStation with a D5000 reagent kit (Agilent). Library concentration was quantified using a Qubit Flex fluorometer and the 1×dsDNA High Sensitivity kit (Thermo Fisher Scientific). Libraries were subsequently sequenced on an Element Aviti platform (Element Biosciences) using the AVITI 2×150 Cloudbreak sequencing kit (Element Biosciences).

Demultiplexing and Quality Control

Demultiplexing was performed using bases2fastq (https://github.com/Elembio/bases2fastq-nf), implemented in Nextflow (Di Tommaso et al., 2017) and executed via Apptainer (Gregory M. Kurtzer et al., 2017). Adapter trimming was enabled during demultiplexing, and single-ended 75 bp reads were called. Because residual adapter contamination remained after bases2fastq processing, likely due to adapters appearing in unconventional positions from rolling circle amplification, we developed a Nextflow pipeline,

edwardbirdlab/element_cleanup (https://github.com/edwardbirdlab/element_cleanup), to further trim Element adapters from raw reads. This pipeline utilized BBDuk v38.96 (https://sourceforge.net/projects/bbmap/) to detect and trim adapters anywhere within reads using a k-mer–based approach. Reads were subsequently quality trimmed and checked for remaining adapter contamination using FastP v0.24.0 (Chen, 2023). For this study, reads were trimmed to a minimum length of 60 bp and an average quality score of 20. Quality control reports generated by FastQC v0.12.1 (Andrews, 2010), FastP, and BBDuk were consolidated and visualized using MultiQC v1.22.2 (Philip Ewels et al., 2016).

Differential Expression Analyses

RNA sequencing data were processed using the nf-core/rnaseq pipeline v3.18.0 (Harshil Patel et al., 2025), a best-practices workflow for RNA-seq data processing and quality control. The pipeline was executed with Nextflow v24.04.2 and Singularity v1.3.1. Raw sequencing reads were initially assessed for quality using FastQC v0.12.1. Low-quality bases were trimmed with Trim Galore! V0.6.10 (Krueger et al., 2023). Trimmed reads were aligned to the *C. sonorensis* reference genome and eGAPx annotation (Chapter 2) using STAR v2.7.11b (Dobin et al., 2013). Transcript quantification was calculated with Salmon v1.10.3 (Patro et al., 2017) and transcripts per million (TPM) normalized counts were generated. Comprehensive quality assessment included RSeQC v5.0.2 (Wang et al., 2012) for read distribution and junction saturation; Qualimap v2.3 (García-Alcalde et al., 2012) for alignment quality metrics and feature coverage; and dupRadar v1.32.0 (Sayols et al., 2016) to evaluate PCR duplication rates relative to gene expression levels.

Differential expression analysis was performed using DESeq2 v1.46.0 (Love et al., 2014) in R v4.2 (Core Team, 2021). Transcript-level quantifications were summarized to the gene level

using the tximport v1.34.0 (Soneson et al., 2015) package prior to import into DESeq2. The effects of BTV and VSV infections on transcript levels were evaluated independently and compared to the PBS-injected controls by fitting separate models for each condition. Principal component analysis was conducted using the prcomp function, using rlog transformed counts generated with DESeq2. Sample correlation heatmaps were generated from pairwise Pearson correlations calculated using the cor function and visualized using the pheatmap v1.0.12 (Kolde R, 2019). For both datasets, a Wald test was conducted using the design formula ~ Treatment. Differential expression was assessed by contrasting infected samples against uninfected controls, with adjusted p-values calculated using the Benjamini-Hochberg method to control for multiple testing. Two statistical thresholds were applied to define significance: strict and loose. The strict threshold tested whether gene expression was significantly different from an absolute $\log_2$ fold change (FC) of 0.58 (1.5-fold), using a null hypothesis of $|\log_2 FC| = 0.58$ and an adjusted p-value < 0.05. The loose threshold tested for any significant change from baseline expression, using a null hypothesis of $|\log_2 FC| = 0$ and an adjusted p-value < 0.05. This approach yielded two sets of differentially expressed genes (DEGs). Volcano plots were generated with EnhancedVolcano v1.24.0 (Blighe, 2018) using the loose DEG set. To identify genes consistently regulated in both BTV and VSV infections, concordance analysis was performed by plotting average $\log_2$ fold changes from the BTV experiment against those from the VSV experiment. Genes with $|\log_2 FC| > 1$ in both conditions were considered concordant if their expression vectors fell within 15° of the 45° (upregulated) or 225° (downregulated) axis, indicating similar magnitude and direction of regulation across both viral infections.

Term Enrichment

To better understand the biological relevance of the DEGs, orthologs between *C. sonorensis* and *Drosophila melanogaster* were identified. Using the previously established orthogroups (Chapter 2, Supplementary Table 3.7), each *C. sonorensis* gene was assigned its corresponding *D. melanogaster* ortholog. Orthogroups were defined using OrthoFinder3. For orthogroups containing multiple *C. sonorensis* or *D. melanogaster* proteins, gene trees were used to identify the ortholog with the shortest phylogenetic distance, which was selected as the representative ortholog. A genomic background set was constructed by including all genes tested in both RNA-seq experiments that had a corresponding *D. melanogaster* ortholog (16,260 out of 18,776 total tested transcripts). The six gene sets defined earlier, which included the significantly upregulated (adjp < 0.05 & $\log_2$FC > 0) and downregulated (adjp < 0.05 & $\log_2$FC < 0) genes from both the BTV and VSV experiments, as well as the upregulated and downregulated genes identified as concordant between the two viral infections, were also converted to *D. melanogaster* orthologs. STRING v12.0 analysis (Szklarczyk et al., 2023) was performed for all six gene sets using their *D. melanogaster* orthologs as input. Analyses were conducted through the STRING web interface (https://string-db.org) using default settings. Enrichment analysis was performed in STRING using the full set of 7,705 *D. melanogaster* orthologs as the background for statistical comparison.

Results

Quality Control of RNA-Seq

Initial RNA-seq data were processed to remove Aviti-specific sequencing artifacts, quality filtered and aligned to the *C. sonorensis* reference genome. This resulted in the alignment of 14.0 million (sample B14) to 22.5 million (sample P5) reads per sample. Using gene

annotations generated by eGAPx (Chapter 2), between 90.7% and 92.7% of reads mapped to annotated exons, 3.7%–5.0% to introns, and 3.4%–4.3% to intergenic regions. Duplication bias was assessed with dupRadar, which showed no evidence of duplication related bias. The BTV and VSV experiments tested 10,555 and 10,565 genes, respectively, that met the minimum expression threshold of 10 reads per gene. Principal component analysis (PCA) was used to evaluate transcriptional structure in relation to treatment conditions (Figure 3.1). In the BTV experiment, PC1 accounted for 19.83% of the total variance and showed clear separation between treatment groups (Figure 3.1A). In contrast, the VSV experiment exhibited no clear separation in PC1–PC3, with only PC4 showing structure related to infection status (9.20% variance explained) (Figure 3.1B).

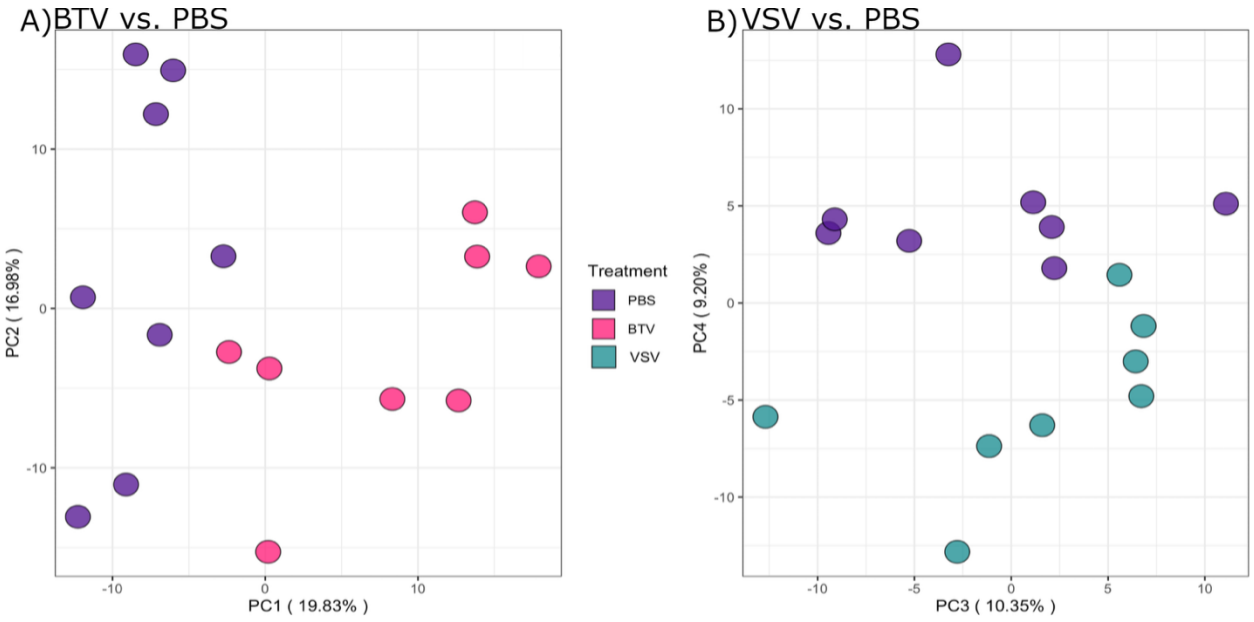


Figure 3.1 Principal component analysis (PCA) of BTV and VSV RNA-seq experiments. A) Ordination of BTV-infected and PBS control samples. Samples show separation along principal component 1 (PC1), which explains 19.83% of the total variance. B) Ordination of VSV-infected and PBS control samples. In this analysis, sample separation related to infection status is captured by principal component 4 (PC4), which accounts for 9.2% of the variance.

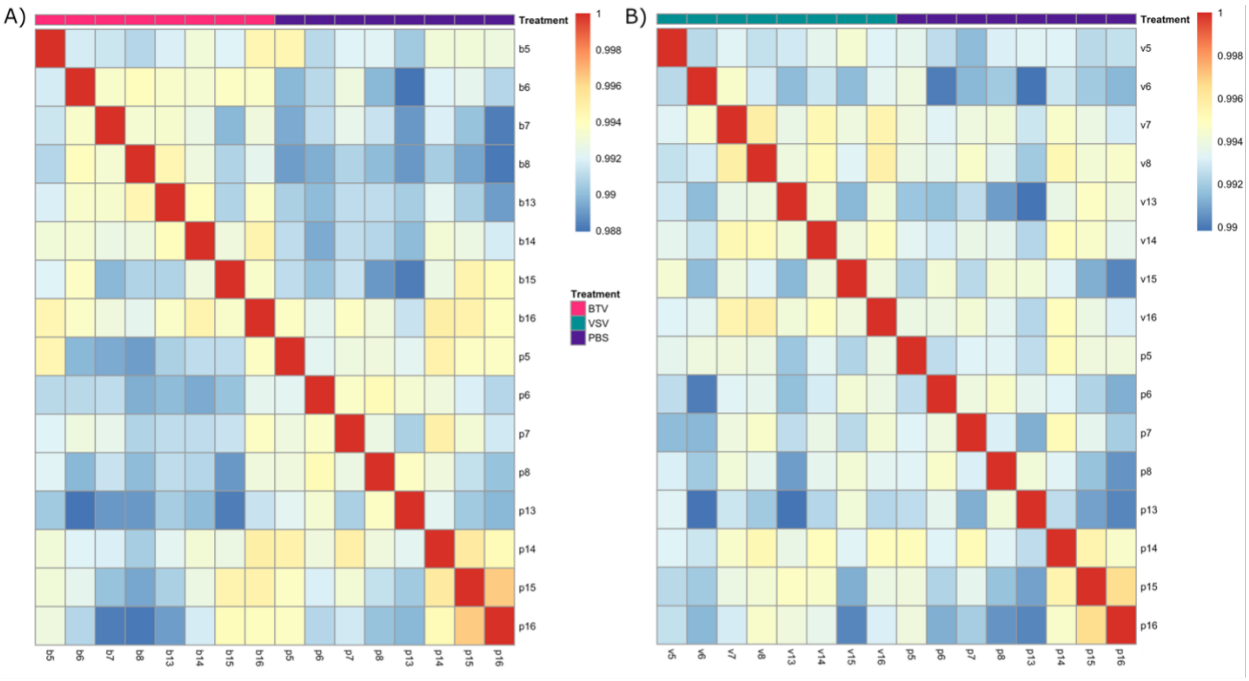


Figure 3.2 Pairwise correlation heatmap based on regularized log-transformed counts. Pairwise Pearson correlation coefficients were calculated from rlog-transformed counts using SummarizedExperiment (Martin Morgan, 2017) and visualized using the Phetmap (Kolde, 2019) packages in R. The color gradient indicates correlation strength between samples. A) BTV-infected vs. PBS injected control samples. B) VSV-infected vs. PBS injected control samples.

To further evaluate sample variability and detect potential outliers, a heatmap of pairwise Pearson correlation coefficients was generated using regularized log-transformed (rlog) counts. Although individual variability was observed within treatment groups, no obvious outliers were identified (Figure 3.2). Finally, dispersion estimates were examined to evaluate the fit of the negative binomial model used by DESeq2. Dispersion plots from both experiments showed that most genes conformed to model expectations without excessive shrinkage, indicating that the fitted models appropriately captured variance in the datasets (Appendix C Figure C.1 & C.2,).

Differential Expression

Differential expression analysis of the BTV-infected versus PBS-injected control midges identified 18 genes under the strict threshold (7 upregulated, 11 downregulated) (Supplementary Table 3.3), and 862 genes under the loose threshold (414 upregulated, 448 downregulated) (Figure 3.3A) (Supplementary Table 3.4). Several biologically relevant genes known to be associated with antiviral or immune responses in insects were identified, including immune-related genes and components of the prophenoloxidase (proPO) cascade. These included *phenoloxidase 8-like* (*PPO8-like*), *phenoloxidase 2-like* (*PPO2-like*), serine protease *snake-like*, and *gram-negative bacteria-binding protein 3-like* (*GNBP3-like*), all of which were upregulated. Additional immune genes also showed elevated expression, including *Dicer-2* (*DCR-2*) from the RNAi pathway, lysosome-associated genes such as *lysozyme C-1-like* (*Lyz1-like*) and *V-type proton ATPase subunit a1-like* (*ATPA6V0A1*), and components of the ESCRT machinery (*CHMP5, CHMP2A*, and *VPS36-like*). Toll pathway members *Toll-like receptor 6* (*TLR6*) and *MY06-like* were also upregulated. In addition to immune-related pathways, genes associated with apoptosis, including *Draper* and *dronc-like*, were also differentially expressed. Among the most strongly upregulated genes was *vacuolar protein sorting 36-like* (*VPS36-like*), with a fold change of 36.0. The highest fold change observed overall was for *protein mab-21-like*, which was upregulated 44.2-fold. Many of the highly downregulated genes, including the four most strongly downregulated DEGs, *cson_001203* (−8.62 FC), *cson_008488* (−9.61 FC), *cson_010602* (−12.48 FC), and *cson_013604* (−14.51 FC), were uncharacterized, making evaluation of the biological importance of these genes difficult.

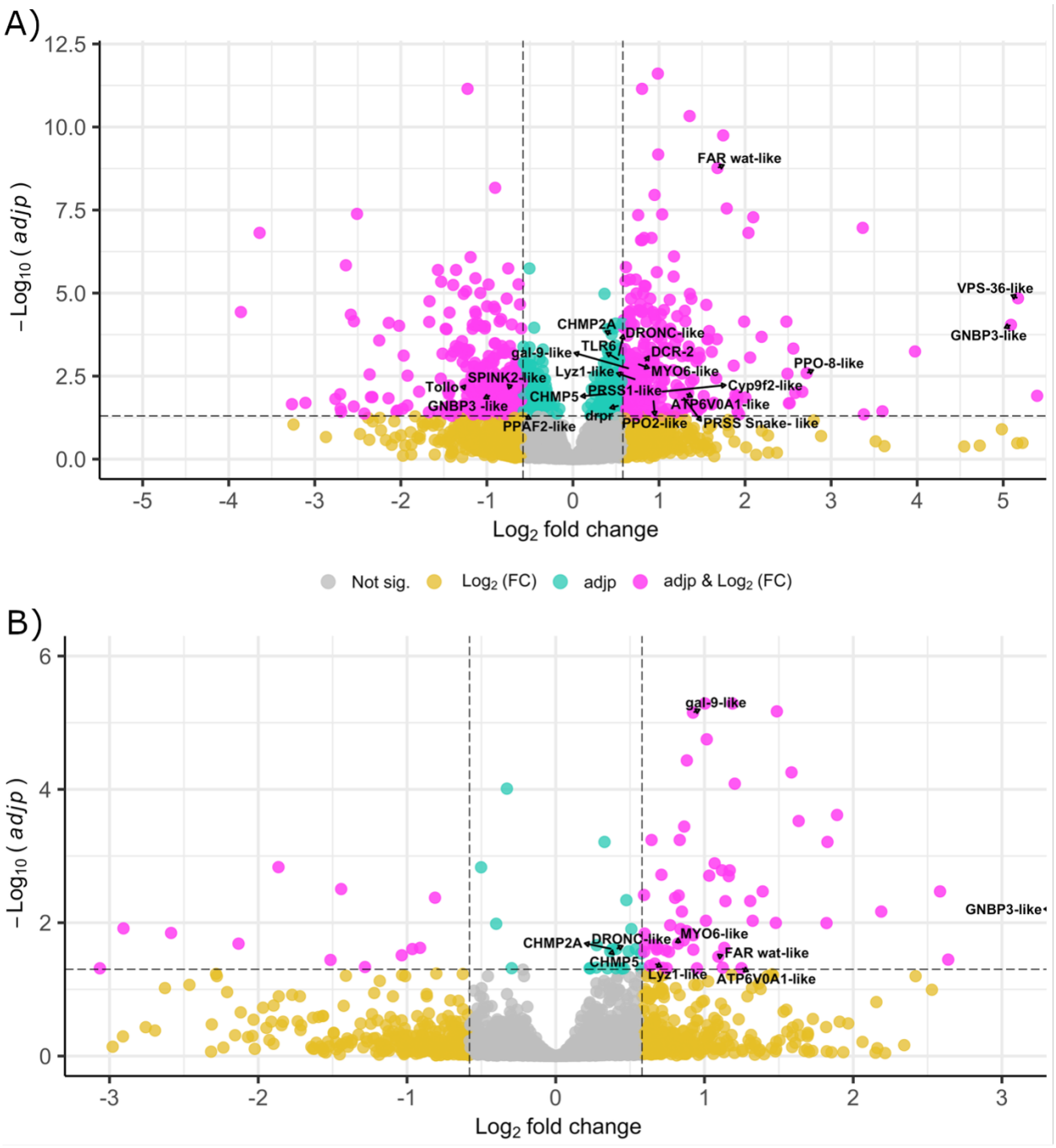


Figure 3.3 Volcano plot of genes showing significant changes in expression following infection with BTV (A) or VSV (B).

Volcano plots were generated with EnhancedVolcano (Blighe, K, 2018) in R. Genes were considered significant if they had an adjusted p-value < 0.05. A fold change cutoff of 1.5-fold ($|\log_2 FC| \geq 0.58$) was applied *post hoc* for visualization. Selected genes of interest are labeled.

As in the BTV experiment, differential expression in the VSV versus PBS comparison was assessed using both strict and loose statistical thresholds. Under the strict threshold, only a single upregulated gene, *cson_011589*, showed a significant change in expression (Supplementary Table 3.5). However, this gene was likely a statistical outlier, as its expression was inconsistent, detected in only two VSV samples, one PBS sample, and two BTV samples. Furthermore, *cson_011589* is uncharacterized and has no identified ortholog in *D. melanogaster*. Using the loose threshold, 96 genes were found to be differentially expressed, comprising 79 upregulated and 17 downregulated genes (Figure 3.3B) (Supplementary Table 3.6). Several of these genes overlapped with those identified in the BTV experiment, including components of the ESCRT pathway (*CHMP2A, CHMP5*), lysozyme gene *Lyz1-like* and immune-related genes such as *GNBP3-like* and *galectin-9-like* (*gal-9-like*). The apoptosis related gene *dronc-like* was also upregulated in response to VSV infection. Of the 96 DEGs identified in the loose VSV dataset, 55 overlapped with the loose DEG set from the BTV experiment (47 upregulated and eight downregulated). A substantial proportion of DEGs in the VSV response were uncharacterized: 29 genes (30.2%) lacked functional annotation, and five were identified as long non-coding RNAs (lncRNAs). For comparison, in the BTV experiment, 179 of 862 DEGs (20.8%) were uncharacterized, and 39 were lncRNAs. Among the 17 downregulated genes in the VSV dataset, seven were uncharacterized, and two were lncRNAs, leaving only nine annotated protein-coding genes. Notably, two of these were histone H3 genes: *cson_005632* (–2.71FC) and *cson_007247* (–2.05FC).

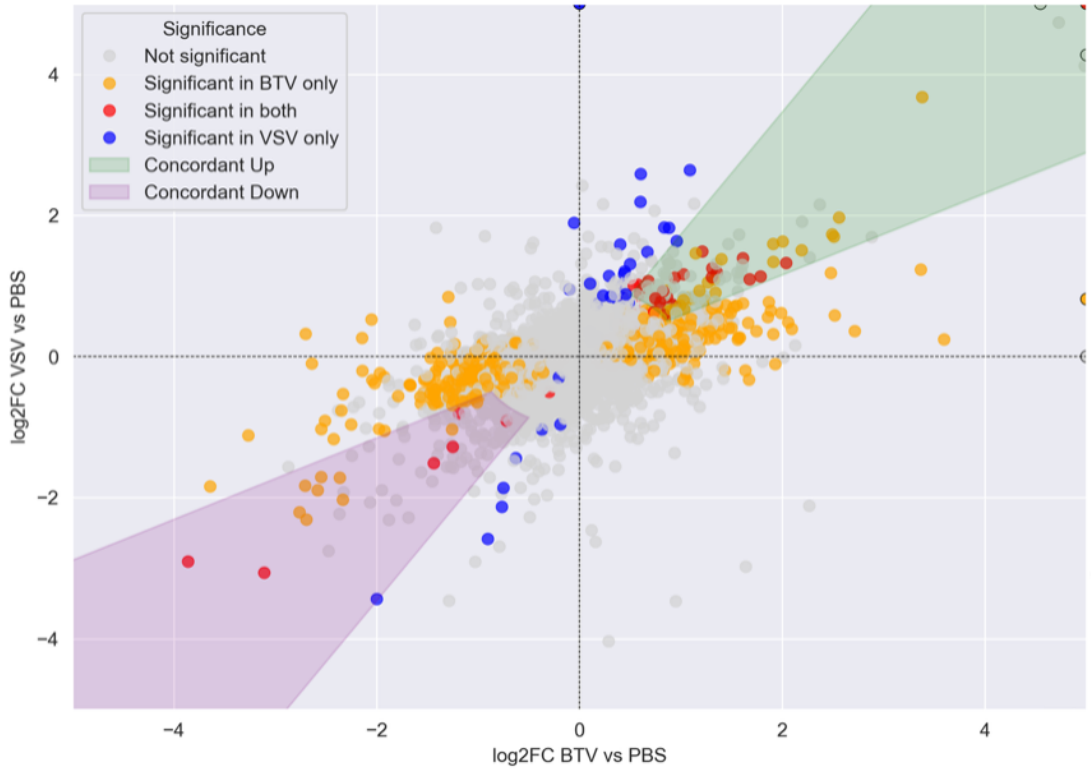


Figure 3.4 Concordance analysis and comparison of expression of key genes across virus infected and PBS injected control midges.

Average Log_2(FC) of BTV experiment plotted against average Log_2(FC) of VSV experiment for every gene examined (n = 10,555). Significance is defined as an adjusted p-value ≤ 0.05, with a test of Log_2(FC) significantly different from 0. Concordance was selected based on a cutoff of Log_2(FC) > 1 (upregulated) or < 1 (downregulated) and within 15 degrees of perfect concordance (45 and 225 respectively).

Since many of the genes that showed significant changes in expression due to VSV infection were similarly observed in the BTV gene set, an expression concordance analysis was conducted to identify genes that shared differential expression between the treatments that may not have been originally identified as significant. Plotting the average Log_2(FC) of genes in the BTV experiment vs. the VSV experiment showed that genes with significant changes in expression in both experiments appeared to be concordant, having equal relative expression level in both experiments (Figure 3.4A). We identified genes in near-perfect concordance between

experiments as those with an $|Log_2(FC)| > 1$ in both experiments and expression directions within 15 degrees of perfect concordance. This resulted in two gene sets, concordant upregulated (n = 40), and concordant downregulated (n = 44). Additionally, genes like *Dicer-2*, which was not identified as significant in the VSV experiment, show very similar patterns of expression across the replicates when observing normalized gene counts. This trend does not extend to some pathways, like the ProPO pathway, where genes like *PPO-8-like*, *PPO-2-like*, and *PRSS-snake-like* do not show similar expression patterns across the viral treatments. As seen with the BTV and VSV significant gene sets, a large portion of the concordant gene sets (32 of the 84 genes; 38.1%) were made up of uncharacterized proteins. Additionally, this gene set contained 18 lncRNA's, leaving only 34 characterized protein coding genes.

Functional Characterization

To gain deeper biological insights into the differentially expressed genes, *D. melanogaster* orthologs were identified for all possible transcripts in the *C. sonorensis* genome. While 16,259 out of 18,776 (86.6%) mRNAs had an identified *D. melanogaster* ortholog, some sets of DEGs did not have as high of an identification rate. In the BTV experiment orthologs for 376 out of 448 (83.92%) of the upregulated DEGs and 351 out of 414 (84.78%) downregulated DEGs were found. In the VSV experiment, orthologs for 58 out of the 79 (73.42%) upregulated and 10 out of the 17 (58.82%) downregulated DEGs were identified. In the set of concordant genes, orthologs were recognized for 21 out of 40 (52.5%) upregulated and 20 out of 43 (46.51%) downregulated genes. Not all the identified orthologs had known functions or proteins, with only 7,705 out of 16,259 identified orthologs having functional representation in StringDB. This led to only three gene sets in the BTV up/down regulated genes, and the upregulated genes from VSV, containing a large enough gene set to perform functional enrichment.

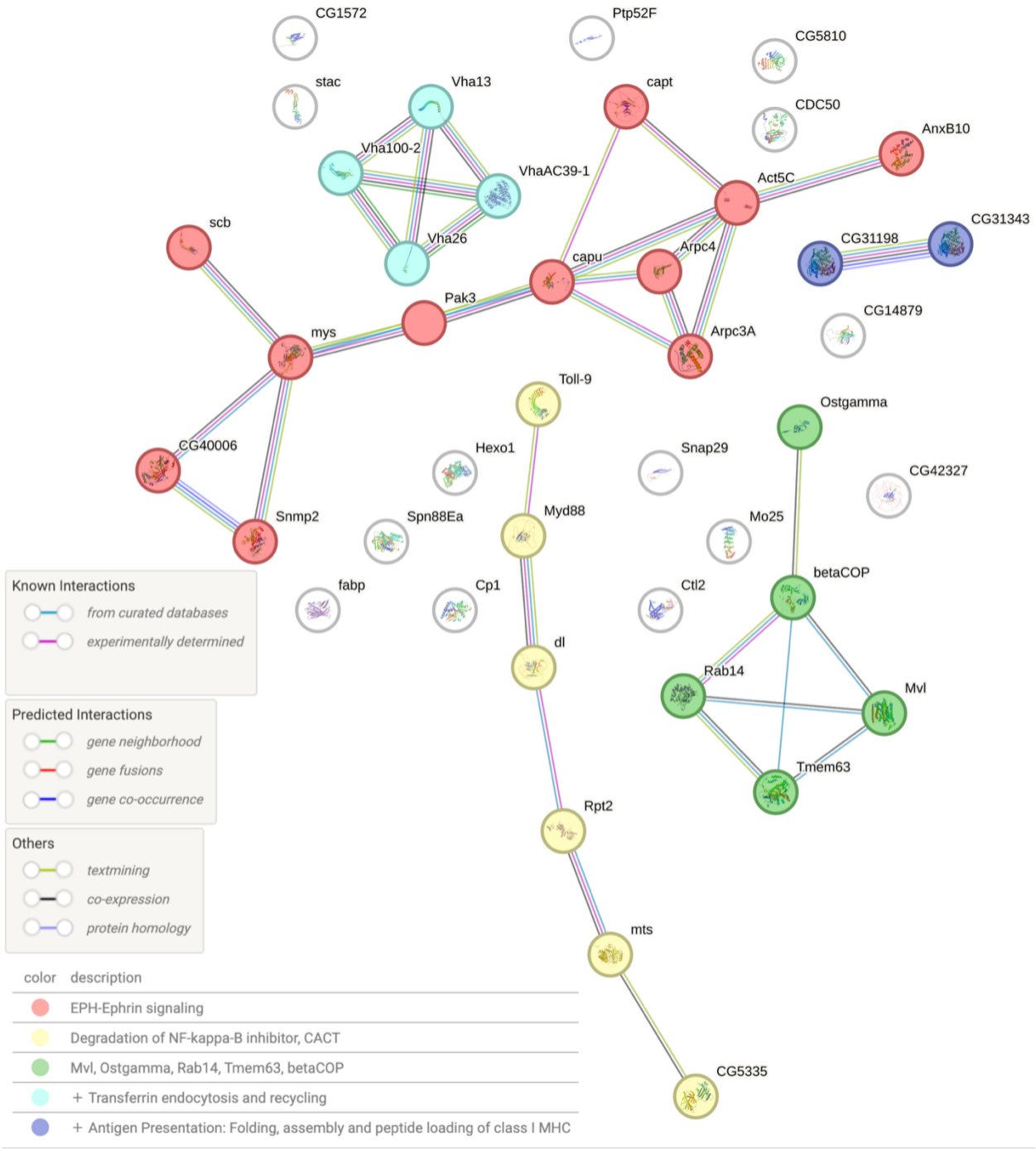


Figure 3.5 Interaction networks of differentially-expressed *C. sonorensis* genes associated with the insect immune response.

Genes associated with the innate immune system (Reactome Pathway: DME-168249) that were upregulated in response to infection with BTV. Interactions between genes are shown as edges and colored according to the interaction type. Genes were clustered by k-means (k = 5).

The enrichment of terms from BTV infected midges revealed many functions and pathways that were differentially expressed (Supplementary Table 3.8 & 3.9). In the upregulated gene set, gene ontology (GO) terms related to transport including transmembrane transporter activity (FDR = 0.002) and transmembrane transport (FDR = 0.0027) were enriched. GO terms related to cellular organization were also observed, including basement membrane organization (FDR = 0.0241), actin filament organization (FDR = 0.0236), and supramolecular fiber organization (0.0241). Additionally, several immune related pathways including the KEGG pathways ECM-receptor interaction (7 of 10 genes, FDR = 0.00048) and phagosome (9 of 43 genes, FDR = 0.0229), and the reactome pathway innate immune system (42 of 446 genes, FDR = 0.0087). The innate immune system pathway revealed the presence of five different clusters of gene associates (Figure 3.5). This includes 11 DEGs associated with the Ephrin-Eph signaling pathway, six DEGs associated with a toll signaling pathway, and two DEGs associated with antigen detection. This cluster analysis also identified a cluster of genes (*mvl*, *ostgamma*, *rab14*, *tmem63*, and *betaCOP*) that are related to neutrophil degranulation.

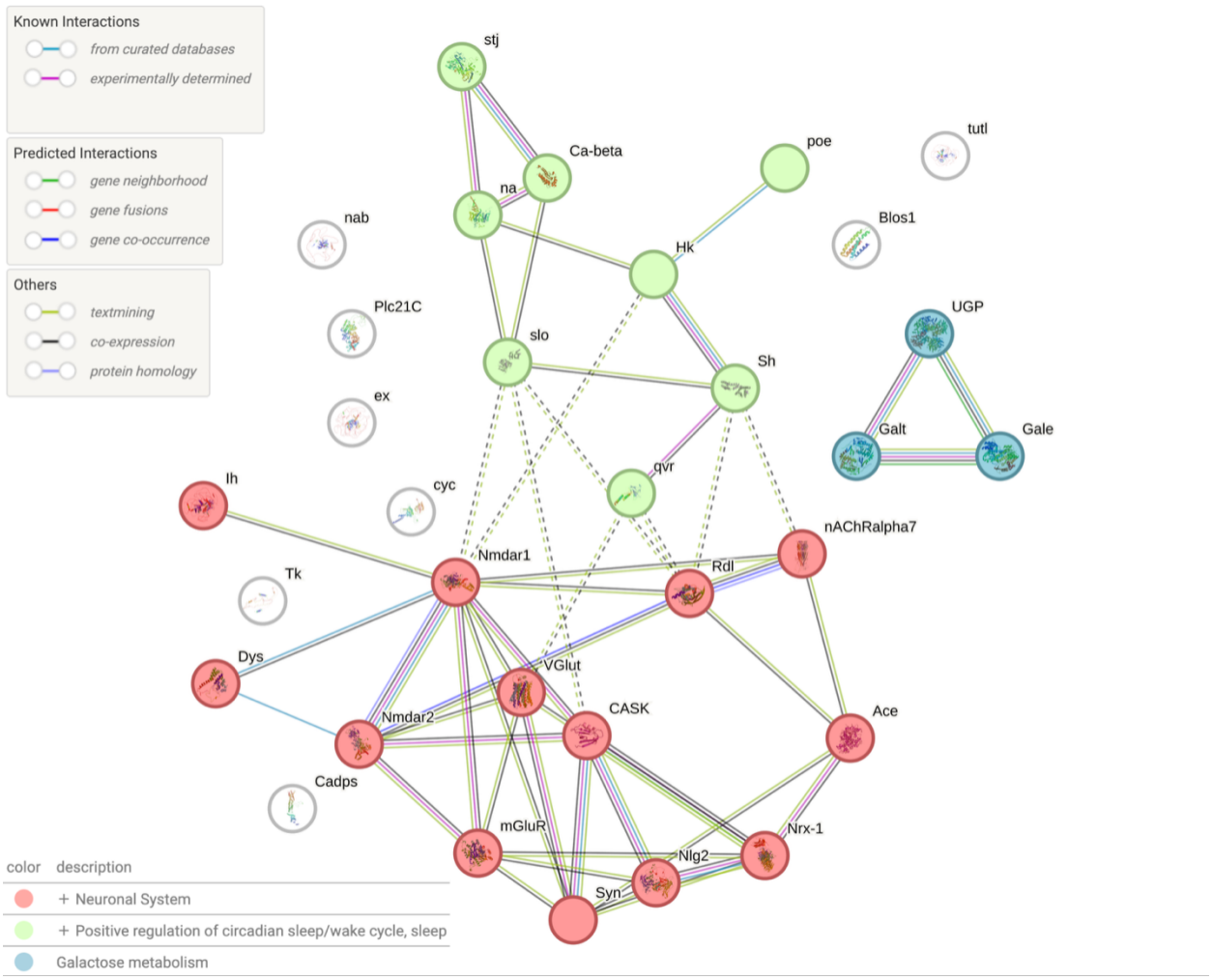


Figure 3.6 Interaction network of Downregulated *Culicoides* genes associated with abnormal behavior.

Gene orthologs associated with abnormal behavior phenotypes (Monarch FBcv:0000387) that were down regulated in response to infection with BTV in *C. sonorensis*. Interaction between genes is shown as edges and colored according to interaction type. Genes were clustered by k-means (k = 3) and edges between clusters are shown as dotted lines.

Enrichment analysis of the downregulated DEGs identified 96 GO terms. Of particular interest, this set included multiple GO terms related to neural activity, including chemical synaptic transmission (33 DEGs, FDR = 3.16e-7), regulation of membrane potential (18 DEGs, FDR = 1.43e-5), modulation of chemical synaptic transmission (22 DEGs, 3.79e-05), and many

others. Additionally, many DEGs related to metabolic pathways were identified, with the KEGG metabolic pathways (53 DEGs, FDR = 0.01) representing many of them. Of these DEGs 36 are associated with amino-acid metabolism, 12 with starch and sucrose metabolism, and five with glycerophospholipid metabolism.

In the downregulated gene set, several different Monarch phenotype clusters were also enriched, including chemical resistant, electrophoretic variant, abnormal behavior, abnormal neurophysiology, abnormal locomotor behavior, fertile, and uncoordinated. Of particular interest, the gene cluster for abnormal behavior (32 DEGs, FDR = 0.0078) showed clustering to three main functional components (Figure 3.6). Cluster 1 represents 13 genes related to neural systems and the modulation of chemical synaptic transmission, cluster 2 contains eight genes related to the positive regulation of circadian rhythm, and voltage-gated cation channels, and cluster 3 has three genes related to galactose metabolism. Nineteen of the DEGs were also associated with the chemical resistant Monarch cluster (FDR = 0.0078) which had the strongest signal of all the phenotype pathways.

The set of DEGs that were upregulated during VSV infection were also tested for term enrichment (Supplementary Table 3.1). Only one biological process, smooth septate junction assembly (3 DEGs, FDR = 0.0433), and three cellular component terms: Smooth septate junction (4 DEGS, FDR = 0.0014), ESCRT III complex (3 DEGs, FDR = 0.0191), and apicolateral plasma membrane (4 DEGs, FDR = 0.0191), were found to be enriched. No terms from GO Molecular Function, KEGG Pathways, Reactome Pathways or Monarch Phenotypes were found. Both the up- and downregulated gene sets from the concordance analysis were too small to perform functional enrichment. Both the VSV downregulated set of genes and the concordant upregulated genes showed no interactions between any of their genes in the String DB. However,

the downregulated set of concordant genes showed three groups of genes with edges, with 1 cluster representing three histones (*His2A, Hist4r*, and *His1*), the second cluster representing two odorant binding proteins (*Obp56d* and *Opb44a*), and the third cluster containing two genes related to carboxylic ester hydrolase activity (*CG13562* and *CG17292*).

Discussion

Our data show that viral infection elicits a complex transcriptional response in *C. sonorensis*. Many of the genes involved were uncharacterized or lacked defined orthologs in well-studied dipterans such as *D. melanogaster* (Nayduch et al., 2019; Scroggs et al., 2023). BTV infection induced a substantially broader response than VSV infection, with over eight times as many genes differentially expressed compared to the same control group. This finding aligns with previous studies examining arboviral infection in *C. sonorensis*. For example, the transcriptional responses of the *C. sonorensis* W8 cell line to VSV infection was previously characterized (Scroggs et al., 2023), and significant transcriptional changes were only observed one hour post-infection, with no significant differences at later time points. In a separate study, infection with EHDV produced a much stronger transcriptional response, with 2,401 unigenes showing a greater than $Log_2(FC) > 1$ in expression (Nayduch et al., 2019). This was substantially greater than changes in expression observed in our results; however, differences in viral species, infection route (oral vs. injection), and experimental design (including the earlier time point in the EHDV study, and infection via oral feeding) may contribute to these discrepancies. A common theme across all three transcriptomic studies is the large proportion of uncharacterized DEGs, many of which lack functional orthologs in other dipterans. *Culicoides* is an understudied lineage of dipteran vectors that diverged from other hematophagous dipterans hundreds of

millions of years ago (Wiegmann et al., 2011), which may contribute to the high proportion of uncharacterized genes and their prevalence among those related to viral infection.

Many of the upregulated DEGs in BTV and VSV infected midges were related to immune responses. The upregulated gene set from BTV showed significant enrichment of immune genes, including pathways like Toll, which was also upregulated in response to EHDV infection (Nayduch et al., 2019). Among the most highly upregulated genes in BTV infection was vacuolar protein sorting associated 36-like (*VPS36*), previously characterized as playing a key role in limiting viral infection in other systems. During porcine epidemic diarrhea virus (PEDV) replication in porcine cells, VPS36 interacts with ORF3 to limit replication (Kaewborisuth et al., 2019), with PEDV replication being significantly repressed in cells overexpressing *VPS36*. Key immune genes such as *GNBP* and *PPO* were also upregulated in BTV infections, as seen with EHDV-2 infections (Nayduch et al. 2019). Additionally, increased expression of genes associated with the ProPO pathway were observed in all three infection treatments (BTV, VSV, and EHDV). However, other innate immune effectors previously identified as upregulated during EHDV-2 infection (Nayduch et al. 2019, including *defensin* and *attacin*) were not able to be analyzed in this study, as they were not present in the annotation generated by eGAPx, and therefore were not available for read mapping. Identification of these genes will require further gene modeling in order to determine their expression status during BTV or VSV infection.

While the small number of DEGs in VSV infection limited the scope of functional enrichment analysis, several pathways were enriched in upregulated genes, including smooth septate junction and apicolateral plasma membrane terms. These pathways suggest cell-to-cell adhesion responses, potentially reflecting an inflammatory immune response. This finding is

consistent with the previous *in vitro* study, Scroggs et al. 2023, using *C. sonorensis* cell lines, which similarly showed little to no differential gene expression at later timepoints post-VSNJV infection. Our viral injection method simulates these late-stage infections, skipping the early midgut barrier response, suggesting that the limited transcriptional response observed here may reflect the establishment of a persistent infection.

The concordant gene set revealed genes with similar expression patterns across both viral infections, suggesting a core antiviral response in *C. sonorensis*. Notably, *glutathione S-transferase 1-like* (*gst1-like*) was significantly differentially expressed in BTV infection and, while not reaching significance in VSV, was identified in the concordant set. This finding aligns with previous research demonstrating the importance of *gst1-like* in vector competence. Specifically, *C. sonorensis* individuals that overexpressed *gst1-like* were refractory to BTV infection (Morales-Hojas et al., 2018). To further clarify the role of *gst1-like* in vector competence, future studies might use RNA interference (RNAi) knockdown or CRISPR-Cas9 knockout methods to test the effects of *gst1-like* silencing on viral replication. The availability of a chromosome-scale genome assembly (Chapter 2) for *C. sonorensis* also allows for analysis like quantitative trait locus (QTL) mapping to reveal if *gst1-like* may operate as part of a gene cluster or regulatory network controlling vector competence, potentially revealing other candidate genes and pathways that underlie the complex phenotype of viral resistance in this key disease vector.

Previous studies identified significant differential expression of vision-associated genes (with the majority being downregulated) in *C. sonorensis* 36 h post-infection via feeding with EHDV-2 (Nayduch et al., 2019). In contrast, neither BTV nor VSV infections in this study, performed via injection, showed enrichment for vision genes in their respective DEG sets or concordant gene sets. Interestingly, however, the gene with the highest upregulation in BTV

infected midges was *mab-21-like*, which is involved in eye development. Additionally, whereas vision-related pathways were not enriched, BTV infection resulted in significant downregulation of genes associated with neurological and behavioral functions, which is also in accordance with findings in EHDV-2 infected midges (Nayduch et al. 2019). These findings suggest that the significant changes in gene expression observed in BTV infection may alter midge behaviors or responses to stimuli compared to uninfected individuals. Previously, midges infected with BTV were less likely to be found in traps using UV light as an attractant than uninfected midges (McDermott et al 2015) suggesting arboviral infection may alter phototaxis. Additionally, changes in mosquito behaviors following infection with arboviruses have been documented, including alteration of feeding and probing behaviors on hosts (Jameson et al. 2024). The changes in transcript levels may provide insight into the molecular mechanisms that underlie these behavioral changes and lead to a better understanding of their impact on attraction to hosts, trapping efficiency, behavior, and how these changes affect the accuracy of risk assessments for livestock or wildlife populations.

Despite limited functional annotation, lncRNAs have emerged as important regulators of immune responses during viral infections, acting as effectors that modulate other immune genes (Ginn et al., 2021; Valanne et al., 2019). More significantly for vector biology, lncRNAs have been directly implicated in regulating vector competence in *Aedes aegypti*, functioning as both pro- and anti-viral replication factors (Belavilas-Trovas et al., 2023). The functional roles of lncRNAs remain poorly characterized and are underrepresented in current databases. We identified numerous lncRNAs that were significantly up- and downregulated following viral infection. Notably, lncRNAs comprised a substantial proportion (21.4%, n = 18) of genes showing concordant expression patterns between VSV and BTV infections, suggesting their

importance in core antiviral responses or effects associated with arbovirus infection in *C. sonorensis*. Searching the differentially expressed lncRNAs from this study against RNAcentral (The RNAcentral Consortium et al., 2019) yielded no significant matches, indicating these sequences have not been previously documented or characterized. Accordingly, these findings highlight lncRNAs as a promising area for future functional characterization as well as potential targets for modulation of vector competence.

The primary limitation of this study was the substantial variability observed between individual midges. Individual transcriptomes were selected because they allowed verification of infection status and quantification of viral load for each midge, which is not possible with pooled samples (Supplementary Table 3.2). While this approach provided high confidence in sample selection, it resulted in considerable inter-individual variability that, particularly in VSV infections, outweighed the effects of viral infection itself, as demonstrated by PCA analysis and r-log clustering (Figs. 1 and 2). Despite these challenges, we obtained biologically relevant insights with n = 8 individuals per treatment. However, future studies may benefit from biological pooling or increased sample sizes to better account for individual variability. Additionally, injections were chosen over oral feeding because infection rates following blood meal feeding were inconsistent, whereas injection provided a reliable infection method with minimal infectious dose variability. Whereas controls and treatments received injections, this approach bypasses the natural midgut barrier and does not recapitulate the natural progression of infection. Although this study provided valuable insights into late-stage effects of viral infection in *C. sonorensis*, future investigations could benefit from methodological modifications. A potentially more effective approach would involve oral infection through blood meals, followed by verification of infection in individuals and subsequent pooling into biological replicates. This

strategy could better balance the need for infection verification with reduced inter-individual variability while maintaining biological relevance.

In summary, we demonstrate that viral infection in *C. sonorensis* elicits a broad and complex transcriptional response. BTV infection triggered a substantially larger and more diverse response than VSV, including the upregulation of well-characterized innate immune pathways as well as virus-specific responses. In addition to immune responses, BTV infection appeared to alter the expression of neural and behavioral genes, suggesting mechanisms through which BTV may cause virus-induced changes in host behavior. The high proportion of uncharacterized genes and lncRNAs differentially expressed after viral infection also highlights the unique biology of *C. sonorensis* and suggests that aspects of its antiviral response may be genus- or species-specific. Further exploration of other *Culicoides* species will be essential in understanding genetic factors influencing vector competence and virus transmission in this important group of arboviral vectors.

References

Alderink, F. J. 1984. Vesicular stomatitis epidemic in Colorado: Clinical observations and financial losses reported by dairymen. Prev. Vet. Med. 3: 29-44.

Alkhamis, M. A., C. Aguilar-Vega, N. M. Fountain-Jones, K. Lin, A. M. Perez, and J. M. Sánchez-Vizcaíno. 2020. Global emergence and evolutionary dynamics of bluetongue virus. Sci. Rep. 10: 21677.

Andrews, S. 2010. FastQC. Computer software.

Belavilas-Trovas, A., S. Tastsoglou, S. Dong, M. Kefi, M. Tavadia, K. D. Mathiopoulos, and G. Dimopoulos. 2023. Long non-coding RNAs regulate *Aedes aegypti* vector competence for Zika virus and reproduction. PLoS Pathog. 19: e1011440.

Bellone, R., P. Lechat, L. Mousson, V. Gilbart, G. Piorkowski, C. Bohers, A. Merits, E. Kornobis, J. Reveillaud, C. Paupy, M. Vazeille, J.-P. Martinet, Y. Madec, X. De Lamballerie, C. Dauga, and A.-B. Failloux. 2023. Climate change and vector-borne diseases: A multi-omics approach of temperature-induced changes in the mosquito. J. Travel Med. 30: taad062.

Blighe, K. 2018. EnhancedVolcano. Computer software.

Chen, S. 2023. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta 2: e107.

Cheng, G., Y. Liu, P. Wang, and X. Xiao. 2016. Mosquito defense strategies against viral infection. Trends Parasitol. 32: 177-186.

Core Team. 2021. R: A language and environment for statistical computing. Computer software.

Di Tommaso, P., M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame. 2017. Nextflow enables reproducible computational workflows. Nat. Biotechnol. 35: 316-319.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. 2013. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29: 15-21.

Drolet, B. S., K. W. Mills, E. L. Belden, and J. O. Mecham. 1990. Enzyme-linked immunosorbent assay for efficient detection of antibody to bluetongue virus in pronghorn (*Antilocapra americana*). J. Wildl. Dis. 26: 34-40.

Drolet, B. S., L. M. Reister, T. D. Rigg, P. Nol, B. K. Podell, J. O. Mecham, K. C. VerCauteren, P. A. Van Rijn, W. C. Wilson, and R. A. Bowen. 2013. Experimental infection of white-tailed deer (*Odocoileus virginianus*) with northern European bluetongue virus serotype 8. Vet. Microbiol. 166: 347-355.

Ewels, P., M. Magnusson, S. Lundin, and M. Käller. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32: 3047-3048.

García-Alcalde, F., K. Okonechnikov, J. Carbonell, L. M. Cruz, S. Götz, S. Tarazona, J. Dopazo, T. F. Meyer, and A. Conesa. 2012. Qualimap: Evaluating next-generation sequencing alignment data. Bioinformatics 28: 2678-2679.

Gaydos, J. K., D. E. Stallknecht, D. Kavanaugh, R. J. Olson, and E. R. Fuchs. 2002. Dynamics of maternal antibodies to hemorrhagic disease viruses (Reoviridae: Orbivirus) in white-tailed deer. J. Wildl. Dis. 38: 253-257.

Gethmann, J., C. Probst, and F. J. Conraths. 2020. Economic impact of a bluetongue serotype 8 epidemic in Germany. Front. Vet. Sci. 7: 65.

Ginn, L., M. La Montagna, Q. Wu, and L. Shi. 2021. Diverse roles of long non-coding RNAs in viral diseases. Rev. Med. Virol. 31: e2198.

Häsler, B., K. S. Howe, E. Di Labio, H. Schwermer, and K. D. C. Stärk. 2012. Economic evaluation of the surveillance and intervention programme for bluetongue virus serotype 8 in Switzerland. Prev. Vet. Med. 103: 93-111.

Hayek, A. M., B. J. McCluskey, G. T. Chavez, and M. D. Salman. 1998. Financial impact of the 1995 outbreak of vesicular stomatitis on 16 beef ranches in Colorado. J. Am. Vet. Med. Assoc. 212: 820-823.

Jameson, S. B., L. D. Vaughan, J. E. de Verges, B. H. Carter, G. L. Dobek, B. Londoño-Renteria, and D. M. Wesson. 2024. Electropenetrography with alternating current reveals *in situ* changes of *Aedes aegypti* probing behaviors associated with Dengue virus infection. Comp. Med. 74: 284-294.

Jones, R. H., R. D. Roughton, N. M. Foster, and B. M. Bando. 1977. *Culicoides*, the vector of epizootic hemorrhagic disease in white-tailed deer in Kentucky in 1971. J. Wildl. Dis. 13: 2-8.

Kaewborisuth, C., Y. Yingchutrakul, S. Roytrakul, and A. Jongkaewwattana. 2019. Porcine epidemic diarrhea virus (PEDV) ORF3 interactome reveals inhibition of virus replication by cellular VPS36 protein. Viruses 11: 382.

Kolde, R. 2019. pheatmap: Pretty Heatmaps. Computer software.

Kurtzer, G. M., V. Sochat, and M. W. Bauer. 2017. Singularity: Scientific containers for mobility of compute. PLoS One 12: e0177459.

Krueger, F., F. James, P. Ewels, E. Afyounian, M. Weinstein, B. Schuster-Boeckler, G. Hulselmans, and Sclamons. 2023. FelixKrueger/TrimGalore: V0.6.10 - add default decompression path. Computer software.

Labadie, T., E. Sullivan, and P. Roy. 2020. Multiple routes of bluetongue virus egress. Microorganisms 8: 965.

Love, M. I., W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15: 550.

McDermott, E. G., C. E. Mayo, A. C. Gerry, D. Laudier, N. J. MacLachlan, and B. A. Mullens. 2015. Bluetongue virus infection creates light averse *Culicoides* vectors and serious errors in transmission risk estimates. Parasit. Vectors 8: 460.

Mills MK, Ruder MG, Nayduch D, et al. 2017. Dynamics of epizootic hemorrhagic disease virus infection within the vector, *Culicoides sonorensis* (Diptera: Ceratopogonidae). Mans BJ, editor. PLOS ONE. 12:11.

Morgan, M., and V. Obenchain. 2017. Summarized Experiment. Computer software.

Morales-Hojas, R., M. Hinsley, I. M. Armean, R. Silk, L. E. Harrup, A. Gonzalez-Uriarte, E. Veronesi, L. Campbell, D. Nayduch, C. Saski, W. J. Tabachnick, P. Kersey, S. Carpenter, and M. Fife. 2018. The genome of the biting midge *Culicoides sonorensis* and gene expression analyses of vector competence for bluetongue virus. BMC Genomics 19: 624.

Mullens, B. A., and E. T. Schmidtmann. 1982. The Gonotrophic Cycle of *Culicoides variipennis* (Diptera: Ceratopogonidae) and its implications in age-grading field populations in New York State, USA. J. Med. Entomol. 19: 340-349.

Nayduch, D., V. Shankar, M. K. Mills, T. Robl, B. S. Drolet, M. G. Ruder, E. D. Scully, and C. A. Saski. 2019. Transcriptome response of female *Culicoides sonorensis* biting midges (Diptera: Ceratopogonidae) to early infection with epizootic hemorrhagic disease virus (EHDV-2). Viruses 11: 473.

Patel, H., J. Manning, P. Ewels, M. U. Garcia, A. Peltzer, R. Hammarén, O. Botvinnik, A. Talbot, G. Sturm, nf-core bot, M. Zepper, D. Moreno, P. Vemuri, M. Binzer-Panchal, E. Greenberg, silviamorins, L. Pantano, R. Syme, G. Kelly, and P. Zhou. 2025. nf-core/rnaseq: Nf-core/rnaseq v3.19.0 - Tungsten Turtle. Computer software.

Patro, R., G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14: 417-419.

Pelzel-McCluskey, A., B. Christensen, J. Humphreys, M. Bertram, R. Keener, R. Ewing, L. W. Cohnstaedt, R. Tell, D. P. C. Peters, and L. Rodriguez. 2021. Review of vesicular stomatitis in the United States with focus on 2019 and 2020 outbreaks. Pathogens 10: 993.

Sweeney, B. A., A. I. Petrov, B. Burkov, R. D. Finn, A. Bateman, M. Szymanski, W. M. Karlowski, J. Gorodkin, S. E. Seemann, J. J. Cannone, R. R. Gutell, P. Fey, S. Basu, S. Kay, G. Cochrane, K. Billis, D. Emmert, S. J. Marygold, and K. P. Williams. 2019.

RNAcentral: A hub of information for non-coding RNA sequences. Nucleic Acids Res. 47: D221-D229.

Rozo-Lopez, P., B. Drolet, and B. Londoño-Renteria. 2018. Vesicular stomatitis virus transmission: A comparison of incriminated vectors. Insects 9: 190.

Rozo-Lopez, P., B. Londono-Renteria, and B. S. Drolet. 2020. Venereal transmission of vesicular stomatitis virus by *Culicoides sonorensis* midges. Pathogens 9: 316.

Rushton, J., F. Macchioni, L. Chelucci, and B. Torracca. 2015. Economic impact of bluetongue: A review of the effects on production. Vet. Ital. 51: 401-406.

Sayols, S., D. Scherzinger, and H. Klein. 2016. dupRadar: A Bioconductor package for the assessment of PCR artifacts in RNA-Seq data. BMC Bioinformatics 17: 428.

Severson, D. W., and S. K. Behura. 2016. Genome investigations of vector competence in *Aedes aegypti* to inform novel arbovirus disease control approaches. Insects 7: 58.

Scroggs, S. L. P., E. J. Bird, D. C. Molik, and D. Nayduch. 2023. Vesicular stomatitis virus elicits early transcriptome response in *Culicoides sonorensis* cells. Viruses 15: 2108.

Soneson, C., M. I. Love, and M. D. Robinson. 2015. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. F1000Res. 4: 1521.

Szklarczyk, D., R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo, P. Bork, L. J. Jensen, and C. von Mering. 2023. The STRING database in 2023: Protein-protein association networks and functional

enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 51: D638-D646.

Tabachnick, W. J. 1996. *Culicoides variipennis* and bluetongue-virus epidemiology in the United States. Annu. Rev. Entomol. 41: 23-43.

Tanya, V. N., E. C. Greiner, and E. P. J. Gibbs. 1992. Evaluation of *Culicoides insignis* (Diptera: Ceratopogonidae) as a vector of bluetongue virus. Vet. Microbiol. 32: 1-14.

Valanne, S., T. S. Salminen, M. Järvelä-Stölting, L. Vesala, and M. Rämet. 2019. Immune-inducible non-coding RNA molecule lincRNA-IBIN connects immunity and metabolism in *Drosophila melanogaster*. PLoS Pathog. 15: e1007504.

Wang, L., S. Wang, and W. Li. 2012. RSeQC: Quality control of RNA-seq experiments. Bioinformatics 28: 2184-2185.

Wiegmann, B. M., M. D. Trautwein, I. S. Winkler, N. B. Barr, J.-W. Kim, C. Lambkin, M. A. Bertone, B. K. Cassel, K. M. Bayless, A. M. Heimberg, B. M. Wheeler, K. J. Peterson, T. Pape, B. J. Sinclair, J. H. Skevington, V. Blagoderov, J. Caravas, S. N. Kutty, U. Schmidt-Ott, and D. K. Yeates. 2011. Episodic radiations in the fly tree of life. Proc. Natl. Acad. Sci. USA 108: 5690-5695.

Chapter 4 - BALROG-MON: A Metagenomic Framework that Facilitates the Use of House Flies as Xenosurveillance Tools for Monitoring Pathogens and Antimicrobial Resistance

Prepared as: Edward Bird, Luke Brendel, Kristopher Silver, Dana Nayduch. BALROG-MON: A Metagenomic Framework that Facilitates the Use of House Flies as Xenosurveillance Tools for Monitoring Pathogens and Antimicrobial Resistance

Abstract

The global rise of antimicrobial resistance (AMR) calls for the development of new surveillance systems capable of tracking pathogen reservoirs and resistance genes across diverse environments. This study presents a xenosurveillance framework that employs house flies (*Musca domestica*) as bioindicators for the detection of environmental AMR and pathogens coupled with BALROG-MON, a tailored metagenomic analysis pipeline. House flies were collected from an urban dumpster and a rural cattle feed facility and long-read metagenome sequencing was performed with Oxford Nanopore technology. Assembly-free analysis of these low-coverage, complex metagenomes was enabled through the application of the BALROG-MON pipeline, which integrated tools for predicting multiple antimicrobial resistance genes (ARG) and mobile genetic elements (MGE). Metagenomic comparisons between locations revealed that flies sample discrete microbial communities resulting in site-specific bacterial signatures. Urban flies had metagenomes with greater ARG diversity, including clinically-relevant resistance genes to carbapenems and beta-lactams, while rural fly metagenomes were enriched with tetracycline resistance genes aligned with widespread use of this antibiotic in cattle disease management. ESKAPE pathogens were present in flies from both environments, and

77% of their associated ARGs were predicted to be plasmid-encoded, suggesting the potential for horizontal ARG transfer. Indicator bacteria were successfully correlated with flies from each environment, with *Pseudomonas fragi* (indicating food spoilage) in the urban flies and *Pantoea agglomerans* (indicating grain association) in rural flies. The combination of taxonomic classification, plasmid prediction, and ARG annotation delivered robust resistome characterization from bacteria carried on house flies. This study validates house flies as effective environmental bioindicators that consolidate microbial communities from diverse local niches into a single convenient sample. This framework delivers a scalable and affordable surveillance alternative that demands little technical capacity that improves upon conventional environmental sampling techniques. Together with the BALROG-MON pipeline's capacity to generate comprehensive information from complex metagenomic data, this system is a valuable contribution to One Health AMR surveillance strategies.

Introduction

For over a century, the use of antibiotics to treat infectious diseases in humans and animals has dramatically reduced morbidity and mortality. However, the widespread overuse and/or misuse of antibiotics in human medicine, livestock production, and agriculture has accelerated the emergence of antimicrobial resistance (AMR) (Akram et al., 2023). With bacterial AMR responsible for 1.27 million global human deaths in 2019 (Murray et al. 2022) and a significant contributor to global animal death rates (One Health Joint Plan of Action, 2022–2026 2022, Jacobsen et al. 2023, Babo Martins et al. 2024), it is now recognized as a critical global threat to human and animal health. Current estimates indicate that antimicrobial resistance will lead to more than 10 million fatalities in 2050 if the rise of AMR continues at current rates (GBD 2021 Antimicrobial Resistance Collaborators 2024). Addressing the

widespread impact of AMR is essential and requires a One Health approach that coordinates efforts across clinical, agricultural, and environmental sectors (Velazquez-Meza et al., 2022). This framework acknowledges that resistance can emerge and circulate beyond clinical settings, underscoring the importance of investigating environmental and agricultural reservoirs of AMR.

Selective pressure from antimicrobial use enables bacteria to readily acquire antimicrobial resistance genes (ARGs), often encoded on mobile genetic elements (MGEs) which can be transferred to other bacteria (Partridge et al., 2018). Therefore, whereas preventing the overuse of antibiotics is important, this strategy alone is not sufficient to halt the emergence of resistance. MGEs are subject to horizontal gene transfer (HGT), particularly via plasmids and integrative conjugative elements, and play a crucial role in accelerating the spread and emergence of resistance. With the current prevalence of resistant bacterial pathogens and ARGs, HGT across microbial taxa has been implicated to be the major driver of the AMR crisis (Collignon et al., 2018).

House flies (*Musca domestica L.*) are synanthropic insects that thrive in both human- and animal-centric environments and have cosmopolitan distribution across the globe. House flies, particularly female flies, are attracted to and seek out protein and microbe-rich substrates, including animal waste, animal feed, and decomposing organic matter. These resources act as a nutrient source for adult and larval houseflies, and as such are a preferred location for oviposition (Nayduch et al., 2023). As adult flies feed and oviposit in these microbe-rich sites, they become burdened with the local microbiota on their body and in their gut. Surveys of flies from farms, hospitals, markets, and urban areas have consistently isolated multidrug-resistant *Escherichia coli*, *Klebsiella sp.*, *Providencia sp.*, *Staphylococcus sp*. and other genera harboring extended-spectrum β-lactamase, tetracycline and colistin resistance genes (Akter et al., 2020; Alam &

Zurek, 2004; Neupane et al., 2020; Pickens et al., 2025; Usui et al., 2013; Yang et al., 2024a). Additionally, the microbial communities of house flies correlate tightly with their local environment, and change depending on the environment they inhabit (Neupane et al., 2023; Neupane & Nayduch, 2022). Because flies freely move between different sources, the microbial community they carry represents a complex assemblage from these local niches.

Metagenomic surveillance has become a powerful tool that allows identification of the full complexity of ARGs in the environment without relying on culture-based methods. This approach leverages affordable, high-throughput sequencing technologies to perform shotgun sequencing of all DNA present in a sample. By avoiding the need to culture bacteria, metagenomics bypasses the selection biases associated with traditional isolation techniques (Danko et al., 2021). Further, metagenomic approaches enable the detection and characterization of ARGs from both culturable and unculturable organisms, while still providing contextual information about their association with both MGEs and potential pathogens. Metagenomic surveillance has been applied to a diverse collection of samples to study the environmental resistome, including wastewater, soil, animal feed, and hospital settings (Pillay et al., 2022). Notably, this method has also been used to characterize the microbiome and resistome of house flies from human-centric locations, revealing that they harbor a wide array of ARGs and human pathogens (Yang et al., 2024a). The increasing affordability of sequencing technologies has facilitated the performance of large-scale metagenomic surveys that provide insights into the complex dynamics of AMR across diverse environments.

In this study, we leveraged the natural filth-seeking biology of house flies as a sampling tool to streamline AMR/ARG and pathogen monitoring. We propose that house flies can serve as an environmental proxy for the microbial community of a sampling site, integrating diverse

localized habitats and resources (e.g., animals, feed, and animal waste or garbage) into a single, easily-collected sample. Instead of sampling each of these niches separately, flies provide a pooled snapshot of the bacterial composition of an entire site, enabling the identification of both ARGs and pathogens present in the environment. To analyze these complex metagenomes, we utilize BALROG-MON (Bird et al., 2025), a novel Nextflow pipeline developed to detect ARGs, MGEs, pathogens, and the correlations between them in high-complexity or low-coverage metagenomic samples. This study had two main goals: first, to demonstrate the utility of house flies as sentinels for pathogens and AMR in the environment; and second, to validate the application of BALROG-MON in extracting relevant insights from complex metagenomic datasets. Together, these aims contribute to a scalable and One Health–oriented surveillance model in which the ubiquitous house fly, combined with next-generation metagenomics, improves the detection of existing, hidden and/or emerging microbial threats.

Methods

Sample Collection

Adult house flies were collected from two distinct sites: an urban dumpster setting adjoining several restaurants (urban site; Manhattan, KS), and a cattle stocker feedyard (rural site; Kansas State University Beef Stocker Unit). Sample collection was conducted using a sweep net disinfected with 10% bleach, with sampling performed by direct sweeps above the dumpster surface and cattle feed bunks. Following sweep netting, flies were immobilized on ice for sex determination. Individual female house flies were preserved in 100 µL of DNA/RNA Shield (Zymo Research, Irvine, CA) and stored at -20°C until extraction.

Sample Extraction & Sequencing

Flies were homogenized using bashing beads (2 mm, Zymo Research) in a MP Biomedicals FastPrep-24 5G for 5 s. The homogenate was centrifuged at 10,000 × g and the supernatant collected for DNA extraction. DNA extraction was performed using the BIOMICS DNA Miniprep (Zymo Research) column kit following the manufacturer's recommended protocol. The extracted DNA was analyzed for purity using a NanoDrop 2000C spectrophotometer (Thermo Fisher Scientific, Waltham, MA) and quantified using a Qubit 3.0 fluorometer (Thermo Fisher Scientific) with the 1x broad range DNA kit (Thermo Fisher Scientific). Long-read sequencing libraries were prepared using the Oxford Nanopore ligation sequencing kit V14 according to manufacturer guidelines. Sequencing was performed on a PromethION R10.4.1 flow cell using an Oxford Nanopore GridION for 48 h. Base calling was conducted using the Dorado r10.4.1 super-accurate model (https://github.com/nanoporetech/dorado).

Bioinformatic Analysis

Data processing was executed using BALROG-MON (Bird et al., 2025) in assembly-free mode with default settings unless otherwise specified. Initial quality control included read trimming and adapter removal using Porechop v0.2.4 (Wick et al., 2017) and Chopper v0.7.0 (De Coster & Rademakers, 2023). Host and human DNA decontamination was performed using Minimap2 v2.26 (H. Li, 2018, 2021) with the NCBI RefSeq house fly genome assembly (GCF_030504385.1) and human genome (GCA_000001405.15_GRCh38). Unmapped reads were extracted using Samtools v1.7.17 (Danecek et al., 2021), with quality metrics visualized through FastQC v0.12.1 (https://github.com/s-andrews/FastQC) and MultiQC v1.22.3 (Ewels et al., 2016). Pathogen detection utilized Kraken2 v2.1.3 (Lu et al., 2022; Wood et al., 2019) with

the precompiled PlusPF Database (https://benlangmead.github.io/aws-indexes/k2), and relative abundance was estimated using Bracken v2.9 (Lu et al., 2017). Seqtk v1.4 (https://github.com/lh3/seqtk) converted reads to FASTA format, with sequence statistics generated via Quast (Gurevich et al., 2013). Sequence origin prediction was performed using Plasmer v23.04.20 (Zhu et al., 2023). Antimicrobial resistance was analyzed using AMRFinder v4.0.19 (Feldgarden et al., 2021), ResFinder v4.4.2 (Bortolaia et al., 2020), and RGI v6.0.3 (McArthur et al., 2013), with results standardized using hAMRonization (Mendes et al., 2024).

Post Processing and Analysis

As a significant proportion of the predicted ARGs identified by BALROG-MON using the MultiAMR module produced non-functional protein sequences, we screened all predicted ARGs for proper open reading frames (ORFs) using a custom python tool, pseudo-hAMR (https://github.com/edwardbirdlab/edwardbirdlab-tools). For each predicted ARG, pseudo-hAMR extracted the coding sequence based on the coordinates and strand information from AMRFinder, ResFinder, or RGI, and assessed whether the sequence contained an in-frame internal stop codon. For sequences predicted from nucleotide homology, the ORF was inferred based on the reading frame that produced the longest uninterrupted sequence without an internal stop. Genes containing an internal stop within the predicted sequence were excluded from downstream analysis. Additionally, the presence or absence of start and stop codons in the expected positions was recorded. Whether a sequence was partial, due to its location at the end of a contig, was also noted. Predicted protein sequences were retained for downstream validation and characterization of intact AMR genes.

Kraken2 results were filtered to include only taxa with reads ≥10, with samples considered positive for a given taxa based on this threshold. Antibiotic resistance gene hits were

filtered to maintain a minimum 80% query coverage. Results from Kraken2, Plasmer, and hAMRonization were integrated using Python for comprehensive sequence analysis. Taxonomic abundance data generated by Bracken results were imported into R V4.2 to perform a non-metric multidimensional scaling (NMDS) ordination and a permutational multivariate analysis of variance (PERMANOVA) using dplyr, tidyverse, and vegan packages (Oksanen et al., 2025; Wickham et al., 2025). Bray-Curtis dissimilarity metric was generated with the vegdist function, and NMDS ordination was observed (k = 2) using megaMDS with 100 random starts and the monoMDS engine. PERMANOVA was performed using the adonis2 function to test if the groups differed significantly using the dissimilarity matrix.

Results

Read QC and Host Decontamination

Sequencing yielded an average of 1.4–3.8 million reads per sample, with a mean quality score of Q36 after base calling. Read lengths varied by sample, ranging from 2,499 to 4,999 bp. During trimming and adapter removal, 5.09%–28.00% of reads were discarded due to low sequence quality. Subsequent removal of house fly host sequences resulted in 96.72%–99.28% of reads being filtered out (map quality > 0). Further filtering for human contamination removed an additional 0.80%–1.24% of the remaining reads. After quality control, 8,806–36,519 reads per sample remained, with an average Q score of 39–46 and a final read length ranging from 1,499 to 2,499 bp. A summary of these results is provided in Supplementary Table 4.1.

Community Analysis

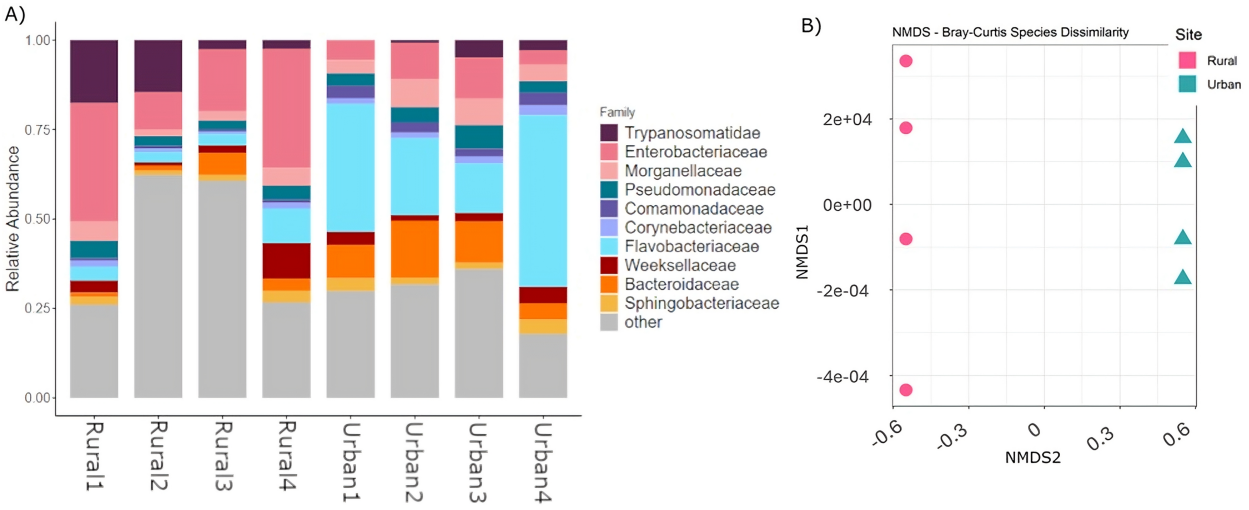


Figure 4.1 Microbial community composition of fly metagenomes.
Microbial communities of house flies from rural and urban environments are shown. A) Relative abundances of the top 10 microbial families identified by Kraken2 and Braken. B) Non-metric Multi-dimensional Scaling (NMDS) ordination of microbial communities using Bray-Curtis species dissimilarity to account for presence/absence and abundance.

Bacterial community composition differed between rural flies and urban flies (Figure 4.1A), highlighting a higher bacterial diversity in rural flies compared to urban flies. Urban fly samples were characterized by a high relative abundance of Flavobacteriaceae, whereas Enterobacteriaceae were prominent in rural fly samples. Additionally, clustering of samples by collection site was observed in the NMDS ordination. The observed species dissimilarity was statistically significant (PERMANOVA, $p = 0.05$), demonstrating that observed bacterial populations formed two distinct groups corresponding to their respective environments.

Pathogen Detection

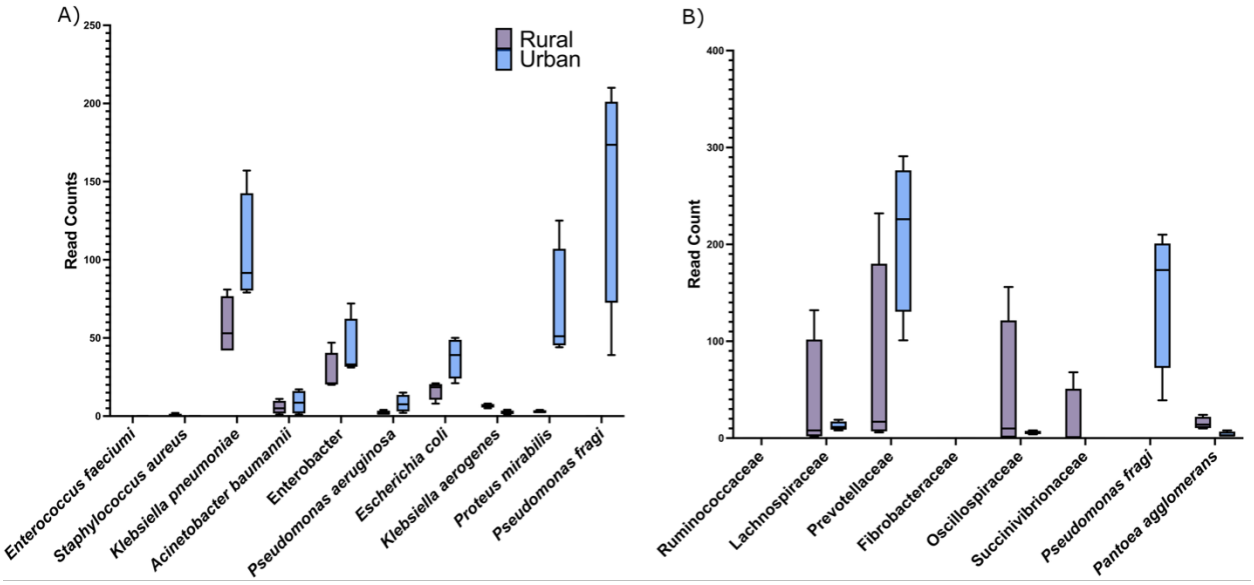


Figure 4.2 Read counts of pathogenic and indicator organisms present in urban and rural house fly metagenomes.

Read-based identification of bacterial taxa in rural and urban house fly metagenomes are shown as read counts for each sample type (n = 4 fly pools). A) Read counts of ESKAPE pathogens and other selected pathogenic species. B) Read counts of indicator organisms.

Kraken2 assigned taxa to an average of 39.35% of reads in rural samples (range: 3,759–12,454 reads) and 62.67% in urban samples (range: 4,767–25,495 reads). Additionally, an average of 13.21% of rural reads and 42.94% of urban reads were classified under the taxon Bacteria. To investigate pathogen presence, we analyzed reads assigned to ESKAPE pathogens (*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa*, and *Enterobacter* species) (Rice, 2008) as well as other pathogenic species such as *Escherichia coli*, *Klebsiella aerogenes*, and *Proteus mirabilis*. Figure 4.2 illustrates the number of reads assigned to each pathogen in rural and urban samples. Read counts varied widely between taxa, with *K. pneumoniae* and *Enterobacter sp*. being the most frequently detected taxa across samples. *A. baumannii* and *P. aeruginosa* had comparatively

fewer assigned reads, while *E. faecium* and *S. aureus* had the lowest read counts. Based on the previously defined threshold, zero rural and zero urban samples tested positive for *E. faeciumi*, zero rural and zero urban for *S. aureus*, four rural and four urban for *K. pneumoniae*, one rural and two urban for *A. baumannii*, zero rural and one urban for *P. aeruginosa*, and four rural and four urban for *Enterobacter sp*. Outside of ESKAPE taxa, many other pathogenic bacterial taxa were identified, with a select few including *E. coli* being positive in three rural and all urban samples, and *P. mirabilis* being detected in zero rural and four urban samples. *K. aerogenes* was detected at a low level in all the rural samples but did not pass the threshold to be considered positive. A summary of these results is provided in Supplementary Table 4.2.

Indicator Organisms

Several bacterial taxonomic groups were analyzed to assess whether specific organisms could serve as indicators of a sample's environmental origin. A sample was considered positive for a given bacterial taxon if at least 10 reads were assigned to that taxon. Read counts for various taxa at different taxonomic levels are shown in Figure 4.2B. Flies collected from the rural site carried indicator taxa assigned to bacterial families commonly associated with rumen digestion, including *Ruminococcaceae*, *Lachnospiraceae*, *Prevotellaceae*, *Fibrobacteraceae*, *Oscillospiraceae*, and *Succinivibrionaceae*. The other tested enteric families, *Ruminococcaceae* and *Fibrobacteraceae* were not detected in any samples. *Lachnospiraceae* were detected in three urban and two rural samples, while *Prevotellaceae* were detected in four urban and three rural samples. *Oscillospiraceae* and *Succinivibrionaceae* were observed only in rural samples, with *Oscillospiraceae* positive in two and *Succinivibrionaceae* positive in only one sample. Additionally, *Pantoea agglomerans*, an opportunistic pathogen linked to grain dust exposure, was identified in all four rural samples and no urban samples, with an average of 15.5 reads per

rural sample. In urban flies, *Pseudomonas fragi*, a potentially pathogenic bacterium associated with milk spoilage, was identified in all four samples at high relative abundance with an average of 120 reads per sample.

Sequence origin

Sequence classification by Plasmer into plasmid, chromosomal, and unclassified categories revealed a consistent pattern across both rural and urban samples. On average, 20.59% of sequences in rural samples and 21.87% in urban samples were classified as plasmid-derived. Chromosomal sequences were less prevalent, comprising 4.28% of rural and 7.10% of urban sequences per sample. Unclassified sequences dominated all samples, with 71.03% in urban and 75.13% in rural samples remaining unclassified by Plasmer. The "too short" category, representing sequences shorter than 500 bp and therefore unsuitable for reliable classification, accounted for a negligible fraction, with a maximum of 0.21% being below 500 bp in sample rural 3. A detailed summary of these results is provided in Supplementary Table 4.3.

ARG database hits & localization

ARG analysis yielded 269 ARG predictions, representing 74 unique ARGs. Of these, 133 ARGs were identified as pseudogenes and removed from downstream analysis. All pseudogenes were detected by ResFinder only. Certain ARGs were more prone to pseudogenization than others: 48 (100%) of *tet(X)* genes, 42 (100%) of *blaMUS-1* genes, 12 (60%) of *tet(M)* genes, and nine (75%) of *tet(Q)* genes were affected. Additionally, some taxa contained more pseudogenes than others, with 90 (84%) of ARGs from *Myroides* species identified as pseudogenes. Several enteric species (e.g., *Proteus mirabilis, Companilactobacillus zhachilii, Bacteroides thetaiotaomicron, Companilactobacillus paralimentarius, Lactobacillus johnsonii*) were overrepresented in the pseudogene set. Two drug classes, folate pathway antagonists and

amphenicols, were completely removed from the analysis due to pseudogenization. After pseudogene removal, CARD/RGI identified an average of 2.5 unique ARGs per rural sample (range: 1–5) and 9.5 unique ARGs per urban sample (range: 5–15). ResFinder identified fewer unique ARGs, with 0 ARGs in rural samples and two total ARGs in urban samples (one each in urban samples 1 and 4). AMRFinderPlus produced slightly more ARG counts than CARD/RGI, identifying an average of 2.5 unique ARGs in rural samples (range: 0–5) and an average of 10 unique ARGs in urban samples (range: 6–15). Because each tool captures a slightly different set of ARG classifications, results were standardized using hAMRonization to integrate predictions across databases. After this step, the number of detected ARGs increased substantially, with rural samples averaging 4.5 unique ARGs per sample (range: 1–9) and urban samples averaging 19.5 unique ARGs per sample (range: 12–30). Harmonized ARG sequences were cross-referenced with Plasmer's classification to determine the predicted genomic origin of each ARG. Plasmid-classified sequences represented the most prevalent source: 41.96% of rural ARGs (range: 0–8 ARGs per sample) and 53.33% of urban ARGs (range: 8–22) originated from plasmids. Chromosomal ARGs were the next most common, accounting for 45.16% in rural samples (range: 1–11) and 25.71% in urban samples (range: 4–11). A smaller proportion of ARGs originated from unclassified sequences, representing 12.9% in rural samples (range: 2–2) and 20.95% in urban samples (range: 1–10). A detailed summary of these results is provided in Supplementary Table 4.4.

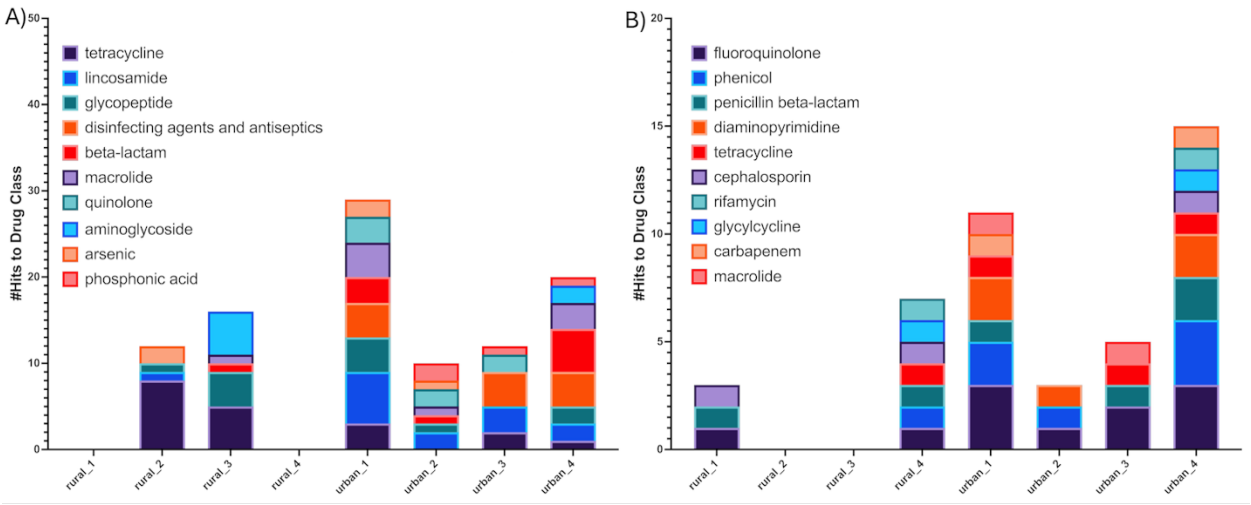


Figure 4.3 Distribution of drug classes from single and multi-drug ARGs present in house fly metagenomes.

Prevalence of drug classes of single- and multi-drug class ARGs in rural and urban house fly metagenomes are shown, highlighting the top 10 most prevalent drug classes associated with the ARGs identified across flies from both environments. A) Single-class ARGs and B) Multi-class ARGs, and their distribution across each sample type (rural or urban fly pools; n = 4 per pool).

To account for the variable number of drug classes associated with an ARG, categories were based on their association with either a single or multiple drug classes (Supplementary Table 4.5 & 4.6). Single-class ARGs confer resistance to only one antibiotic class, while multi-class ARGs are linked to two or more classes. This distinction prevents ARGs with broad resistance profiles from disproportionately influencing overall prevalence patterns. ARGs associated with disinfectants and antiseptics were observed to be more prevalent in urban flies, whereas tetracycline resistance was predominant in rural fly samples. Two of the rural samples (rural 1 and rural 4) contained no ARGs associated with only a single drug class (Figure 4.3A). Analysis of multi-class ARGs revealed a more balanced distribution of ARGs across the top 10 classes, with fluoroquinolone resistance being the most prevalent across both sample types. The

same rural samples with no single resistance genes contained multi-class ARGs; however, rural 2 and rural 3 had no multi-class ARGs. Interestingly, ARGs conferring resistance to carbapenems were detected in three urban fly samples, and urban flies carried a greater variety and quantity of ARGs than rural samples (Figure 4.3B).

Pathogen Predicted ARGs

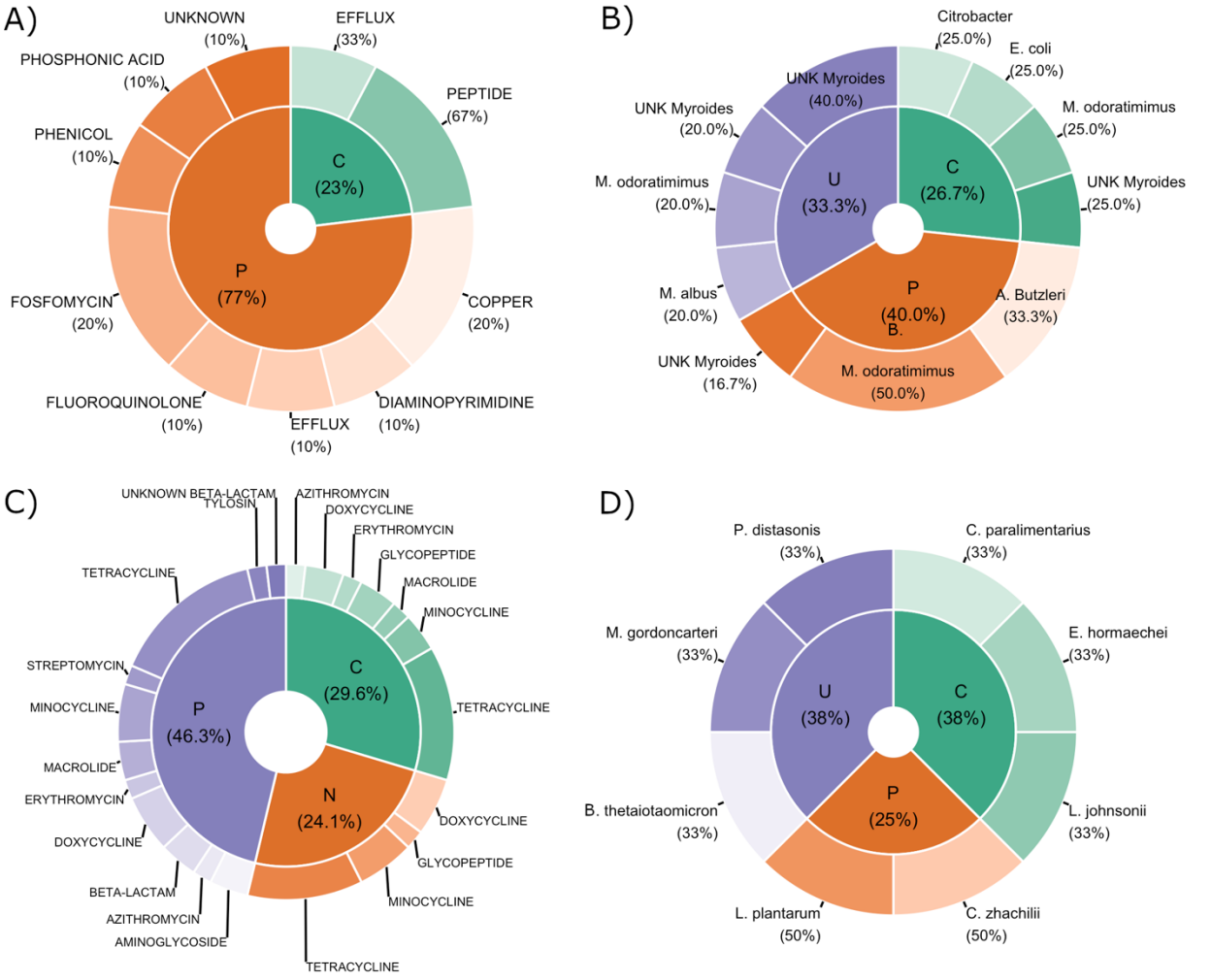


Figure 4.4 Distribution of ARGs and pathogenic taxa identified in house fly metagenomes in relation to element type

Putative origins of ARGs in urban and rural house fly metagenomes are shown. Donut plots show the percentage of ARGs classified as Chromosomal (C), Plasmid (P), or Unknown (U) in origin on the inner pie chart. Predicted taxonomic origin of ARGs (Kraken2) are displayed on the

outer ring in panels B&D, with the predicted antibiotic class of ARGs displayed on the outer ring of panels A&C. A) Resistance profile of ESKAPE pathogens from urban house fly samples. B) Taxonomic distribution of B-lactamase-associated ARGs in urban house fly samples. C) Distribution of ARG classes from enteric microbiota from rural house fly samples. D) Taxonomic distribution of tetracycline-associated ARGs from rural house fly samples.

Results from BALROG-MON, including pathogen identification, plasmid classification, and ARG annotation, were integrated to enable an in-depth analysis of the microbial resistome in fly samples. In urban flies, several ESKAPE pathogens were identified, and many ARGs were predicted to originate from these taxa. Among ARGs associated with ESKAPE pathogens, 77% were predicted to be plasmid-encoded, while 23% were chromosomally encoded (Figure 4.4A). A diverse range of ARGs was observed, including genes conferring resistance to fluoroquinolones and phenicols. Notably, urban flies harbored a high proportion of β-lactamase resistance genes, which were distributed across plasmid (40%), chromosomal (26.7%), and unclassified (33.3%) sequence categories in relatively equal proportions (Figure 4.4B). *Myroides* species contributed substantially to β-lactamase resistance, with hits present in all three categories. In rural flies, a broad representation of enteric microbes was observed, many of which were predicted to carry diverse ARGs (Figure 4.4C). Plasmids represented the largest class of ARG carriers (46.3%), followed by chromosomal (29.6%) and unclassified sequences (24.1%). Tetracycline resistance genes were the most frequently identified, including derivatives such as minocycline and doxycycline, and were present across plasmids/chromosome. Overall, rural flies carried ARGs with substantial tetracycline resistance (Figure 4.4D), distributed among plasmid (25%), chromosomal (38%), and unclassified (38%) origins.

Discussion

This study demonstrates that house flies effectively sample their environment and serve as bioindicators of local microbial communities, including pathogenic bacteria and ARGs.

Through their natural behavior, flies seek out microbe-rich substrates including waste, feed, and decaying organic matter, where they encounter diverse bacterial populations and accumulate a representative snapshot of the local microbial landscape. Our metagenomic analyses revealed clear site-specific differences in microbial community composition between urban and rural environments, with specific taxa serving as indicators of each setting. Urban flies, collected from a dumpster, showed a high abundance of *Pseudomonas fragi*, an opportunistic foodborne pathogen associated with dairy spoilage, indicating recent contact of flies with decomposing food waste. Rural flies collected near a cattle stocker unit harbored bacterial taxa known to be associated with manure, animal gut microbiota, and grain dust, reflecting contact with cattle-associated environments. Notably, *Pantoea agglomerans* was the only indicator organism consistently detected across all rural fly samples, suggesting a strong association between the flies and the cattle feed at this operation. These distinct microbial markers demonstrate that flies capture hallmarks of their local environments, making them valuable tools for monitoring environmental pathogen reservoirs and understanding microbial ecology at the human-animal-environment interface.

All flies in this study carried ARGs associated with both pathogenic and non-pathogenic bacteria and originating from both mobile and chromosomal elements. However, we observed distinct patterns of ARG incidence and prevalence between urban and rural environments that reflect location differences in ARG reservoirs or sources. Urban flies harbored both higher relative quantities and a greater diversity of ARGs, including genes associated with clinically important antibiotic classes such as beta-lactams and carbapenems. Rural flies demonstrated lower overall ARG abundance but were enriched for tetracycline resistance genes, consistent with the historically heavy use of tetracycline-class antibiotics in agriculture (Granados-

Chinchilla & Rodríguez, 2017). The detection of ARGs on both plasmid- and chromosome-associated sequences underscores the potential for these resistance determinants to persist and disseminate within microbial communities. Additionally, several ARGs were linked to ESKAPE pathogens, including those predicted to be plasmid-borne, highlighting the risk for the spread of AMR and the utility of flies for antimicrobial resistance surveillance (Miller & Arias, 2024).

House fly-based surveillance offers several advantages over traditional environmental collection methods. House flies are abundant, globally ubiquitous, and easily collected with minimal cost or training, making this approach accessible and scalable. Their biology drives them to seek out diverse microbe-rich substrates that help target sites with AMR and pathogenic bacteria. When performing AMR surveillance, the prevailing approach has been to focus on a subset of sample types deemed most relevant (Berendonk et al., 2015; Huijbers et al., 2019), given the impracticality of comprehensively sampling the entire environment. While this is a practical and targeted strategy, house flies offer a unique opportunity to establish a standardized sampling method that may serve as a proxy for broader environmental surveillance. As flies naturally traverse these habitats, they provide a single sampling point that captures broad environmental signals without extensive sample types. However, this approach also has important limitations. Individual variations in fly behavior, including choice of food sources, may lead to differences in microbial threat detection. Additionally, while indicator organisms can provide insight into possible environmental sources, the specific origin of detected ARGs or pathogens cannot be definitively identified from fly metagenomes alone. Sequencing challenges in this study included high host DNA contamination, which resulted in low microbial sequence representation and limited metagenomic coverage. This particularly affects detection of threats in environments where pathogens and ARGs may be less abundant, like the rural samples in this

study. Future improvements through differential lysis protocols or adaptive sequencing strategies could help mitigate these limitations.

The effectiveness of the BALROG-MON pipeline was demonstrated through the analysis of these complex metagenomic samples despite challenges from limited metagenomic coverage due to host contamination. The pipeline's modular design enabled data analysis with no signal lost to assembly. Additionally, the integration of multiple ARG prediction tools improved the resistome resolution. Critically, BALROG-MON facilitated association of most of the identified ARGs with potential bacterial hosts and determined their genomic context (plasmid vs. chromosomal), providing comprehensive AMR characterization at each sampling site. Long-read sequencing proved essential for BALROG-MON's effectiveness, enabling direct linkage of ARGs, taxonomic identity, and mobility elements without computationally expensive assembly and binning. This is particularly important in low-coverage metagenomes where short reads often produce fragmented contigs. Additionally, our post-analysis validation scripts discriminated between functional genes and pseudogenes, addressing a critical limitation of mapping-based approaches that can falsely annotate pseudogenes as functional ARGs. The relatively high proportion of non-functional ARGs detected in this study highlights the importance of capturing complete ARG sequences, rather than relying on mapping-based strategies commonly used in metagenomic studies (Langelier et al., 2024; B. Li et al., 2015; Osorio et al., 2023), including the only previous work using house flies for AMR surveillance (Yang et al., 2024b).

Several important areas warrant further investigation to fully illuminate the potential of fly-based surveillance. Expanding sampling to include more flies across broader geographic ranges would provide insight into individual- and population-level variation while enabling

geospatial correlation of fly metagenomes. Although flies can travel up to 20 miles under the right conditions, mark-release-recapture studies suggest they often remain localized (within 1-2 miles of their hatching location) when resources are sufficient, making spatial analysis crucial for optimizing collection strategies (Nayduch et al., 2023). Direct comparison of fly-derived metagenomes to environmental samples (soil, water, manure, surface swabs) from identical locations would help evaluate capture efficiency and identify potentially underrepresented environmental niches. Notably, bacterial communities characterized from flies collected at dairy farms represented the entire microbiome from co-located manure samples along with other taxa acquired from animals, demonstrating that flies are effective sampling tools for these various sources (Neupane et al. 2024). In addition to house flies, exploring other synanthropic insects such as blowflies could further expand surveillance capabilities to include different niches visited by these flies, such as animals or other decaying materials. Technical improvements to BALROG-MON to also include taxon-specific ARG prediction models would improve sensitivity and specificity, quantitative ARG abundance estimation, and integration of metatranscriptomic sequencing to assess active gene expression and identify viable pathogens with active resistance genes carried within flies.

This study establishes house flies as informative bioindicators of environmental AMR and microbial pathogens. By combining long-read metagenomics with our BALROG-MON pipeline, we successfully associated ARGs with specific taxa and determined genomic origins even in challenging, low-coverage samples. This approach offers a scalable, cost-effective AMR surveillance solution requiring minimal technical expertise for sample collection across diverse environments. As global AMR threats escalate, developing enhanced strategies for environmental resistance detection and monitoring becomes increasingly critical. Continued

refinement of both sampling approaches and bioinformatic tools will be essential to fully utilize the potential of fly-based AMR monitoring within the One Health frameworks.

References

Akram, F., M. Imtiaz, and I. U. Haq. 2023. Emergent crisis of antibiotic resistance: A silent pandemic threat to 21st century. Microb. Pathog. 174: 105923.

Akter, S., A. A. M. Sabuj, Z. F. Haque, M. A. Kafi, M. T. Rahman, and S. Saha. 2020. Detection of antibiotic-resistant bacteria and their resistance genes from houseflies. Vet. World 13: 266-274.

Alam, M. J., and L. Zurek. 2004. Association of *Escherichia coli* O157:H7 with houseflies on a cattle farm. Appl. Environ. Microbiol. 70: 7578-7580.

Babo Martins S, Sucena Afonso J, Fastl C, et al. 2024. The burden of antimicrobial resistance in livestock: A framework to estimate its impact within the Global Burden of Animal Diseases programme. One Health. 19:100917.

Berendonk, T. U., C. M. Manaia, C. Merlin, D. Fatta-Kassinos, E. Cytryn, F. Walsh, H. Bürgmann, H. Sørum, M. Norström, M.-N. Pons, N. Kreuzinger, P. Huovinen, S. Stefani, T. Schwartz, V. Kisand, F. Baquero, and J. L. Martinez. 2015. Tackling antibiotic resistance: The environmental framework. Nat. Rev. Microbiol. 13: 310-317.

Bird, E., V. Pickens, D. Molik, K. Silver, and D. Nayduch. 2025. BALROG-MON: a high-throughput pipeline for Bacterial AntimicrobiaL Resistance annOtation of Genomes-Metagenomic Oxford Nanopore. Open Access.

Bortolaia, V., R. S. Kaas, E. Ruppe, M. C. Roberts, S. Schwarz, V. Cattoir, A. Philippon, R. L. Allesoe, A. R. Rebelo, A. F. Florensa, L. Fagelhauer, T. Chakraborty, B. Neumann, G. Werner, J. K. Bender, K. Stingl, M. Nguyen, J. Coppens, B. B. Xavier, and F. M. Aarestrup. 2020. ResFinder 4.0 for predictions of phenotypes from genotypes. J. Antimicrob. Chemother. 75: 3491-3500.

Collignon, P., J. J. Beggs, T. R. Walsh, S. Gandra, and R. Laxminarayan. 2018. Anthropological and socioeconomic factors contributing to global antimicrobial resistance: A univariate and multivariable analysis. Lancet Planet. Health 2: e398-e405.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. 2021. Twelve years of SAMtools and BCFtools. GigaScience 10: giab008.

Danko, D., D. Bezdan, E. E. Afshin, S. Ahsanuddin, C. Bhattacharya, D. J. Butler, K. R. Chng, D. Donnellan, J. Hecht, K. Jackson, K. Kuchin, M. Karasikov, A. Lyons, L. Mak, D. Meleshko, H. Mustafa, B. Mutai, R. Y. Neches, A. Ng, and S. Zubenko. 2021. A global metagenomic map of urban microbiomes and antimicrobial resistance. Cell 184: 3376-3393.

De Coster, W., and R. Rademakers. 2023. NanoPack2: Population-scale evaluation of long-read sequencing data. Bioinformatics 39: btad311.

Ewels, P., M. Magnusson, S. Lundin, and M. Käller. 2016. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32: 3047-3048.

Feldgarden, M., V. Brover, N. Gonzalez-Escalona, J. G. Frye, J. Haendiges, D. H. Haft, M. Hoffmann, J. B. Pettengill, A. B. Prasad, G. E. Tillman, G. H. Tyson, and W. Klimke. 2021. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci. Rep. 11: 12728.

Granados-Chinchilla, F., and C. Rodríguez. 2017. Tetracyclines in food and feedingstuffs: From regulation to analytical methods, bacterial resistance, and environmental and health implications. J. Anal. Methods Chem. 2017: 1-24.

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler. 2013. QUAST: Quality assessment tool for genome assemblies. Bioinformatics 29: 1072-1075.

Huijbers, P. M. C., C.-F. Flach, and D. G. J. Larsson. 2019. A conceptual framework for the environmental surveillance of antibiotics and antibiotic resistance. Environ. Int. 130: 104880.

Jacobsen ABJE, Ogden J, Ekiri AB. 2023. Antimicrobial resistance interventions in the animal sector: scoping review. Front. Antibiot. 2:1233698.

Langelier, C., D. Lu, K. Kalantar, V. Chu, A. Glascock, E. Guerrero, N. Bernick, X. Butcher, K. Ewing, E. Fahsbender, O. Holmes, E. Hoops, A. Jones, R. Lim, S. McCanny, L. Reynoso, K. Rosario, J. Tang, O. Valenzuela, and A. McArthur. 2024. Simultaneous detection of pathogens and antimicrobial resistance genes with the open source, cloud-based, CZ ID pipeline. In Review.

Li, B., Y. Yang, L. Ma, F. Ju, F. Guo, J. M. Tiedje, and T. Zhang. 2015. Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. ISME J. 9: 2490-2502.

Li, H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094-3100.

Li, H. 2021. New strategies to improve minimap2 alignment accuracy. Bioinformatics 37: 4572-4574.

Lu, J., F. P. Breitwieser, P. Thielen, and S. L. Salzberg. 2017. Bracken: Estimating species abundance in metagenomics data. PeerJ Comput. Sci. 3: e104.

Lu, J., N. Rincon, D. E. Wood, F. P. Breitwieser, C. Pockrandt, B. Langmead, S. L. Salzberg, and M. Steinegger. 2022. Metagenome analysis using the Kraken software suite. Nat. Protoc. 17: 2815-2839.

McArthur, A. G., N. Waglechner, F. Nizam, A. Yan, M. A. Azad, A. J. Baylay, K. Bhullar, M. J. Canova, G. De Pascale, L. Ejim, L. Kalan, A. M. King, K. Koteva, M. Morar, M. R. Mulvey, J. S. O'Brien, A. C. Pawlowski, L. J. V. Piddock, P. Spanogiannopoulos, and G. D. Wright. 2013. The comprehensive antibiotic resistance database. Antimicrob. Agents Chemother. 57: 3348-3357.

Mendes, I., E. Griffiths, A. Manuele, D. Fornika, S. H. Tausch, T. Le-Viet, J. Phelan, C. J. Meehan, A. R. Raphenya, B. Alcock, E. Culp, F. Lorenzo, M. S. Haim, A. Witney, A. Black, L. Katz, P. Oluniyi, I. Olawoye, R. Timme, and F. Maguire. 2024.

hAMRonization: Enhancing antimicrobial resistance prediction using the PHA4GE AMR detection specification and tooling. Bioinformatics.

Miller, W. R., and C. A. Arias. 2024. ESKAPE pathogens: Antimicrobial resistance, epidemiology, clinical impact and therapeutics. Nat. Rev. Microbiol. 22: 598-616.

Nayduch, D., S. Neupane, V. Pickens, T. Purvis, and C. Olds. 2023. House flies are underappreciated yet important reservoirs and vectors of microbial threats to animal and human health. Microorganisms 11: 583.

Neupane, S., and D. Nayduch. 2022. Effects of habitat and sampling time on bacterial community composition and diversity in the gut of the female house fly, *Musca domestica* Linnaeus (Diptera: Muscidae). Med. Vet. Entomol. 36: 435-443.

Neupane, S., J. L. Talley, D. B. Taylor, and D. Nayduch. 2023. Bacterial communities and prevalence of antibiotic resistance genes carried within house flies (Diptera: Muscidae) associated with beef and dairy cattle farms. J. Med. Entomol. 60: 1388-1397.

Neupane S, Park Y, Watson DW, et al. 2024. Bacterial communities of house flies from dairy farms highlight their role as reservoirs, disseminators, and sentinels of microbial threats to human and animal health. Insects. 15(9):730.

Neupane, S., K. White, J. L. Thomson, L. Zurek, and D. Nayduch. 2020. Environmental and sex effects on bacterial carriage by adult house flies (*Musca domestica* L.). Insects 11: 401.

Oksanen, J., G. Simpson, G. Blanchet, R. Kindt, P. Legendre, P. Minchin, R. B. O'Hara, P. Solymos, M. H. H. Stevens, E. Szoecs, and H. Wagner. 2025. vegan: Community ecology package. Computer software.

One Health Joint Plan of Action, 2022–2026. 2022. FAO; UNEP; WHO; World Organisation for Animal Health (WOAH) (founded as OIE)

Osorio, V., A. Sabater I Mezquita, and J. L. Balcázar. 2023. Comparative metagenomics reveals poultry and swine farming are hotspots for multidrug and tetracycline resistance. Environ. Pollut. 322: 121239.

Partridge, S. R., S. M. Kwong, N. Firth, and S. O. Jensen. 2018. Mobile genetic elements associated with antimicrobial resistance. Clin. Microbiol. Rev. 31: e00088-17.

Pickens, V., B. Hall, K. Yeater, T. Purvis, E. Bird, G. Brooke, C. Olds, and D. Nayduch. 2025. Bacterial abundance and antimicrobial resistance prevalence carried by adult house flies (Diptera: Muscidae) at Kansas dairy and beef cattle operations. J. Med. Entomol. tjaf052.

Pillay, S., D. Calderón-Franco, A. Urhan, and T. Abeel. 2022. Metagenomic-based surveillance systems for antibiotic resistance in non-clinical settings. Front. Microbiol. 13: 1066995.

Rice, L. B. 2008. Federal funding for the study of antimicrobial resistance in nosocomial pathogens: No ESKAPE. J. Infect. Dis. 197: 1079-1081.

Usui, M., T. Iwasa, A. Fukuda, T. Sato, T. Okubo, and Y. Tamura. 2013. The role of flies in spreading the extended-spectrum β-lactamase gene from cattle. Microb. Drug Resist. 19: 415-420.

Velazquez-Meza, M. E., M. Galarde-López, B. Carrillo-Quiróz, and C. M. Alpuche-Aranda. 2022. Antimicrobial resistance: One Health approach. Vet. World 743-749.

Wick, R. R., L. M. Judd, C. L. Gorrie, and K. E. Holt. 2017. Completing bacterial genome assemblies with multiplex MinION sequencing. Microb. Genomics 3.

Wickham, H., R. François, L. Henry, K. Müller, and D. Vaughan. 2025. dplyr: A grammar of data manipulation. Computer software.

Wood, D. E., J. Lu, and B. Langmead. 2019. Improved metagenomic analysis with Kraken 2. Genome Biol. 20: 257.

Yang, Y., P. Xu, W. He, and F. Tao. 2024a. Metagenomic analysis reveals houseflies as indicators for monitoring environmental antibiotic resistance genes. Environ. Microbiol. Rep. 16: e70032.

Zhu, Q., S. Gao, B. Xiao, Z. He, and S. Hu. 2023. Plasmer: An accurate and sensitive bacterial plasmid prediction tool based on machine learning of shared k-mers and genomic features. Microbiol. Spectr. 11: e04645-22.

Chapter 5 - Conclusion

Vector-borne disease (VBD) is a significant threat to animal health and agriculture, accounting for 17% of human infections. Beyond the contemporary aspect of direct infection from vectors, insects also play complex roles in the scope of human and animal health, spreading and promoting the proliferation of antimicrobial resistant (AMR) bacteria. In this dissertation we focused on advancing omics and bioinformatic resources in two understudied systems, *Culicoides* biting midges and house flies, *Musca domestica*. This work provides the basis for future research and the advancement of VBD research.

Chapter 2 lays the foundation for future omics research in *Culicoides sonorensis,* presenting a high-quality, chromosome-scale genome assembly. We utilized PacBio HiFi long-read, and high-throughput chromosome conformation capture sequencing to produce a telomere-to-telomere genome from a single female *C. sonorensis*. The genome models three chromosomes representing over 99% of the genome, spanning 136.5 Mb. We provide comprehensive annotations modeling 11,367 genes, and 18,776 transcripts. We then characterized this genome in comparison to the two other available *Culicoides* genomes, as well as representatives from every dipteran genera that have chromosome-scale genomes. This analysis identified orthologous proteins between all 30 tested dipteran species, which was used to identify lineage-specific expansions, such as cytochrome P450-d1-like across *Culicoides.* Additionally, these orthogroups allow researchers to incorporate information for better studied systems like *Drosophila melanogaster*, as we do in Chapter 2. As *Culicoides* are phylogenetically separated by at least 200 million years from other well studied dipterans, like *Anopheles* mosquitoes, comparative research may uncover basal characteristics of vector biology. This highlights the potential of research into this clade to advance our understanding of VBD, while also showing the difficulty

in studying understudied systems like *Culicoides*. The genomic resources provided in this study will serve as a foundation for molecular studies of vector competence, immune function, and control strategy development in this economically important vector group.

Chapter 3 leveraged the newly assembled *C. sonorensis* genome to investigate transcriptomic responses to viral infection, revealing host responses to bluetongue virus (BTV) and vesicular stomatitis virus (VSV). BTV infection elicited a more intense transcriptional response than VSV, with 862 differentially expressed genes compared to 96 for VSV. Utilizing the orthogroups identified in Chapter 2, we were able to analyze functional enrichment in response to viral infection, which revealed upregulation of innate immune pathways in both viral treatments and downregulation of neurological and behavioral pathways during BTV infection. The identification of 55 shared differentially expressed genes between both treatments suggests a core response to viral infection, whereas the high proportion of uncharacterized genes highlights the unique biology of *Culicoides* vectors.

The work presented in chapter 2 and 3 will pave the way for deeper exploration into molecular underpinnings of key aspects of midge biology that can be utilized to control *Culicoides* midges and/or the pathogens they transmit. The differential gene expression patterns we identified reveal potential indicators of behavioral changes in infected insects, such as genes involved in circadian rhythm and locomotor activity. This knowledge may contribute to the development of more effective trapping strategies for infected midges, such as adjusting timing or length of trapping due to disruption of circadian rhythm. Additionally, the new genome presented in this dissertation opens the possibility to study the genetic factors of midge reproduction, leading to opportunities for population control through RNA interference approaches targeting essential reproductive genes, or as targets for gene drive systems that could

spread through wild populations. Furthermore, the egg, larval, pupal and adult specific gene expression profiles provide information on genetic factors that may be key to midge development, providing additional targets for population control through the restriction of progress to specific life stages. Our characterization of genes that respond to viral infection provides a starting point for the development of transmission-blocking strategies, where we could enhance the midge's natural antiviral responses or identify genes that bottleneck viral replication that could be further enhanced. Overall, these data and findings will enable us to investigate new avenues of research for the understanding of *Culicoides* biology and how we can mitigate the damage they cause.

Chapter 4 investigated the utility of house flies as environmental bioindicators for bacterial pathogen and antimicrobial resistance (AMR) surveillance, coupled with the development of BALROG-MON (Appendix A), a specialized metagenomic analysis pipeline. We tested this framework by utilizing Oxford Nanopore long-read metagenomic sequencing of house flies collected from urban and rural environments. We observed distinct microbial community and antibiotic resistances gene (ARG) profiles between these two environments. Urban flies exhibited greater ARG diversity including genes associated with resistance to clinically important antibiotics, while rural flies were enriched with ARGs associated with tetracycline, which has been historically used in agricultural settings. The detection of ESKAPE pathogens and the discovery that 77% of ARGs were predicted to be encoded on plasmids demonstrates the potential for this framework to study antibiotic evolution and detect key bacterial drivers of resistance.

These findings demonstrate that flies serve as environmental samplers that can significantly enhance our AMR surveillance toolkit by providing a cost-effective, scalable and

standardized single-point samples for monitoring across diverse landscapes. The integration of fly-based xenosurveillance into broader AMR monitoring networks will enable early detection of emerging resistance patterns from difficult-to-sample environments that may go unnoticed with by-hand collection methods. This framework will allow for detection of novel ARG mutations and/or concerning genotypes that can be further investigated through culture-based isolation, sequencing, and experimental methods. Additionally, we can identify regions or locations playing key roles in the spread and evolution of AMR, allowing for more precise targeting of surveillance efforts. Furthermore, the ability to associate the ARGs we detect with mobile genetic elements, such as plasmids, suggests that flies may allow us to examine the risk of ARGs being spread across taxa via horizontal gene transfer events, one of the main driving factors of the AMR crisis we face today. This fly-based xenosurveillance approach can be integrated with One Health surveillance frameworks, providing a single sampling method that can bridge the gap between the environmental, agricultural, and human spheres into a comprehensive monitoring system.

Throughout this dissertation we demonstrate the value of proper data stewardship practices, and the use of reproduceable bioinformatics. The data produced and described in Chapter 2 were directly used in subsequent work in Chapter 3 and will be useful to the broader *Culicoides* scientific community when uploaded to public databases. Furthermore, standardized reproducible analytical pipelines, like BALROG-MON, allow for the analysis of complex datasets without compromising reusability or data trustworthiness. The data sets, and analytical workflows that were developed in this dissertation, will be preserved in accordance to FAIR practices and will be publically available to other researchers. The *Culicoides* genome represents a global reference that will be reused for comparative genomics, gene function studies, and the

development of targeted control strategies for vector-borne diseases. Together with the BALROG-MON pipeline and house fly xenosurveillance framework, this work demonstrates how robust data and reproducible tools can extend beyond a single study to enable ongoing research in One Health systems, advancing our understanding of the mechanisms of VBD.

Appendix A - BALROG-MON: a high-throughput pipeline for Bacterial AntimicrobiaL Resistance annOtation of Genomes-Metagenomic Oxford Nanopore

Published as: Edward Bird, Victoria Pickens, David Molik, Kristopher Silver, and Dana Nayduch. BALROG-MON: a high-throughput pipeline for Bacterial AntimicrobiaL Resistance annOtation of Genomes-Metagenomic Oxford Nanopore. microPublication Biology.

Abstract

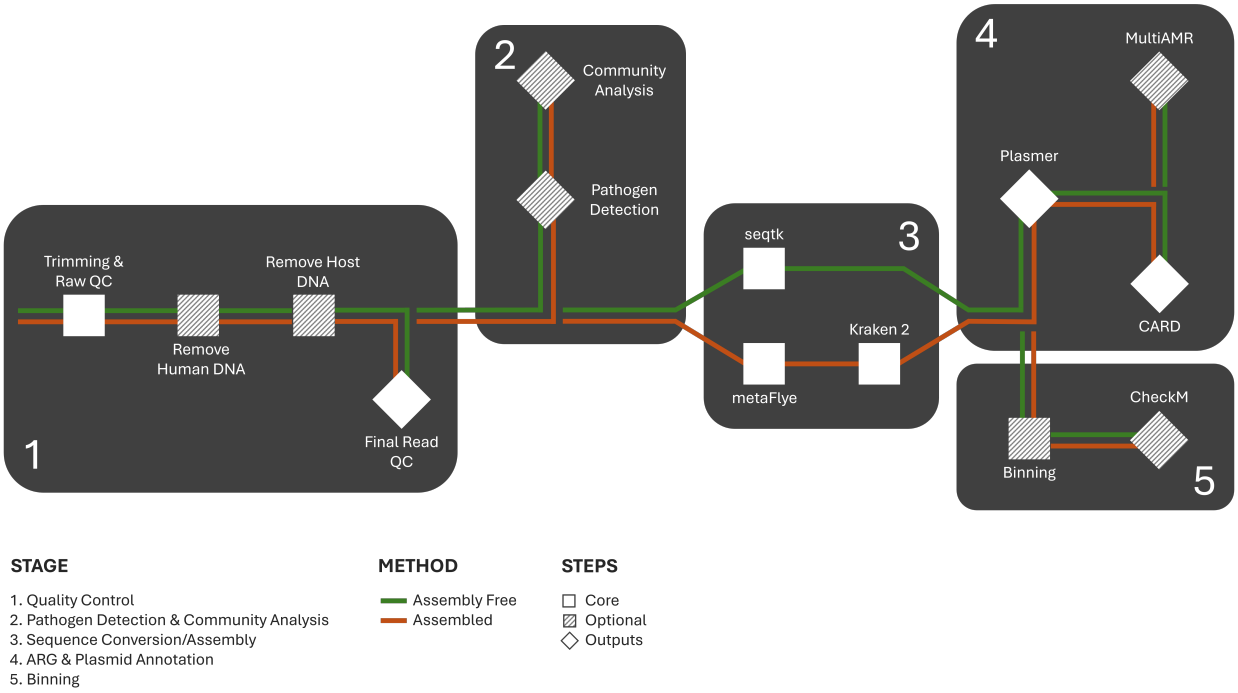
BALROG-MON is a Nextflow pipeline for automated analysis of metagenomic long-read data to detect pathogens, annotate antimicrobial resistance genes (ARGs), link ARGs to specific pathogens, predict ARG origin (e.g., plasmid, chromosomal) and optionally perform steps like community analysis. With both assembly-based and assembly-free workflows, BALROG-MON is applicable to a wide range of sample types with low or high coverage, varying complexities and origins. Optional genome binning provides a comprehensive overview of ARGs within the dataset. BALROG-MON additionally presents results in summarized reports, overall serving as a flexible analysis tool for exploring diverse metagenomic samples for pathogens and antibiotic resistance.

Description

BALROG-MON (Bacterial AntimicrobiaL Resistance annOtation of Genomes – Metagenomic Oxford Nanopore) (https://github.com/edwardbirdlab/BALROG-MON/) automates the analysis of antimicrobial resistance genes (ARGs) from complex metagenomes. Antimicrobial resistance (AMR) is a significant global health challenge exacerbated by the increasing prevalence of resistance to and the declining discovery of new antibiotics. Over the past decade, multiple bioinformatic tools and databases have been developed to identify,

characterize and understand the evolution of ARGs. While many tools focus on isolated bacterial genomes, BALROG-MON analyzes antimicrobial resistance metagenomically, offering a high-throughput approach to study ARGs across entire environments. It also provides insights into AMR in pathogens without the need for culture-based methods, making it a powerful tool for understanding resistance in complex microbial communities and what threats may be present.

Metagenomic sequencing is a recently feasible method that can generate data from which ARGs, microbiomes, and pathogens in the sample can be characterized simultaneously, serving as a more effective and comprehensive method than isolating and culturing bacteria. Typical analysis of metagenomic data involves either an assembly-based approach or a read-based approach, with each having its own benefits and limitations. Metagenomic assembly allows for upstream or downstream investigation of ARGs and provides accurate identification of their origin. However, this approach may lead to information loss, as low-coverage genomes are often not assembled. In contrast, read-based approaches enable mapping of all available data but lack the capability to explore surrounding genomic context or provide accurate taxonomic classifications. To address these challenges, we developed BALROG-MON, a versatile and reproducible Nextflow pipeline for surveying pathogens and ARGs from metagenomic long-read sequencing, offering both “assembled” and “assembly-free” workflow options.



Appendix Figure A.1 BALROG-MON workflow diagram

BALROG-MON enables the option of using an "assembled" or "assembly-free" method for the annotation of ARGs from metagenomic long read sequences.

BALROG-MON v1.0 (DOI: 10.5281/zenodo.14850876) consists of six major steps: (1) Quality Control, (2) Pathogen Detection and Community Analysis, (3) Sequence Conversion/Assembly, (4) ARG and Plasmid Annotation, (5) Binning (*optional*) and (6) Output Collection and Summary. Written in Nextflow, BALROG-MON compartmentalizes each process, allowing users to easily modify or customize steps to suit their analysis needs. All processes are executed with a single command, simplifying the workflow while ensuring scalability. Docker containers manage all dependencies, ensuring reproducibility across different computing environments. During data quality control host sequences and human sequences can be removed, and low-quality bases and adapters are trimmed. Quality-controlled sequences are then either assembled in assembly mode or converted to FASTA format in assembly-free mode. For high-coverage metagenomes, binning can be enabled, allowing reads or contigs to be

grouped for generating metagenomically reconstructed genomes. The resulting sequences are classified as either plasmid or chromosomal in origin. ARGs are annotated using multiple tools, and the outputs are standardized. ARGs, taxonomic classifications, and plasmid classification results are integrated to generate a comprehensive report for each sample. Additionally, all quality control metrics (trimming, host and human sequence removal, final read QC) are summarized in a final report.

Input data and, if also provided, reference genome(s) are first run through data_validator v1.0 (https://github.com/edwardbirdlab/nextflow_input_std), a custom Python script that checks input data for a valid format and, if necessary, reformats FASTQ and FASTA files to the expected format. FastQC v0.12.1 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is run on raw reads to provide statistics on the quality of sequencing, and Porechop v0.2.4 (https://github.com/rrwick/Porechop) and chopper v0.7.0 (https://github.com/wdecoster/chopper) quality control reads. Optionally, reads can be mapped to the human genome (GCA_000001405.15_GRCh38) using minimap2 v2.26 (https://github.com/lh3/minimap2), and then non-human reads extracted using Samtools v1.17 (https://www.htslib.org/). Reads can also be mapped to one or more provided reference genomes, and host sequences removed as described above. Final quality-controlled host- and/or human-depleted reads are run through FastqQC v0.12.1 for final quality metrics.

Quality-controlled reads are classified utilizing Kraken 2 v2.1.3 (https://github.com/DerrickWood/kraken2) and Kraken's pre-built PlusPFP (https://benlangmead.github.io/aws-indexes/k2) database by default. Optionally, species-level composition of the metagenome can be estimated using Bracken v2.9 (https://github.com/jenniferlu717/Bracken) and the same database used in Kraken 2 sequence

identification. KrakenTools v1.2 (https://github.com/jenniferlu717/KrakenTools) then calculates Shannon's alpha diversity and Bray-Curtis dissimilarity (Beta Diversity), as well as a Krona chart and MPA style report.

The sequence processing module workflow differs between BALROG-MON run modes (e.g., "assembly" or "assembly-free"). In assembly-free mode, quality-controlled reads are simply converted from FASTQ to FASTA format with SeqTK v1.4 (https://github.com/lh3/seqtk). In assembly mode, a draft metagenome will be assembled with metaFlye v2.9.3 (https://github.com/mikolmogorov/Flye) and contigs re-identified with Kraken 2 and the PlusPFP database. Finally, in both assembly or assembly-free mode, QUAST v5.2.0 (https://github.com/ablab/quast) is run to gather statistics (N50, Total Length, etc.) on the output sequences.

Sequences are first predicted to be of plasmid or chromosomal origin using Plasmer v23.04.20 (https://github.com/nekokoe/Plasmer), and then renamed to reflect their origin. ARGs are then annotated using Resistance Gene Identifier (RGI) v6.0.3 (https://github.com/arpcard/rgi) and the Comprehensive Antibiotic Resistance Database (CARD) (Alcock et al., 2023) protein homologue model. Optionally, sequences can also be annotated with AMRFinderPlus v4.0.19 (https://github.com/ncbi/amr) and Resfinder v4.4.2 (https://github.com/cadms/resfinder).

If binning is enabled, LRBinner v2.1 (https://github.com/anuradhawick/LRBinner) will bin reads in assembly-free mode, or COMEBin v1.0.3 (https://github.com/ziyewang/COMEBin) will bin contigs in assembly mode. Genome completeness and binning accuracy is then assessed with CheckM v1.2.3 (https://github.com/Ecogenomics/CheckM).

The deliverables from BALROG-MON are summaries that combine and/or visualize the pipeline results. The main output from BALROG-MON is the summary that creates a table based

on the combination of Kraken2 sequence identities, Plasmer sequence classifications (Plasmid/Chromosome), and ARGs. This table allows the user to investigate each ARG and determine the putative origin or location (e.g., plasmid vs. chromosomal). Additionally, all quality control metrics, including those from raw sequence data, trimming, host/human sequence removal, and final read metrics, are summarized using MultiQC v1.22.3 (https://github.com/MultiQC/MultiQC). Lastly, if multiple ARG annotation tools are used, hAMRonization v1.0.2 (https://github.com/pha4ge/hAMRonization) standardizes outputs by harmonizing gene names, recording databases and software versions, and consolidating results into a single output. Additionally, an easy to use web-based report is generated that combines all samples for easy cross sample and cross database comparisons.

BALROG-MON is a robust and adaptable Nextflow pipeline developed for surveying pathogens and ARGs from metagenomic long-read sequencing data. By offering both "assembled" and "assembly-free" workflows, BALROG-MON overcomes the limitations of existing tools, providing researchers with a powerful solution to study ARGs in a wide range of samples, including methods to analyze host associated and low coverage metagenomes. The pipeline incorporates several key steps, including data quality control, read-based sequence identification, ARG and plasmid annotation, and optional metagenomic binning, ensuring a detailed and accessible analyses of complex metagenomic samples. Notable features include versatility in pipeline execution to handle difficult samples, the integration of multiple ARG databases, plasmid prediction capabilities, and optional binning for deeper analysis. BALROG-MON unifies these functionalities into a streamlined, user-friendly platform, simplifying the study of ARGs in metagenomic datasets. As the field of metagenomics advances, BALROG-MON stands poised to play a vital role in elucidating the dynamics of AMR and shaping

strategies to address this global health challenge across human, animal, and environmental domains.

References

Alcock, B. P., W. Huynh, R. Chalil, K. W. Smith, A. R. Raphenya, M. A. Wlodarski, and A. G. McArthur. 2022. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. Nucleic Acids Res. 51: D690-D699.

Andrews, S. 2010. FastQC. Computer software.

Bortolaia, V., R. S. Kaas, E. Ruppe, M. C. Roberts, S. Schwarz, V. Cattoir, and F. M. Aarestrup. 2020. ResFinder 4.0 for predictions of phenotypes from genotypes. J. Antimicrob. Chemother. 75: 3491-3500.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, and H. Li. 2021. Twelve years of SAMtools and BCFtools. GigaScience 10: giab008.

De Coster, W., and R. Rademakers. 2023. NanoPack2: population-scale evaluation of long-read sequencing data. Bioinformatics 39: btad311.

Feldgarden, M., V. Brover, N. Gonzalez-Escalona, J. G. Frye, J. Haendiges, D. H. Haft, and W. A. Klimke. 2021. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci. Rep. 11: 12728.

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29: 1072-1075.

Kolmogorov, M., D. M. Bickhart, B. Behsaz, A. Gurevich, M. Rayko, S. B. Shin, and P. A. Pevzner. 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat. Methods 17: 1103-1110.

Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094-3100.

Li, H. Seqtk. Computer software.

Lu, J. Kraken Tools. Computer software.

Lu, J., F. P. Breitwieser, P. Thielen, and S. L. Salzberg. 2017. Bracken: estimating species abundance in metagenomics data. PeerJ Comput. Sci. 3: e104.

Mendes, I., E. Griffiths, A. Manuele, D. Fornika, S. H. Tausch, T. Le-Viet, and F. Maguire. 2024. hAMRonization: Enhancing antimicrobial resistance prediction using the PHA4GE AMR detection specification and tooling. bioRxiv 2024.03.07.583950.

Parks, D. H., M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25: 1043-1055.

Wang, Z., R. You, H. Han, W. Liu, F. Sun, and S. Zhu. 2024. Effective binning of metagenomic contigs using contrastive multi-view representation learning. Nat. Commun. 15: 234.

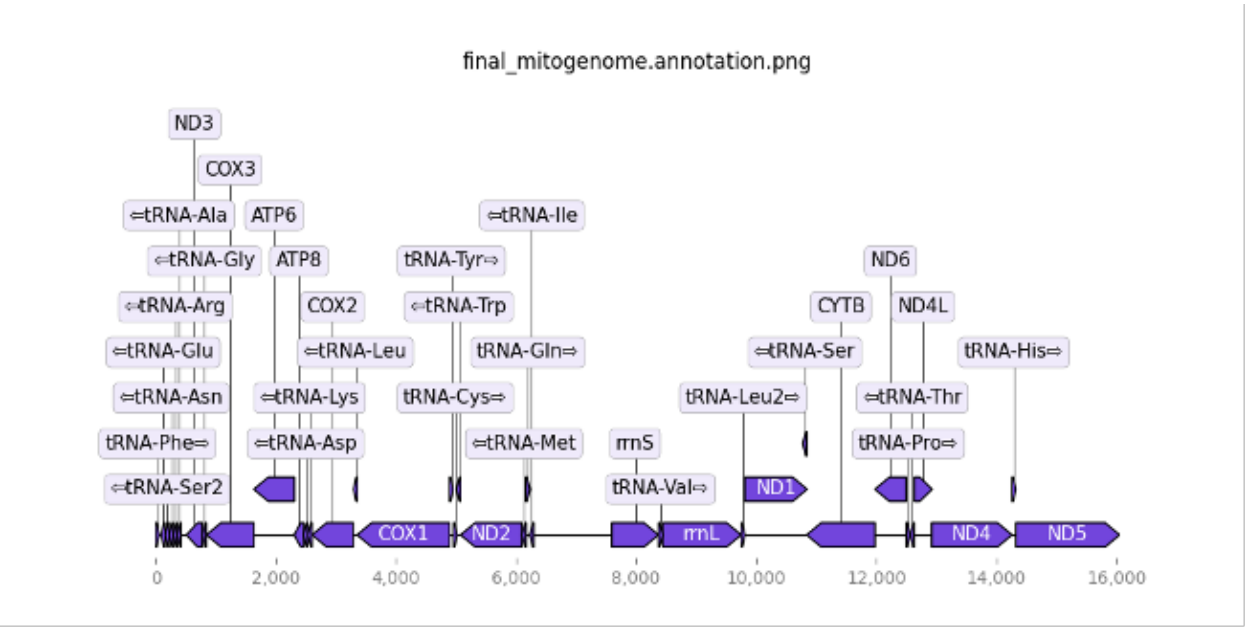
Wick, R., and J. Volkening. Porechop. Computer software.

Wickramarachchi, A., and Y. Lin. 2022. Binning long reads in metagenomics datasets using composition and coverage information. Algorithms Mol. Biol. 17: 3.

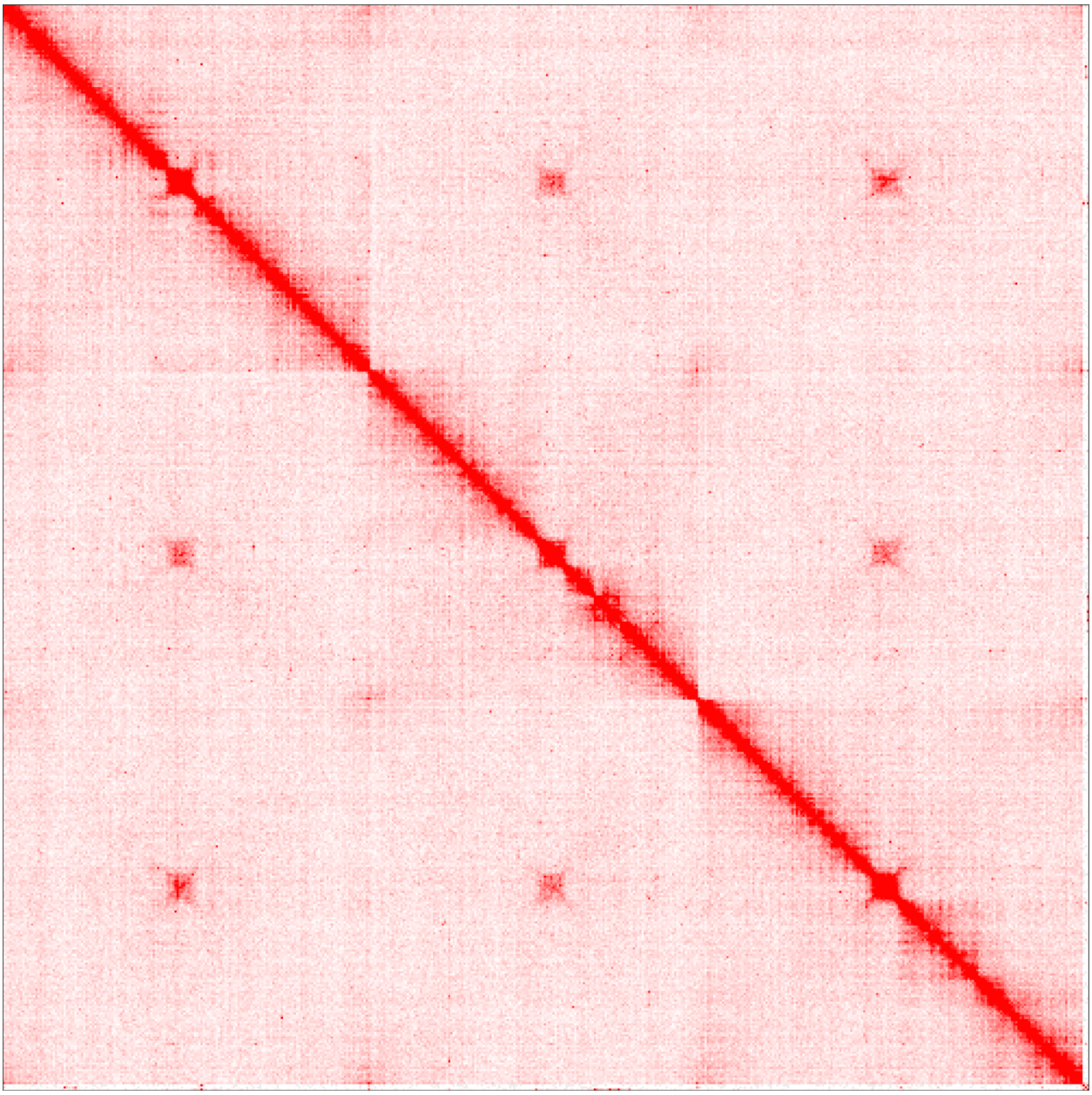
Wood, D. E., J. Lu, and B. Langmead. 2019. Improved metagenomic analysis with Kraken 2. Genome Biol. 20: 257.

Zhu, Q., S. Gao, B. Xiao, Z. He, and S. Hu. 2023. Plasmer: An accurate and sensitive bacterial plasmid prediction tool based on machine learning of shared k-mers and genomic features. Microbiol. Spectr. 11: e04645-22.

Appendix B - Supplemental Information for Chapter 2

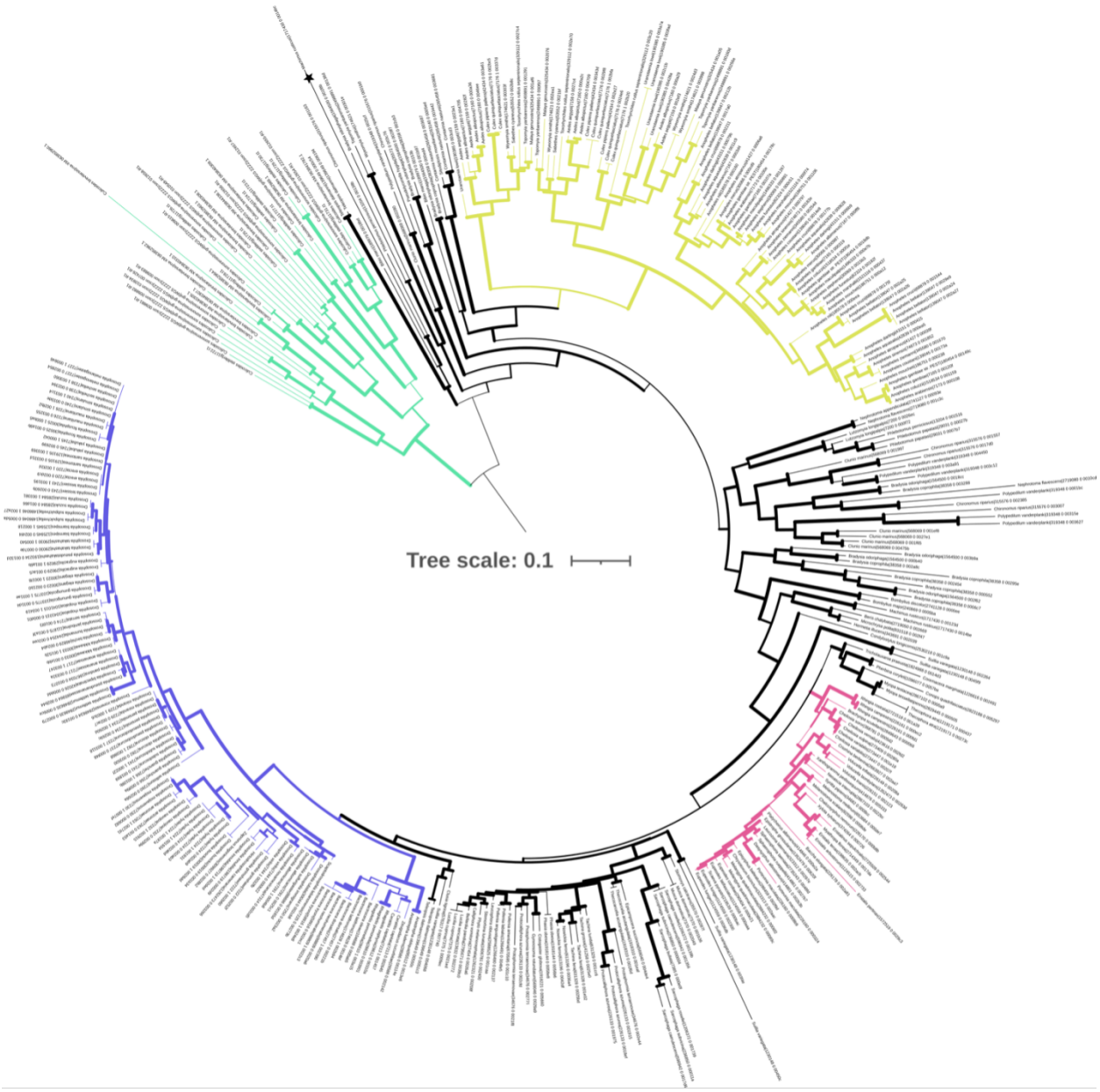


Appendix Figure B.1 Annotation of the *C. sonorensis* circular mitochondrial genome. This figure and annotation were generated by MitoHiFi and represent the best circular mitochondrial genome. All thirty-seven of the expected mitochondrial genes were identified.



Appendix Figure B.2 Hi-C Contact map for *C. sonorensis*

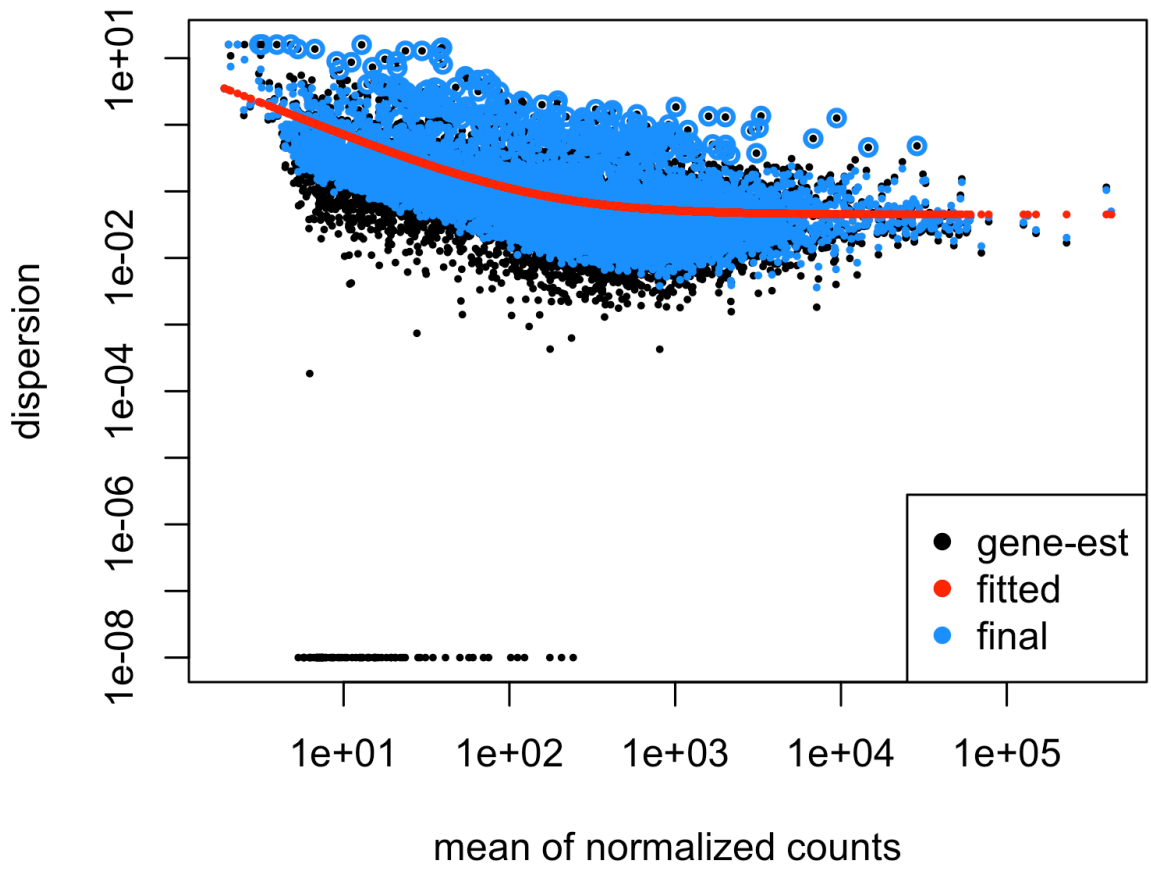
The Hi-C Contact map showing all three chromosomes and unplaced scaffolds (bottom right). DNA sequences are represented on the diagonal, top left to bottom right, and Hi-C contacts between genomic locations are shown in red.



Appendix Figure B.3 Gene tree of dipteran Cyp4d1-like proteins identified in *C. sonorensis*

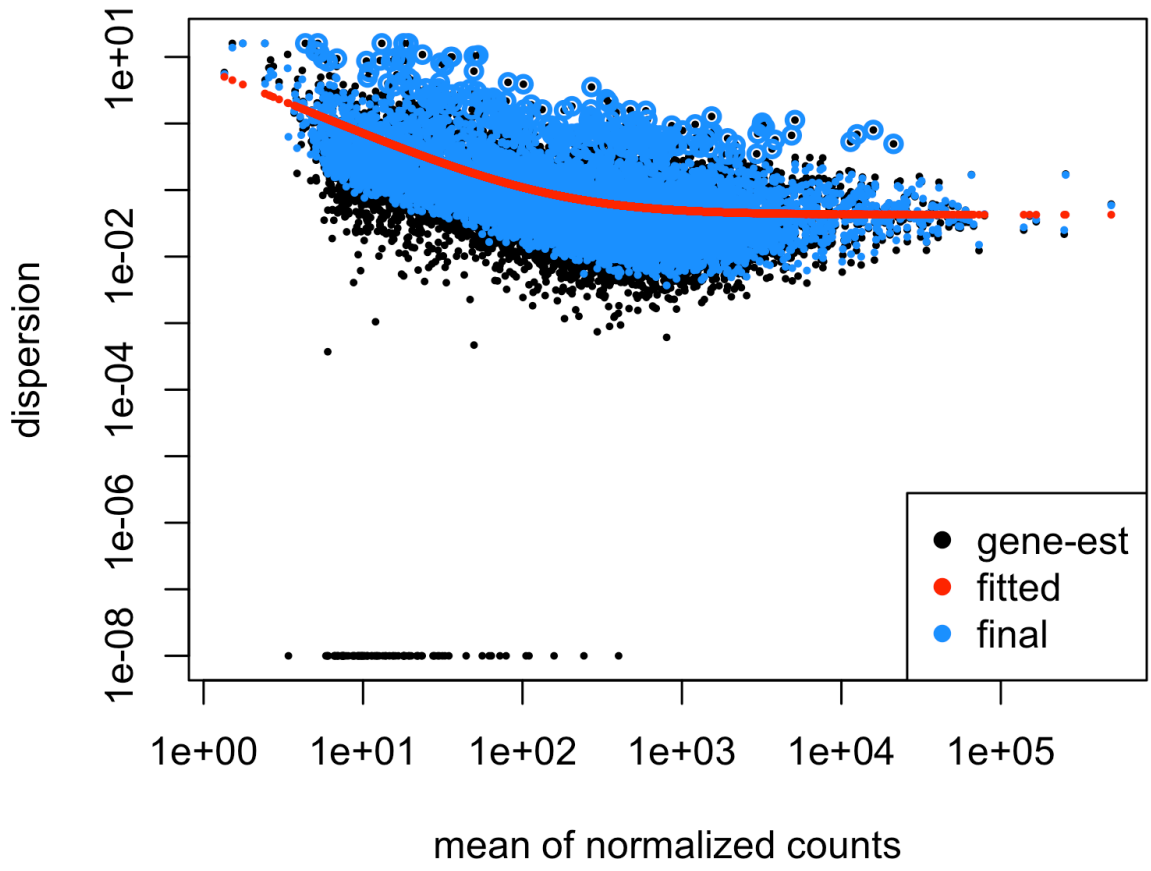
Gene tree of Cyp4d1-like proteins from OG0002693, a *Culicoides*-specific orthogroup identified in this study, and OrthoDB Cyp4d1-like orthogroup (7227_0:002854). *Culicoides* proteins are shown in green, Culicidae in yellow, Syrphidae in pink, and Drosophilidae in blue. Ultra-Fast Bootstrap values are represented by line thickness.

Appendix C - Supplemental Information for Chapter 3



Appendix Figure C.1 Dispersion plot of BTV infected vs. PBS control treatments of *C. sonorensis.*

Plot showing shrinkage and dispersion of gene counts in BTV vs. PBS Deseq2 Model.



Appendix Figure C.2 Dispersion plot of VSNJV infected vs. PBS control treatments of *C. sonorensis.*

.

Plot showing shrinkage and dispersion of gene counts in BTV vs. PBS Deseq2 Model.