

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

An entrée of neurotropic teratogenic viruses with a side of CNS inflammation and ecology for dessert

Permalink

<https://escholarship.org/uc/item/4vg171k9>

Author

Retallack, Hanna E. G.

Publication Date

2020

Peer reviewed|Thesis/dissertation

An entrée of neurotropic teratogenic viruses with a side of CNS inflammation and ecology for dessert

by
Hanna Retallack

DISSERTATION

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

in

Biomedical Sciences

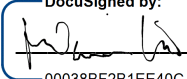
in the

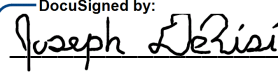
GRADUATE DIVISION

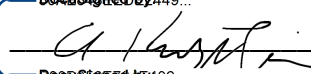
of the

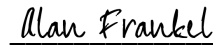
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:

00038BF2B1EE40C... Melanie Ott
Chair

DocuSigned by:

00038BF2B1EE40C... Joseph L. DeRisi

DocuSigned by:

00038BF2B1EE40C... Arnold Kriegstein

DocuSigned by:

A0A99322C05A480... Alan Frankel

Committee Members

Copyright 2020

by

Hanna Retallack

Acknowledgements

I could not have completed this dissertation without the help of many individuals in my professional and personal life. For their guidance and support, I thank my thesis committee, Melanie Ott, Arnold Kriegstein, and Alan Frankel; my PI, Joe; key advisors, Joanne Engel, Anita Sil, and Amy Kistler; my contemporaries in the DeRisi Lab; and my classmates. To my family and extended family, your patience, encouragement, and much listening to dissection details, frustrated complaints and celebration over the smallest successes have not gone unnoticed.

Contributions

Chapter 1

Includes material previously published in:

Retallack H*, Di Lullo E*, Arias C, et al. Zika virus cell tropism in the developing human brain and inhibition by azithromycin. *Proc Natl Acad Sci U S A*. 2016;113(50):14408-14413. doi:10.1073/pnas.1618029113

Chapter 2

Includes contributions from Galina Schmunk, David Shin, Denise Allen (sample acquisition), Tomasz Nowakowski, Arnold Kriegstein (guidance and resources).

Chapter 3

Includes contributions from: Galina Schmunk (sample acquisition, assistance with cell culture and experiments with dissociated microglia), Tomasz Nowakowski (guidance), Tom Hobman (infectious clone of rubella strain M33).

Chapter 4

Includes material previously published in:

Mandel-Brehm C*, Retallack H*, Knudsen GM, et al. Exploratory proteomic analysis implicates the alternative complement cascade in primary CNS vasculitis. *Neurology*. 2019;93(5):e433-e444. doi:10.1212/WNL.00000000000007850

Chapter 5

Includes material previously published in:

Retallack H, Okihiro MS, Britton E, Sommeran SV, DeRisi JL. Metagenomic next-generation sequencing reveals miamiensis avidus (ciliophora: scuticociliatida) in the 2017

epizootic of leopard sharks (*Triakis semifasciata*) in San Francisco bay, California, USA. *J Wildl Dis.* 2019;55(2):375-386. doi:10.7589/2018-04-097

Chapter 6

Includes material previously published in:

Retallack H, Clubb S, DeRisi JL. Genome Sequence of a Divergent Avian Metapneumovirus from a Monk Parakeet (*Myiopsitta monachus*). *Microbiol Resour Announc.* 2019;8(16):e00284-19. Published 2019 Apr 18. doi:10.1128/MRA.00284-19

Chapter 7

Includes contributions from Pdraig Duignan and Barbie Halaska (Marine Mammal Center, specimen collection and photo documentation, seal), Jill Murray (Oklahoma State University, specimen collection, frogs), David Guzman, Sarah Ozawa, and Tracy Drazenovich (UC Davis, specimen collection and photo documentation, rabbit), Eric Chow and Derek Bogdanoff (UCSF CAT, sequencing).

Chapter 8

Manuscript in preparation, publicly available as:

Joshua Batson, Gytis Dudas, Eric Haas-Stapleton, Amy L. Kistler, Lucy M. Li, Phoenix Logan, Kalani Ratnasiri, Hanna Retallack. Single mosquito metatranscriptomics recovers mosquito species, blood meal sources, and microbial cargo, including viral dark matter bioRxiv 2020.02.10.942854; doi: <https://doi.org/10.1101/2020.02.10.942854>

Chapter 9

Includes material previously published in:

DeRisi JL, Huber G, Kistler A, Retallack H, Wilkinson M, Yllanes D. An exploration of

ambigrammatic sequences in narnaviruses. *Sci Rep.* 2019;9(1):17982. Published 2019
Nov 29. doi:10.1038/s41598-019-54181-3

Chapter 10

Includes contributions from Katerina Popova (assistance with design and performance of ribosome profiling assays), Sara Sunshine (library preparation for ribosome profiling), Aaron Brault (CT and Hsu cell lines), Lori Kohlstaedt (mass spectrometry), Amy Kistler (guidance on experimental design), Joe DeRisi (guidance on experimental design and data interpretation).

An entrée of neurotropic teratogenic viruses with a side of CNS inflammation and ecology for dessert

Hanna Retallack

Abstract

The first three chapters of this dissertation explore the pathogenesis of viruses infecting the human developing brain. Specifically, the tropism and consequences of infection by Zika and rubella virus are defined. While Zika virus was found to infect primarily radial glia and astrocytes, and rubella appeared to target microglia, both stimulate a strong response consistent with type I interferon signaling. The next chapter addresses a rare disease that likewise stems from overactivation of immune response, in this case, inflammation of the vessels of the brain. An exploration of abnormal molecular pathways in primary angiitis of the central nervous system revealed dysregulation of the complement system. Chapters five through eight investigate infectious diseases and commensal microbiota of wildlife. These include epizootics among sharks and parrots, and a demonstration of how metagenomic next-generation sequencing can color in an ecological diagram of mosquitoes, their bloodmeal hosts, and the commensals and pathogens they carry. Diving deeper into one component of the mosquito microbiota, the final chapters grapple with a surprising feature of the narnavirus genome. These simple entities, named NAKed-RNA-viruses, turn out to be far more complex than previously recognized.

Table of Contents

Introduction.....	1
REFERENCES FOR INTRODUCTION.....	9
Chapter 1 Zika Virus in the Developing Human Brain: Cell Tropism and Drug Inhibition.....	11
REFERENCES FOR CHAPTER 1	31
Chapter 2 Innate immune response of the developing brain to viruses.....	37
REFERENCES FOR CHAPTER 2	57
Chapter 3 Tropism of rubella virus in the human developing brain.....	61
REFERENCES FOR CHAPTER 3	74
Chapter 4 Exploratory proteomic analysis implicates the alternative complement cascade in Primary CNS Vasculitis.....	76
REFERENCES FOR CHAPTER 4	104
Chapter 5 Metagenomic Next-Generation Sequencing Reveals <i>Miamiensis Avidus</i> (Ciliophora: Scuticociliatida) in the 2017 Epizootic of Leopard Sharks (<i>Triakis</i> <i>Semifasciata</i>) In San Francisco Bay, California, USA.....	109
REFERENCES FOR CHAPTER 5	135
Chapter 6 Genome Sequence of a Divergent Avian Metapneumovirus from a Monk Parakeet.....	140
REFERENCES FOR CHAPTER 6	149

Chapter 7 Wildlife investigations with no infectious agent identified by mNGS.....	151
REFERENCES FOR CHAPTER 7	157
Chapter 8 Single Mosquito Metatranscriptomics Recovers Mosquito Species, Blood Meal Sources, and Microbial Cargo, Including Viral Dark Matter	158
REFERENCES FOR CHAPTER 8	219
Chapter 9 An exploration of ambigrammatic sequences in narnaviruses.....	234
REFERENCES FOR CHAPTER 9	255
Chapter 10 Functional relevance of the narnavirus' unique ambigrammatic genome.....	260
REFERENCES FOR CHAPTER 10	270

List of Figures

Figure 1.1 Tropism of ZIKV for radial glia in the developing human brain.	18
Figure 1.2 ZIKV infects astrocytes in later stages of human brain development.	20
Figure 1.3 Azithromycin (AZ) treatment inhibits ZIKV infection in glial cells.	21
Figure 2.1 Approach: scRNA-Seq on primary human fetal brain tissue after in vitro infection.	40
Figure 2.2 Overall transcriptional response to ZIKV vs. mock.	40
Figure 2.3 Consistency in transcriptional response to ZIKV between replicate samples.	41
Figure 2.4 Cell type classification approach for scRNAseq	41
Figure 2.5 Cell type classification in scRNAseq data.	42
Figure 2.6 Strongest transcriptional response to ZIKV observed in radial glia and astrocytes.	42
Figure 2.7 Distinguishing whether cell types show fundamentally different responses or different amplitudes of the same response.	43
Figure 2.8 Cell type-specific responses to ZIKV	43
Figure 2.9 Putative cell-type specific responses to ZIKV	44
Figure 2.10 Upregulation of interferon-beta (IFNB) after ZIKV infection.	45
Figure 2.11 Similarity of transcriptional response to ZIKV and to IFNB treatment.	45
Figure 2.12 Dramatic increase in ISG15 RNA after ZIKV infection.	48
Figure 2.13 Rapid increase in ISG15 RNA after treatment with recombinant IFNB.	48
Figure 2.14 Timecourse of type I interferon activation in primary neural cells	49

Figure 2.15 Protein upregulation and activation in ISGs after treatment of primary human brain cells with recombinant human IFNB.....50

Figure 2.16 Weak response to IFNB at baseline and no significant difference after attempted knockdown of USP18 RNA by shRNAs.....50

Figure 2.17 Cell death vs. proliferation in primary brain tissue treated with IFNB.52

Figure 2.18 Proliferation in germinal zone after IFNB treatment of primary brain tissue. ..53

Figure 2.19 Relative size of Ki67-strong germinal zone after IFNB treatment of primary brain tissue.54

Figure 2.20 Coarse assessment of progenitors and neurons in primary brain tissue treated with IFNB.....55

Figure 2.21 Coarse assessment of microglia in primary brain tissue treated with IFNB.56

Figure 3.1 Rubella virus capsid protein in microglia after infection of primary human brain tissue64

Figure 3.2 Single cell RNA sequencing of Rubella virus (RV)- and mock-infected organotypic slice culture.....65

Figure 3.3 Rubella capsid in microglia depends on presence of other cell types67

Figure 3.4 Titering and qPCR for RV in tissue at multiple days after inoculation of slice culture.68

Figure 3.5 qPCR for RV in supernatant at multiple days after inoculation of slice culture.....68

Figure 3.6 Titering for RV in supernatant at multiple days after inoculation of Vero cells or primary microglia in co-culture.....69

Figure 3.7 Validation of FISH probes and anti-RV capsid antibody for detection of rubella virus.....	70
Figure 3.8 Validation of negative strand FISH on RV-infected Vero cells.....	70
Figure 3.9 RV (+) strand accumulates in cells with anti-RV capsid.....	71
Figure 3.10 Assessment of RV (-) strand RNA in microglia by FISH.....	71
Figure 3.11 RV M33 strain with inserted GFP, expressed as fusion to p150.	72
Figure 4.1 Demonstrative biopsies from patients with primary angiitis of the CNS (PACNS).	92
Figure 4.2 Unbiased clustering of patients by CSF proteome.	93
Figure 4.3 Comparison of discriminating proteins across cohorts.....	95
Figure 4.4 Enrichment of peptides in extracellular domains in proteins downregulated in PACNS.	96
Figure 4.5 Molecular phenotype in primary angiitis of the CNS (PACNS) CSF is informed by proteomic comparison of CSF between PACNS and noninflammatory control (NIC) cohorts.	98
Figure 5.1 Map and photographs of shark strandings occurring in San Francisco Bay (SF Bay) in spring of 2017.	117
Figure 5.2 Gross observations of lesions in brains of stranded leopard sharks (<i>Triakis semifasciata</i>) in San Francisco Bay in spring of 2017 in which the scuticociliate <i>Miamiensis avidus</i> was involved.	118

Figure 5.3 Molecular identification of the scuticociliate parasite <i>Miamiensis avidus</i> in cerebrospinal fluid from leopard sharks (<i>Triakis semifasciata</i>) in San Francisco Bay in spring of 2017.	121
Figure 5.4 Histology showing protozoa morphologically consistent with the scuticociliate <i>Miamiensis avidus</i> in brain tissues of stranded leopard sharks in San Francisco Bay in spring of 2017.	123
Figure 5.5 Precipitation in regions draining to San Francisco Bay (SF Bay) by year, including 2017 when strandings of leopard sharks (<i>Triakis semifasciata</i>) were associated with infection by <i>Miamiensis avidus</i>	128
Figure 5.6 Field specimen collection, 2017.	130
Figure 5.7 Moribund leopard shark (<i>Triakis semifasciata</i>) stranded in San Mateo, CA, 2019.	131
Figure 5.8 Culture of <i>Miamiensis avidus</i> , fixed and stained for acetylated tubulin and DAPI.	132
Figure 5.9 Z-stack of cultured <i>Miamiensis avidus</i> , showing micronucleus.	134
Figure 6.1 Coverage and phylogenetic analysis of sequence representing a new subgroup of avian metapneumovirus.	143
Figure 6.2 Validation of Diagnostic PCR for aMPV-E.	146
Figure 6.3: Dissection of Quaker parrot (monk parakeet, <i>Myiopsitta monachus</i>) that had died during an outbreak of aMPV-E following sudden onset illness.	147
Figure 6.4: PCR identifies aMPV in additional parrots, by organ.	148

Figure 7.1 Alopecia in female ribbon seal Toboggan (<i>Histiophoca fasciata</i>), stranded at Morro Bay December 2017.....	153
Figure 7.2 Papillomatous anorectal mass of uncertain etiology in a Dutch Dwarf rabbit (<i>Oryctolagus cuniculus</i>).	154
Figure 7.3 Blanchard’s Cricket Frogs to be processed for mNGS.....	155
Figure 8.1 Number of each genus and species of mosquitoes collected across 5 regions in California.....	164
Figure 8.2 Read count distribution across non-host taxonomic groups detected among the set of 148 mosquitoes.	166
Figure 8.3 Quantifying contribution of new viral sequences.....	171
Figure 8.4 Distribution of viruses, Wolbachia, and eukaryotes in single mosquitoes.	173
Figure 8.5 Consensus taxonomic calls of vertebrate contigs for 45 of 60 blood fed mosquitoes collected in Alameda County.	176
Figure 8.6 Previously unrecognized Orthomyxovirus genome segments identified among unaligned “dark matter” contigs using co-occurrence analysis.....	180
Figure 8.7 Phylogenetic analysis of Wuhan mosquito virus 6.	182
Figure 8.8 Unbiased identification of mosquito species.....	208
Figure 8.9 Breakdown of reads for each sample into broad categories.....	209
Figure 8.10 Analysis of peribunya-like virus showing completeness.	210
Figure 8.11 Schematic describing method for branch length contribution calculation. ..	211
Figure 8.12 Distribution of mosquitoes within the study in which no, one, or multiple viral lineages were detectable.	211

Figure 8.13 An example of viruses with similar bulk abundance but different prevalence.....	212
Figure 8.14 Prevalence of each virus by mosquito species.....	212
Figure 8.15 Co-occurrence analysis to identify additional viral segments.....	213
Figure 8.16 Methods for co-occurrence analysis.	214
Figure 8.17 Genome diagrams for newly proposed viral segments.	215
Figure 8.18 RdRp-based maximum likelihood tree spanning the quaranjaviruses in this study for which 8 segments were recovered.	216
Figure 8.19 Phylogenetic relationship of narnavirus RdRp and Robin segments.	217
Figure 8.20 Lack of association between abundance of narnavirus and fungi in individual mosquitoes.....	218
Figure 9.1 Ambigrammatic sequences in narnaviruses.....	236
Figure 9.2 Labelling conventions used in this paper for reading frames.....	237
Figure 9.3 Maximum likelihood tree of amino-acid sequences for RNA-dependent RNA polymerase (RdRp) of 42 representative narnaviruses, identified by homology to the narnaviruses observed in culture, <i>Culex</i> narnavirus 1 and <i>Phytophthora infestans</i> virus 4.	238
Figure 9.4 Probability distribution for ORF lengths in narnavirus-like sequences.	245
Figure 10.1 Description of <i>Culex</i> narnavirus 1 (CxNV1) in <i>Culex tarsalis</i> (CT) cell line..	263
Figure 10.2 Persistence of CxNV1 RdRp and Robin RNA depends on active RdRp.	265
Figure 10.3 Ribosome profiling of CT cells shows a unique pattern on transcripts of CxNV1 RNAs.....	268

List of Tables

Table 3.1 Number of cells from single cell RNA sequencing of Rubella virus or mock-infected organotypic slice culture with (+) or without (-) RV RNA.....	65
Table 4.1 Demographics and clinic features of the PACNS, RCVS, and NIC cohorts.....	88
Table 5.1 Histopathologic lesions in stranded and captive sharks from SF Bay.	120
Table 5.2 Histopathologic lesions in stranded and captive sharks from San Francisco Bay.....	122
Table 6.1: Primers for diagnostic testing for aMPV-E (based on MK49199 genome sequence)	145
Table 7.1 Sequencing metrics for ribbon seal	154
Table 7.2 Sequencing metrics for frogs.....	156

Introduction

There is a statistic for every disease that makes it the greatest threat to human health. For example: 1) Rapid global spread of Zika virus from the outbreak's peak in 2016 has led to local transmission by mosquitoes in 87 countries as of July 2019 (WHO 2019). Even with the best therapy, nearly half of patients with Primary Angiitis of the Central Nervous System (PACNS) relapse with debilitating neurologic symptoms (Hutchinson et al. 2010a). Likewise, there is an explanation for every biological feature that makes it the most fundamental to understanding life: 1) Translation is the process that makes every protein in your body. 2) The interferon-response pathway is the master regulator of a cell's initial response to viral infection. I don't claim that these statistics and explanations motivate the various projects discussed in this dissertation. I only claim that these projects are worthwhile for having fascinated me, having taught me, and having revealed a smidgen more about biology. If these projects had to fit under a topical umbrella, it would be "Infectious Disease and Immune Response". But in reality, experiments and happenstance steer a winding route. A traditional scientific contextualization can be found at the beginning of each chapter. The remainder of the introduction discusses cross-cutting themes, often tying together disparate projects that resulted from simply following my curiosity.

In various pathogen-centered projects, the **IMMUNE RESPONSE** of the host played a key role. When studying Zika virus infection of the developing human brain (Chapter 1), a simple transcriptional experiment showed the tremendous immune response of the cells

during infection of the tissue. These data opened a new line of inquiry, asking how the innate immune response to viruses could interfere with neural development (Chapter 2). Indeed, genetic syndromes suggest that the tissue-level response may be damaging in itself (Livingston and Crow 2016), and so a myopic focus on virally-infected cells alone likely ignores important mechanisms of pathogenesis. Of course, ethical limitations prevent experiments with the fully intact human immune system, and so absence of systemic immune cells is a key limitation of the model system I utilized for Zika virus (organotypic slice culture of primary human fetal brain tissue).

With other pathogens too, the immune system proved critical. I observed that rubella virus was primarily found inside the innate immune cells of the brain: microglia (Chapter 3). Since microglia are phagocytic, it was especially important to determine whether the virus actually replicated inside this cell type. Separately, examining the histology of inflamed brain tissue from deceased sharks (Chapter 5), inflammatory cells were morphologically similar to the unicellular pathogen we suspected of causing the inflammation, necessitating DNA identification of genetic material from the pathogen, *Miamiensis avidus*. In a metagenomic survey of mosquitoes (Chapter 8), we analyzed putative viral sequences for a characteristic signature of being targeted by the mosquito's antiviral defense systems, to support our conclusion that these sequences were indeed viral. As many infectious disease scientists will surely agree, the interaction between pathogen and host is a rich area for exploration. While one might start with a pathogen-focused question, to ignore the immune response would be to blind oneself to a mountain of gold.

Several projects in this dissertation focus on human diseases that affect different **POPULATIONS**. All demographics are susceptible to Zika virus (Chapter 1) and rubella virus (Chapter 3), but pregnant women are especially relevant since these two viruses can cross the placenta and impact the fetus' development. As for geographical distribution, Zika virus infection is most prevalent in tropical countries where the mosquito species that transmit the virus (mostly *Aedes* spp.) are most abundant (Santos and Meneses 2017). Conversely, the impact of rubella virus is greatest in countries whose health system infrastructure and public policy have not yet achieved widespread administration of the highly effective vaccine (Grant et al. 2017). For PACNS, a non-infectious disease discussed in Chapter 4, the burden falls on a tiny population of individuals, with an estimated prevalence of 2.4 cases per 1,000,000 person-years (Salvarani et al. 2007). While these cases are typically identified in countries with advanced medical practices, the disease may be under-recognized in low-resource settings.

For the two neurotropic teratogenic viruses in this dissertation, consider the **TIMING** of my research relative to the emergence and description of the virus in human populations. Zika and rubella show a striking contrast in this regard. Rubella virus was first recognized to cause congenital cataracts in 1941, and the epidemic of birth defects was curbed by highly effective vaccines deployed in the late 1960s (Gregg 1941; Banatvala and Brown 2004). As the vaccine rose, scientific interest waned, and my experience in this project is highly influenced by reviewing decades-old literature that relies on experimental design and interpretation for lack of modern techniques; by deep freezers of generous colleagues whose research program had shut down; and by asking a seemingly

simple question that had not been answered while specimens were easily obtainable:
what cell types of the developing brain are susceptible to infection?

In contrast, asking the same question for Zika virus, in January of 2016 at the peak of the epidemic, was a totally different experience. The enormous and sudden interest in the virus affected the project in many ways, from the source of viral isolates, to the large number of person-hours on the project, to the urgency of possible treatment, to publication. I learned that “getting scooped” is not a useful threat. Far more productive was the accelerated cycle of posit and respond, where multiple groups publish results to similar questions, confirming or modifying what was known before, until the consensus or lingering debate offers the best answers as a whole scientific community and clarifies what is likely biologically true and what is incongruent.

During completion of this dissertation, the same accelerated pace is being driven by the new emerging virus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Given the local needs, my efforts were best spent in operational public health response rather than scientific discovery. Obviously new epidemics can force or incentivize changing the direction of a research program. Still, the old pathogens have much to teach us. Recent investigations into a type of uveitis have shown that rubella virus can persist for decades in immunologically privileged sites such as the eye, causing immune responses that degrade vision decades later (Doan et al. 2016; de Groot-Mijnes et al. 2006; Suzuki et al. 2010). Intriguingly, it appears that the newsworthy RNA viruses, Zika and Ebola, may also persist in similar sites in the body (Varkey et al. 2015; Petridou et al. 2019). It is hard to predict the consequences of persistent replication under immunological pressure for a

particular virus. But the more cases we study, old and new, the more we understand in case a rapidly-mutating persistent virus were to appear.

The projects in this dissertation examine biology at many different **SCALES**, using corresponding **TECHNICAL APPROACHES**. At the smallest scale, I examine RNA structure of a narnavirus and the movement of ribosomes that translate its genome (Chapter 10); the investigation of interferon response pathways focuses on molecular events connecting the outside to the inside of a cell, and between cell types in a tissue (Chapter 2); the cell types that comprise an organ, the brain, are the focus of the Zika and rubella projects (Chapter 1, Chapter 3); identifying a pathogen's target organs within a whole organism played a major role in the description of shark and parrot epizootics (Chapter 5, Chapter 6) and understanding the impact of environmental conditions on sharks' susceptibility to infection (Chapter 5), or the interaction between mosquitoes, their bloodmeal hosts, and their microbiota (Chapter 8) completes the span up to the ecological level. Of course, certain technical approaches lend themselves best to each scale. Ironically, I tackled the largest-scale, ecological questions with technology that looks at the smallest building blocks of life: the genetic code. The power of metagenomic next-generation sequencing to understand living organisms in the context of an entire ecosystem is impressive.

Every project in this dissertation required analysis of large datasets from mass spectrometry (Chapter 4) or from next-generation sequencing (NGS). I utilized NGS to verify viral strains (Chapter 1, Chapter 3, Chapter 10), discover new pathogens (Chapter 5, Chapter 6, Chapter 7), and perform primary analyses such as examining transcriptional responses or investigating viral phylogenetics (Chapter 2, Chapter 8, Chapter 9). Although

many tools exist to simplify standard workflows for NGS data, I found it incredibly valuable to learn some basics in command line tools, server functioning, Bash, Python, and R, in order to manipulate and analyze data independently. I am particularly excited about ongoing projects that utilize these skill sets, developing tools for genetic screening and in vivo monitoring using components of arenaviruses. Without the computational fluency and familiarity from prior opportunities, I could not have easily imagined let alone design these projects, which have been especially fun as a chance to get creative.

Finally, it is impossible to discuss the topics, approaches, and implications of the science without discussing the **TEAMS** that made the discoveries. On some projects, teams were structured traditionally, such as working with a post-doc in the same lab on PACNS (Chapter 4), or joining forces with labs specializing in human brain development for Zika and rubella (Chapter 1 and Chapter 3). In international collaborations related to Zika virus, I met scientists working in incredibly strong infectious diseases research programs in Brazil. Many worked at institutions whose administrative organization impacted day-to-day science in unfamiliar ways. In several projects I worked alongside people whose jobs have many responsibilities beyond scientific discovery: private harbormasters and veterinarians, community volunteers and government employees at county and state agencies. These individuals offered crucial context and a wealth of knowledge about their subject matter that complemented my molecular investigations.

Three collaborators in particular stand out as examples of scientific excellence outside the traditional academic structure. Mark Okihiro, at California Fish and Wildlife, hooked me on a problem in our own backyard: dead leopard sharks washing up on

beaches of San Francisco Bay (Chapter 5). He taught me how to necropsy a shark in the field, and through these necropsies and histology, with stellar photo-documentation, proposed a route of invasion for the pathogen we found. Susan Clubb, a veterinarian specializing in birds, had a keen insight when multiple outbreaks of common infections at her facility were difficult to control, emailing us “I think there is an underlying virus that may be immunosuppressive.” Together, we identified a new strain of avian metapneumovirus (Chapter 6), which is closely related to viruses that are serious issues in the commercial poultry industry where they cause immunosuppression leading to devastating secondary infections. Finally, Eric Haas-Stapleton, at the Alameda County Mosquito Abatement District, demonstrated for us the low-tech bottle-assay used to assay mosquitoes’ resistance to insecticides, and had great vision for how to use high-tech methods for tracking, providing practical guidance to our imaginative proposal to use the insect-specific viruses carried normally by the mosquito to track populations and interactions (Chapter 8).

The people who have contributed directly to the science are acknowledged in each chapter. I expect that innumerable others, not mentioned here, will continue to challenge my science and point out the odd and the overlooked nooks of their biological surroundings.

No single theme clearly wins as an organizing principle for the chapters that follow. The projects are rather like the chambers of an anthill. Connecting corridors allow passage between chambers that may share a purpose, or inhabitants, or physical dimensions, or date of creation. Together, the chambers and corridors build a structure

that sustains an active and evolving colony. And knowing that the “completion” of one scientific project always begets many more, like the underground nest, there is always room for indefinite expansion.

References for Introduction

- Banatvala, J., and Brown, D. (2004). Rubella. *The Lancet* 363, 1127–1137.
- Doan, T., Wilson, M.R., Crawford, E.D., Chow, E.D., Khan, L.M., Knopp, K.A., O'Donovan, B.D., Xia, D., Hacker, J.K., Stewart, J.M., et al. (2016). Illuminating uveitis: metagenomic deep sequencing identifies common and rare pathogens. *Genome Med* 8, 90.
- Grant, G.B., Reef, S.E., Patel, M., Knapp, J.K., and Dabbagh, A. (2017). Progress in Rubella and Congenital Rubella Syndrome Control and Elimination — Worldwide, 2000 – 2016. *MMWR Morb. Mortal. Wkly. Rep.* 66, 1256–1260.
- Gregg, N.M. (1941). Congenital Cataract Following German Measles in the Mother. In *Problems of Birth Defects*, T.V.N. Persaud, ed. (Dordrecht: Springer Netherlands), pp. 170–180.
- de Groot-Mijnes, J.D.F., de Visser, L., Rothova, A., Schuller, M., van Loon, A.M., and Weersink, A.J.L. (2006). Rubella Virus Is Associated With Fuchs Heterochromic Iridocyclitis. *American Journal of Ophthalmology* 141, 212-214.e1.
- Hutchinson, C., Elbers, J., Halliday, W., Branson, H., Laughlin, S., Armstrong, D., Hawkins, C., Westmacott, R., and Benseler, S.M. (2010). Treatment of small vessel primary CNS vasculitis in children: an open-label cohort study. *The Lancet Neurology* 9, 1078–1084.

- Livingston, J., and Crow, Y. (2016). Neurologic Phenotypes Associated with Mutations in TREX1, RNASEH2A, RNASEH2B, RNASEH2C, SAMHD1, ADAR1, and IFIH1: Aicardi–Goutières Syndrome and Beyond. *Neuropediatrics* 47, 355–360.
- Petridou, C., Bonsall, D., Ahmed, A., Roberts, M., Bell, C., de Cesare, M., Bowden, R., Graham, V., Bailey, D., Simpson, A., et al. (2019). Prolonged Zika Virus RNA Detection in Semen of Immunosuppressed Patient. *Emerg. Infect. Dis.* 25, 1598–1600.
- Salvarani, C., Brown, R.D., Calamia, K.T., Christianson, T.J.H., Weigand, S.D., Miller, D.V., Giannini, C., Meschia, J.F., Huston, J., and Hunder, G.G. (2007). Primary central nervous system vasculitis: analysis of 101 patients. *Ann Neurol.* 62, 442–451.
- Santos, J., and Meneses, B.M. (2017). An integrated approach for the assessment of the *Aedes aegypti* and *Aedes albopictus* global spatial distribution, and determination of the zones susceptible to the development of Zika virus. *Acta Tropica* 168, 80–90.
- Suzuki, J., Goto, H., Komase, K., Abo, H., Fujii, K., Otsuki, N., and Okamoto, K. (2010). Rubella virus as a possible etiological agent of Fuchs heterochromic iridocyclitis. *Graefes Arch Clin Exp Ophthalmol* 248, 1487–1491.
- Varkey, J.B., Shantha, J.G., Crozier, I., Kraft, C.S., Lyon, G.M., Mehta, A.K., Kumar, G., Smith, J.R., Kainulainen, M.H., Whitmer, S., et al. (2015). Persistence of Ebola Virus in Ocular Fluid during Convalescence. *N Engl J Med* 372, 2423–2427.
- WHO (2019). Zika Epidemiology Update (World Health Organization).

Chapter 1 Zika Virus in the Developing Human Brain: Cell

Tropism and Drug Inhibition

Authors:

Hanna Retallack^{1,*}, Elizabeth Di Lullo^{2,3,*}, Carolina Arias^{1,4}, Kristeene A. Knopp¹, Matthew T. Laurie¹, Carmen Sandoval-Espinosa^{2,3}, Walter R. Mancía Leon^{2,3}, Robert Krencik^{5,6}, Erik M. Ullian⁵, Julien Spatazza^{2,7}, Alex A. Pollen^{2,3}, Caleigh Mandel-Brehm¹, Tomasz J. Nowakowski^{2,3}, Arnold R. Kriegstein^{1,2,†}, & Joseph L. DeRisi^{1,8,†}

Affiliations:

· Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, CA 94158, USA.
Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California, San Francisco, San Francisco, CA 94143, USA.

· Department of Neurology, University of California, San Francisco, San Francisco, CA 94158, USA.

· Department of Molecular, Cellular, and Developmental Biology, University of California Santa Barbara, Santa Barbara, CA 93106, USA.

· Department of Ophthalmology, University of California San Francisco, San Francisco, CA 94122, USA.

· Center for Neuroregeneration, Department of Neurosurgery, Houston Methodist Research Institute, Houston, TX, 77030, USA.

· Department of Neurological Surgery, University of California, San Francisco, San Francisco, California 94143, USA.

· Howard Hughes Medical Institute, Chevy Chase, MD, USA.

*These authors contributed equally to this work.

†Corresponding authors. Email: J.L.D. (Joe@derisilab.ucsf.edu) and A.R.K. (kriegsteina@stemcell.ucsf.edu)

Material previously published in:

Retallack H*, Di Lullo E*, Arias C, et al. Zika virus cell tropism in the developing human brain and inhibition by azithromycin. *Proc Natl Acad Sci U S A*. 2016;113(50):14408-14413. doi:10.1073/pnas.1618029113

Abstract:

The rapid spread of Zika virus (ZIKV) and its association with abnormal brain development constitute a global health emergency. Congenital ZIKV infection produces a range of mild to severe pathologies, including microcephaly. To understand the pathophysiology of ZIKV infection, we used models of developing brain that faithfully recapitulate the tissue architecture in early- to mid-gestation. We identify the brain cell populations that are most susceptible to ZIKV infection in primary human tissue, provide evidence for a mechanism of viral entry, and show that a commonly used antibiotic protects cultured brain cells by reducing viral proliferation. In the brain, ZIKV preferentially infected neural stem cells, astrocytes, oligodendrocyte progenitor cells, and microglia, whereas neurons were less susceptible to infection. These findings suggest mechanisms for microcephaly and other pathologic features of infants with congenital ZIKV infection that are not explained by neural stem cell infection alone, such as calcifications in the cortical plate. Furthermore, we find that blocking the glia-enriched putative viral entry receptor, AXL, reduced ZIKV infection of astrocytes *in vitro* and genetic knockdown of AXL in a glial cell line nearly abolished infection. Finally, we evaluate 2,177 compounds, focusing on drugs safe in pregnancy. We show the macrolide antibiotic, azithromycin, reduced viral proliferation and viral-induced cytopathic effects in glial cell lines and human astrocytes. Our characterization of infection in developing human brain clarifies the pathogenesis of congenital ZIKV infection and provides the basis

for investigating possible therapeutic strategies to safely alleviate or prevent the most severe consequences of the epidemic.

Significance statement:

Zika virus is a mosquito-borne flavivirus that has rapidly spread through the Americas and has been associated with fetal abnormalities, including microcephaly. To understand how microcephaly develops, it is important to identify which cell types of the developing brain are susceptible to infection. We use primary human tissue to show that radial glia and astrocytes are more susceptible to infection than neurons, a pattern that correlates with expression of a putative viral entry receptor, AXL. We also perform a screen of FDA-approved compounds, with an emphasis on drugs known to be safe in pregnancy. We identify an antibiotic, azithromycin, that reduces viral proliferation in glial cells, and compare its activity to daptomycin and sofosbuvir, two additional drugs with anti-ZIKV activity.

Main Text:

A correlation between congenital exposure to the mosquito-borne and sexually transmitted Zika flavivirus (ZIKV) and the increased incidence of severe microcephaly suggests a causal relationship between ZIKV infection and neurodevelopmental abnormalities (Patricia Brasil et al. 2016; Mlakar et al. 2016). Yet the mechanisms of infection and specifically which cell populations are vulnerable to ZIKV during the course of human brain development remain unclear. Major insights have been drawn from *in vitro* models of human brain development and primary mouse tissues. In the developing

mouse brain, ZIKV has been shown to infect radial glia and neurons (C. Li et al. 2016), while studies in human pluripotent stem cell (hPSC)-derived neural cells highlighted widespread infection and apoptosis of neural progenitor cells (Tang et al. 2016; Qian et al. 2016). Because these models do not fully recapitulate the developmental events and cell types present during human brain development, these results may not faithfully represent ZIKV-induced pathology *in vivo*.

During human brain development, radial glia cells, the neural stem cells, give rise to diverse types of neuronal and glial cells including neurons, oligodendrocytes, and astrocytes, in a temporally controlled pattern. We reasoned that identifying cell types that are especially vulnerable to viral infection would facilitate studies of the viral lifecycle including entry mechanisms and host cell requirements. Building on studies that suggested enriched expression of the candidate entry factor AXL could confer vulnerability to ZIKV entry (Hamel et al. 2015; Onorati et al. 2016; S. Liu et al. 2016), we used AXL expression levels to predict that radial glia, astrocytes, microglia, and endothelial cells would be particularly vulnerable to infection (Nowakowski et al. 2016). A recent study highlighted the utility of *ex vivo* models using primary human tissue samples to analyze the consequences of ZIKV infection in human prenatal brain (Onorati et al. 2016). Here we further utilize primary tissue samples from distinct stages of brain development, corresponding to periods of peak neurogenesis and early gliogenesis.

Determining the tropism of ZIKV for specific cell types will help identify suitable cellular models for investigating potential therapeutic interventions. While development of a vaccine could provide a long-term solution to the current ZIKV epidemic, there remains

an unmet clinical need to identify drugs that can limit or prevent the consequences of congenital infection. A recent screen of a subset of FDA-approved compounds against ZIKV in hepatic cells identified several anti-cancer, anti-microbial, anti-parasitic, and anti-fungal drugs with anti-ZIKV activity (Barrows et al. 2016). Another screen based on human neural progenitor cells identified an anti-fungal drug and several scaffold compounds for further development (Xu et al. 2016). However, the majority of compounds with anti-ZIKV activity from these screens are contraindicated or of unknown safety during pregnancy. Furthermore, two promising candidates that might be safe during pregnancy, daptomycin and sofosbuvir, showed variable effectiveness by cell type (Onorati et al. 2016; Barrows et al. 2016; Sacramento et al. 2016). Combining unbiased screens of approved compounds with comparisons of top candidates with known antiviral activity may quickly narrow the search for drugs that could mitigate the effects of congenital ZIKV infection.

Here, we assessed ZIKV cell tropism in the developing human brain and performed a drug screen on relevant cell types targeted by the virus with an emphasis on drugs known to be safe in pregnancy. We found that radial glia, and later in development, astrocytes, were especially vulnerable to ZIKV infection. By screening FDA-approved compounds for anti-ZIKV activity in a glial cell line with features of both cell types, we also found that the common antibiotic azithromycin prevented viral production and virus-mediated cell death, which we further validated in human astrocytes.

Results

To determine the cell populations most susceptible to ZIKV infection we investigated the infectivity of ZIKV in the developing human brain using organotypic cultures from primary human tissue. We exposed human cortical tissue slices to three strains of ZIKV: Cambodia 2010 (ZIKV-CAM), Brazil 2015 (ZIKV-BR), and Puerto Rico 2015 (ZIKV-PR), cultured them for 72 hours (h), and detected infection by immunostaining for the flavivirus envelope protein (ENV), an approach that we validated in cultured cells (Fig. S1). Infection in tissue was confirmed by immunostaining for the viral RNA-dependent RNA polymerase, nonstructural protein 5 (NS5), present only during viral replication. In samples from mid-neurogenesis (13-16 post-conception weeks (pcw)), we observed high rates of infection in the ventricular and subventricular zones (Figure 1.1, and Fig. S2). We found that the virus preferentially infected both ventricular (vRG) and outer radial glia (oRG) cells (Figure 1.1, A to F, and Fig. S2). Interestingly, we observed clusters of infected radial glia (Fig. S2B), which may reflect local viral spread. A minor fraction of cells positive for ENV at these stages included postmitotic neurons (Fig. 1H), and microglia (Figure 1.1 I). We observed similar patterns of infection across ZIKV strains (Fig. S2). We also observed a small but significant increase in cell death of ENV+ cells as compared to ENV- cells in ZIKV-infected tissue or mock-infected tissue (Fig. S3).

At later stages of development (after 17 pcw), we observed infection and viral replication throughout the developing cortex, including the cortical plate and subplate, with production of infectious virus by 48 hpi (Figure 1.2, and Fig. S4). Among cortical

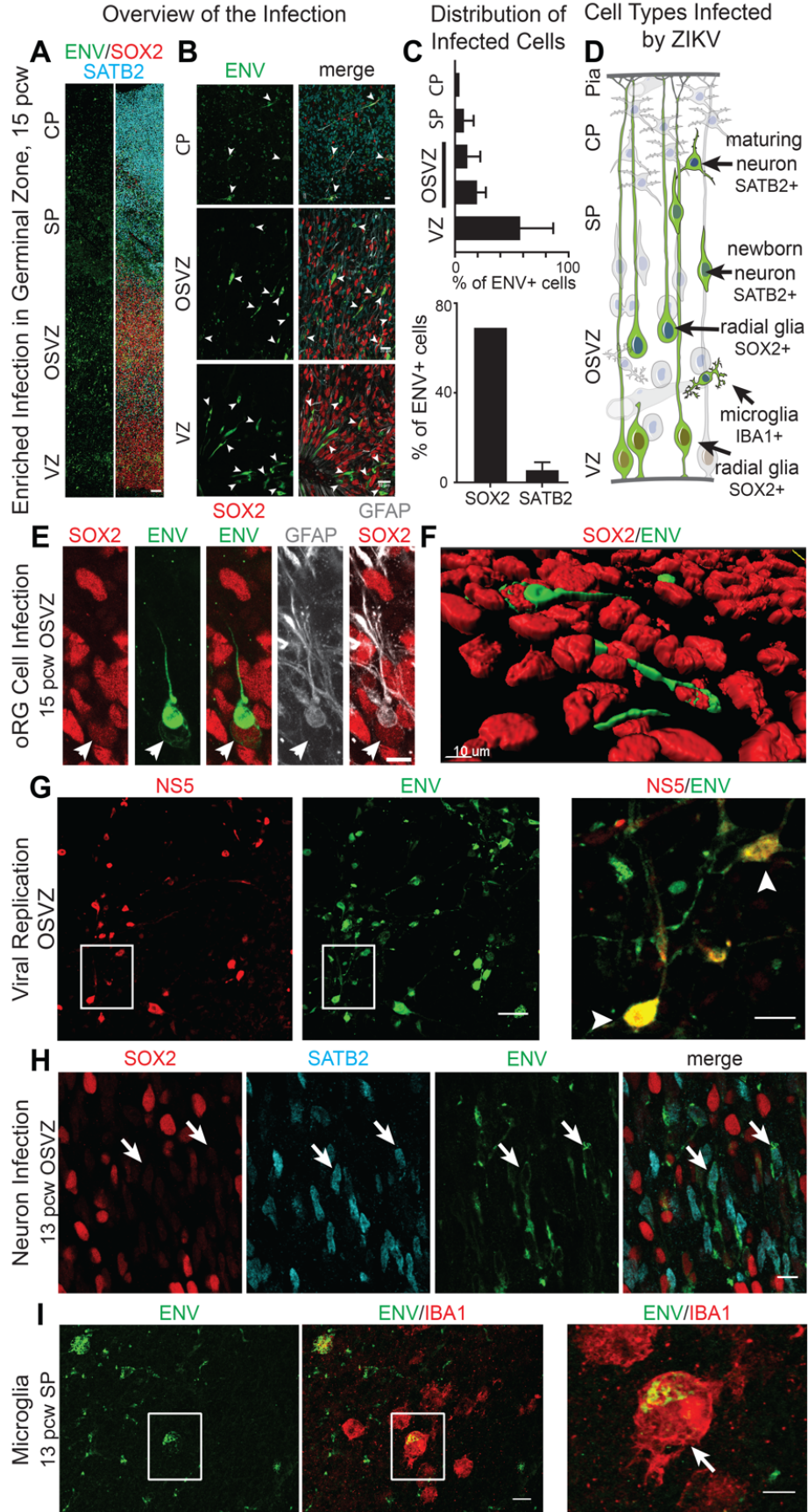


Figure 1.1 Tropism of ZIKV for radial glia in the developing human brain.

Human cortical organotypic brain slices were infected with ZIKV-BR and cultured for 72 h. (A to B) Low magnification overview of ZIKV infection detected by ENV (green) within the cortex. (A) ENV staining was analyzed with respect to region and cell type. Scale bar 100 μ m. (B) High magnification of (A). Notably, ENV staining (arrowheads) appears to preferentially enriched in the ventricular zone (VZ) and outer subventricular zone (OSVZ). SP - subplate, CP - cortical plate, pcw - post-conception weeks. Scale bar 20 μ m. (C) Quantification of ENV positive cells by region (top) and cell type (bottom) at 13-14 pcw. $n = 2$. mean \pm SD (see Methods, error bar not shown where shorter than line thickness (top panel, CP, and bottom panel, SOX2)). (D) Schematic summary of cell types observed to be susceptible to ZIKV infection (green) in the developing human brain during mid-neurogenesis. (E) High magnification overview of a ZIKV-infected radial glial cell in the OSVZ. Scale bar 10 μ m. (F) Three dimensional reconstruction of (E), highlighting intracellular presence of ENV signal. Scale bar 10 μ m. (G) ENV and NS5 signal in OSVZ cells (arrowheads) suggested replicating ZIKV-PR. Scale bar 20 μ m. (H) Immature neurons (SATB2+, blue) infected with ZIKV (arrows). Scale bar 20 μ m. (I) Microglia (IBA1+) immunopositive for ENV. High magnification panel (right) shows ENV+ microglia with amoeboid morphology (arrow), typical of activated microglia. Scale bar 10 μ m.

plate cells, we observed a high rate of infection in astrocytes as distinguished by their location, morphology, and immunoreactivity with the glial markers GFAP and SOX2 (Figure 1.2, A, B and D, and Fig. S4, A, B, C, and D). We also observed cells immunoreactive for both ENV and the microglial marker IBA1, indicating microglial infection or phagocytosis of other ZIKV infected cells (Figure 1.2, and Fig. S4 G and H). This ENV+/IBA1+ microglial population was quantified at $7\pm 1\%$ of ENV+ cells, and represented $7\pm 2\%$ of the total IBA1+ population ($n = 4$, 15-22 pcw, see Methods). We further observed infection of oligodendrocyte precursor cells (OPCs) (Figure 1.2G and Fig. S4I), but limited infection of neurons (Figure 1.2, B and D, and Fig. S4, A and J). This pattern of infectivity was consistent across ZIKV strains (Fig. S4) and matched viral tropism predicted by AXL receptor expression (Nowakowski et al. 2016).

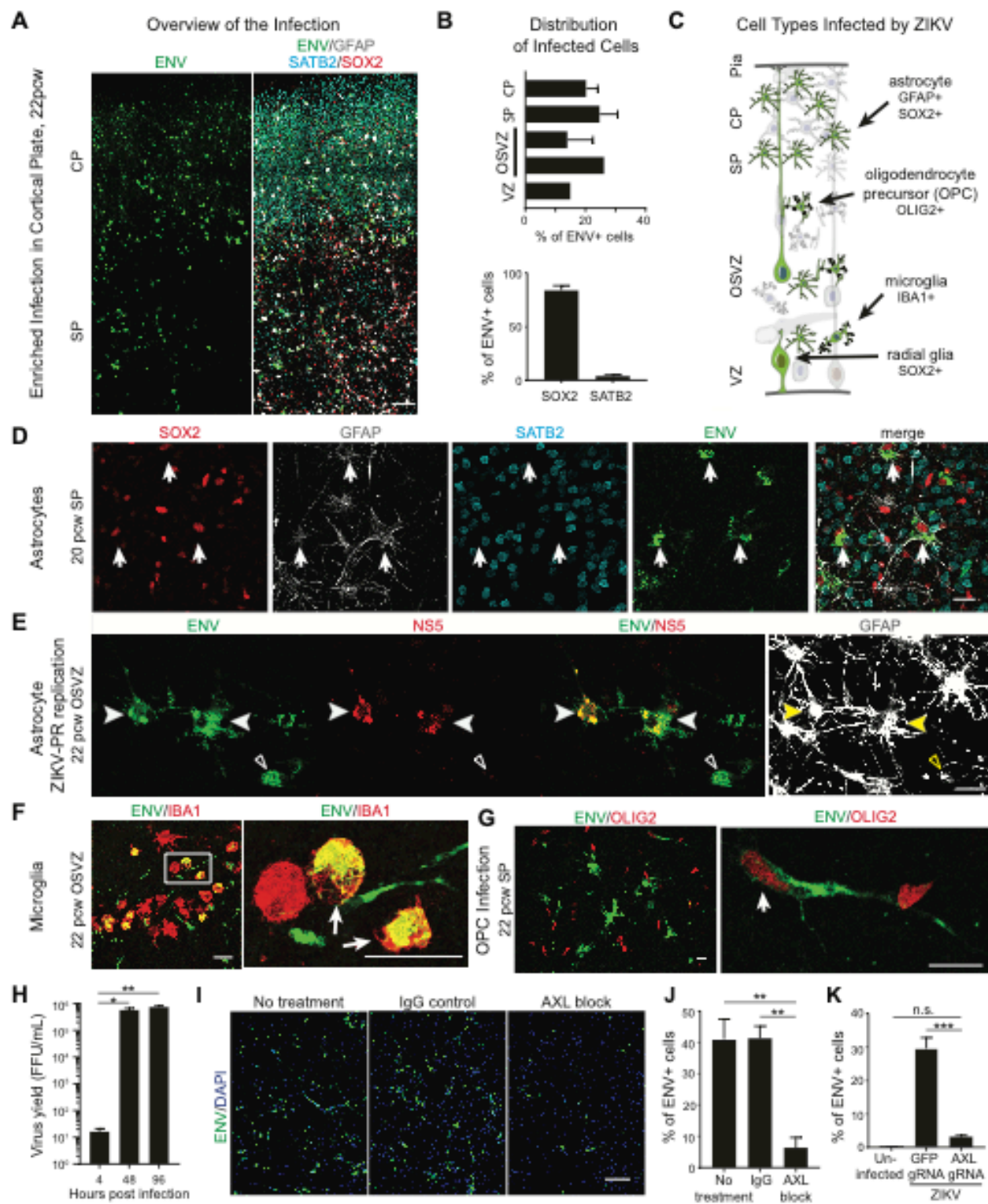


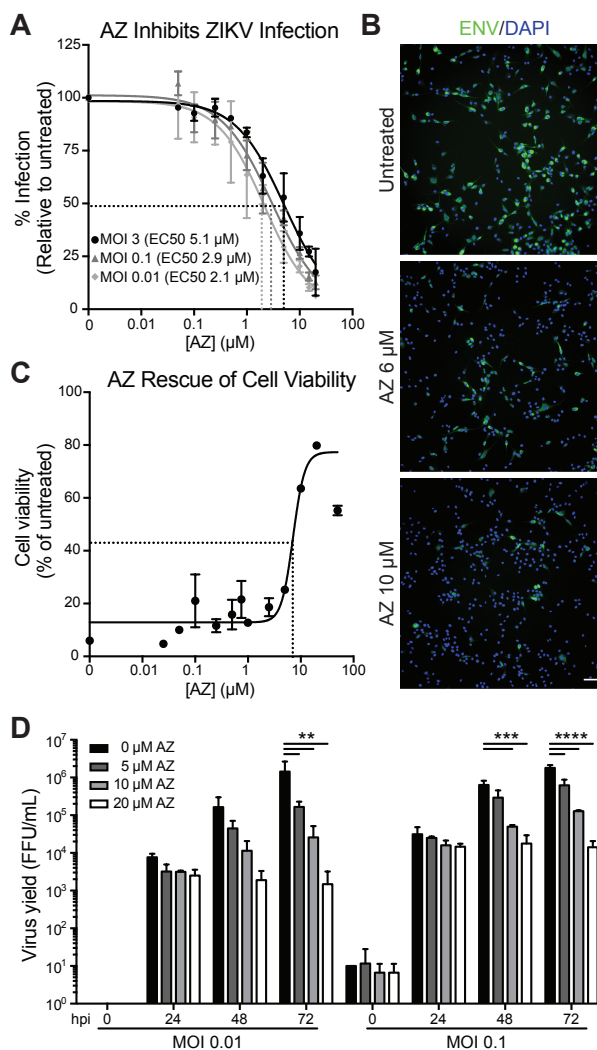
Figure 1.2 ZIKV infects astrocytes in later stages of human brain development.

(A) Low magnification overview of ZIKV infection detected by ENV (green) within human organotypic cortical slices during late neurogenesis/gliogenesis. Scale bar 100 μ m. (B) Quantification of ENV positive cells by region (top) and cell type (bottom) at 20-22 pcw. $n = 2$. mean \pm SD (see Methods, error bar not shown where shorter than line thickness (top panel, VZ and lower OSVZ)). (C) Schematic summary of cell types observed to be susceptible to ZIKV infection (green). (D to E) Immunohistochemical analysis reveals ZIKV infection in astrocytes by positivity for ENV (arrows, D), or ENV and the non-structural protein NS5 indicating active viral replication (filled arrowheads, E). Scale bars 20 μ m. (F) Microglia co-labeled with ENV. Scale bar 50 μ m. (G) ZIKV infection of oligodendrocyte precursor cells (OPCs). Scale bar 20 μ m. (H) Viral production in 19 pcw cortical slices, quantified by focus forming assay from combined homogenized tissue and conditioned media at 4, 48, and 96 hours post infection; mean \pm SEM, 2 independent biological replicates with 2 technical replicates for each timepoint; one-way ANOVA with Tukey's multiple comparisons test, * $p \leq 0.05$, ** $p \leq 0.01$; see also Fig. S4F. (I to J) Analysis of ZIKV-BR infection in the presence of AXL blocking antibody in hPSC-derived astrocytes (see Methods). Note reduced ENV staining with AXL block compared to IgG control. Scale bar 100 μ m. (J) Quantification of experiment represented in (I); see also Fig. S5A; mean \pm SEM, $n = 3$, one-way ANOVA with Tukey's multiple comparisons test, ** $p \leq 0.01$. (K) ZIKV-PR infection after knockdown of AXL using U87-dCas9 lines expressing either GFP gRNA (non-targeting control) or AXL gRNAs (dCas9-mediated knockdown); see also Fig. S5B; mean \pm SEM, 2 biological replicates in cell lines generated with independent transductions; two-way ANOVA with Tukey's multiple comparisons test, n.s. - not significant, *** $p \leq 0.001$.

To test the possible role of AXL in mediating ZIKV entry into human astrocytes, we infected hPSC-derived astrocytes (Krencik et al. 2015; 2011) in the presence of a non-activating antibody specific for the extracellular domain of AXL. Blocking the AXL receptor substantially reduced infection (Figure 1.2, I and J, and Fig. S5A). To further test the requirement of AXL for ZIKV infection of glial cells, we used the U87 glioblastoma line that expresses high levels of astrocyte marker genes and AXL (S. J. Liu et al. 2016). U87 cells were readily infected with ZIKV, with strong virus production at 48 hpi (Fig. S1) and

Figure 1.3 Azithromycin (AZ) treatment inhibits ZIKV infection in glial cells.

(A) U87 cells were treated with increasing concentrations of AZ and infected with ZIKV-PR at varying MOIs (0.01, 0.1 and 3, as indicated). The percent of infected cells at 48 hpi was determined by flow cytometry of cells immunostained for ENV, and normalized to untreated cells (for raw data see Fig. S6A). EC50 values for AZ-mediated reduction of ZIKV infection were 5.1 μM for an MOI of 3 ($n = 2$), 2.9 μM for an MOI of 0.1 ($n = 2$) and 2.1 μM for an MOI of 0.01 ($n = 2$); mean \pm SD. (B) Representative images of U87 cells treated with AZ and infected with ZIKV-PR at an MOI of 3 (as in A). At 48 hpi cells were immunostained for ENV protein (green) and cellular DNA (DAPI, blue). Scale bar 100 μm . (C) Rescue of cell viability with AZ. U87 cells were pre-treated with AZ for 1 h and then infected with ZIKV-PR at an MOI of 10 in the presence of AZ. Cell viability was measured at 72 hpi using the CellTiter-Glo luminescence assay. The EC50 value for the AZ-mediated rescue of cell viability was 7.1 μM . The data point at the highest concentration of AZ (50 μM) showed reduced cell viability, likely due to drug toxicity (see Fig. S6C). $n = 2$, mean \pm SD. (D) Decrease of virus production with AZ treatment. U87 cells were pre-treated with AZ for 1 h and then infected with ZIKV-PR at an MOI of 0.1 or 0.01 in the presence of AZ. Quantification of virus yield in conditioned media was performed by focus forming assay at 0, 24, 48, and 72 hpi; mean \pm SD with $n = 2$ for each MOI, two-way ANOVA with Tukey's multiple comparisons testing, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$.



robust cytopathic effect at 72 hpi (Figure 1.3C and Fig. S6D). We then used CRISPRi to knock down AXL in this cell line (see Materials and Methods, validated by Western blot in Fig. S5B), and observed a substantial decrease in infection (Figure 1.2K), confirming the importance of this receptor for ZIKV infection in this cell type.

Given that AXL is a receptor tyrosine kinase with signaling pathways that could be involved in innate immune responses (Rothlin et al. 2007), we tested whether the kinase activity of AXL was relevant for the decrease in infection observed in the knockdown line. After pre-treatment with a small molecule inhibitor, R428, we observed no decrease in infection at up to 1 μM , which is >70 fold the half-maximal effective concentration (EC50) for AXL kinase inhibition (Fig. S5C) (Holland et al. 2010). Although we did observe a decrease in infection at 3 μM R428, this high concentration of >200 fold the EC50 likely created off-target effects. Together, these results suggest that AXL has an important role in glial cell infection that depends more on its extracellular domain than on its intracellular kinase activity.

There is a pressing need to identify pharmacological compounds that can diminish the effects of ZIKV infection in relevant human cell types. We performed a screen of 2177 clinically approved compounds (2016 unique) by monitoring inhibition of virus-dependent cell death at 72 hpi in Vero cells. While our screen revealed compounds that rescued cell viability, including antibiotics and inhibitors of nucleotide and protein synthesis, many showed toxicity in Vero or U87 cells, or are contraindicated during pregnancy (Tables S1, S2, S3, S4). We focused on further characterization of the macrolide antibiotic azithromycin (AZ), which rescued ZIKV-induced cytopathic effect with low toxicity in our primary screens, and is generally safe during pregnancy (Lin et al. 2013). AZ dramatically reduced ZIKV infection of U87 cells at an EC50 of 2-3 μM at multiplicities of infection (MOIs) of 0.01-0.1, as evaluated by ENV staining (Figure 1.3, A and B, and Fig. S6A). We further established a relationship between EC50 and baseline

infection rate (Fig. S6B), and showed that even at >60% infection, AZ consistently reduced infection at concentrations ten- to twenty-fold below the half-maximal toxicity concentration (TC50) of 53 μM (Fig. S6, A and C). AZ treatment also rescued cell viability (Figure 1.3C, Fig. S6D) and decreased viral production (Figure 1.3D). Finally, we found that AZ substantially reduced infection in hPSC-derived astrocytes without toxicity at the effective concentration (EC50 15 μM at 72% baseline infection) (Fig. S6, E to G). To compare AZ with compounds identified in previous screens, we evaluated the anti-ZIKV activity of daptomycin and sofosbuvir in U87 cells (EC50 2.2 μM and 12.4 μM respectively) (Fig. S6H). We observed that treatment with daptomycin was insufficient to lower the percentage of infected cells below 46% even at the highest dose in this cell type (20 μM) (Fig. S6H), whereas AZ and sofosbuvir treatment decreased ZIKV infection from 78% to below 5% infection at 20 μM and 50 μM respectively. These results highlight AZ as a potential tool compound against ZIKV infection in glial cells.

Discussion

The rapid spread of ZIKV and its link to fetal abnormalities, including microcephaly, have created a global health crisis. Understanding viral tropism for specific cell types in the developing brain furthers our understanding of the pathophysiology of ZIKV-associated microcephaly, and provides a basis for investigating antiviral drugs in a relevant cell type. Our findings offer several novel aspects. In particular, we show ZIKV tropism for astrocytes in addition to radial glia in primary developing human brain, demonstrate the importance of AXL for ZIKV infection of glial cells, and identify a

common antibiotic with anti-ZIKV activity, AZ, which we compare to two other drugs with anti-ZIKV activity that may be safe in pregnancy.

Our finding that radial glia are preferentially infected during early neurogenesis is consistent with experiments in cultured primary human brain cells (Hanners et al. 2016), developing mouse cortex (C. Li et al. 2016; Brault et al. 2016), and primary human organotypic brain slice culture (Onorati et al. 2016). These studies also reported overall survival of infected radial glia, in contrast to *in vitro* derived neural stem cells that undergo apoptotic cell death following infection (Tang et al. 2016; Qian et al. 2016; Cugola et al. 2016). Cell lines derived from primary neural progenitors have variably shown infection with substantial apoptosis (Onorati et al. 2016) or persistence (Hanners et al. 2016). In our organotypic slice culture we observe a small increase in apoptosis of infected cells. The discrepancy in levels of apoptosis in dissociated versus tissue cell culture may reflect differences in gene expression, maturation, or experimental conditions. Besides causing cell death, ZIKV infection could also affect cell cycle progression (C. Li et al. 2016; Cugola et al. 2016), differentiation, or the migration and survival of newborn neurons – mechanisms thought to underlie genetic causes of microcephaly and lissencephaly (Thornton and Woods 2009). Tissue disorganization in organotypic slice culture suggests these non-cell death-mediated mechanisms may contribute to clinical phenotypes (Onorati et al. 2016), but this remains to be confirmed by directly analyzing cell behavior.

The high rate of infection in astrocytes at later developmental ages, many of which contact microcapillaries, could link our understanding of initial infection with clinical findings of cortical plate damage. For example, after prolonged infection, viral production

in astrocytes could lead to a higher viral load in the cortical plate causing infection of additional cortical cell types, and astrocyte loss could lead to inflammation and further damage, even in uninfected cells. Widespread cell death *in vivo*, which may take days to weeks to occur and is therefore outside the time frame of our experimental paradigm, is expected given clinical reports of band-like calcifications in the cortical plate, cortical thinning, and hydrocephalus (Mlakar et al. 2016; Hazin et al. 2016). Based on their susceptibility to ZIKV infection and a central role in brain tissue homeostasis, human astrocytes provide a good cellular model for further investigation of mechanisms of viral entry and a platform for testing the efficacy of candidate therapeutic compounds.

Our observation that blocking or knocking down the AXL receptor prevents infection of human astrocytes, but that blocking intracellular kinase activity does not, suggests that the extracellular domain of AXL contributes to ZIKV infection whereas AXL signaling is dispensable. This extends comparable findings in endothelial cells to a cell type relevant for understanding microcephaly (Hamel et al. 2015; S. Liu et al. 2016), but does not address other viral receptors that may be important for ZIKV infectivity in other cell types, or rule out a role for AXL signaling in the context of a full immune response *in vivo*. While AXL knockout mice can be readily infected with ZIKV, disruption of the blood brain barrier in these mice could lead to atypical routes for infection of the brain (Miner et al. 2016).

In addition to characterizing brain cell tropism, we also sought to identify possible therapeutic candidates with known safety profiles, especially in pregnancy. Several compounds expected to inhibit ZIKV were identified by our drug screen. These positive

controls include the protein synthesis inhibitor cycloheximide, nucleic acid synthesis inhibitors such as mycophenolate derivatives, and intercalating compounds like doxorubicin and homidium bromide. We additionally identified compounds that are known to be safe in pregnancy, including azithromycin (AZ). AZ is recommended for the treatment of pregnant women with sexually transmitted infections or respiratory infections due to AZ-susceptible bacteria (Workowski and Bolan 2015; DHHS Panel on Opportunistic Infections in HIV-Infected Adults and Adolescents 2014). Adverse events have not been observed in animal reproduction studies, and studies in pregnant women show no negative effects on pregnancy outcome or fetal health associated with AZ (Sarkar et al. 2006). Orally administered AZ has been shown to reach concentrations of $\sim 2.8 \mu\text{M}$ in the placenta, and is rapidly transported to amniotic fluid and umbilical cord plasma in humans (Ramsey et al. 2003; Sutton et al. 2015). Moreover, AZ accumulates in fetal tissue and in the adult human brain at concentrations from $4\text{-}21 \mu\text{M}$ (Jaruratanasirikul et al. 1996; Kemp et al. 2014). Together, these pharmacokinetic studies suggest that AZ could rapidly accumulate in fetal tissue including the placenta *in vivo* at concentrations comparable to those that inhibit ZIKV proliferation in culture. Nonetheless, it remains unknown whether these *in vitro* results would be recapitulated in humans.

We further compared AZ with two promising drug candidates that might be safe in pregnancy and have reported anti-ZIKV activity in cell culture: daptomycin and sofosbuvir. Our dose-response curves are in agreement with the documented activity of sofosbuvir in human neuroepithelial stem cells (Onorati et al. 2016), and extend the activity of daptomycin previously seen in HuH-7 and HeLa cells (Barrows et al. 2016) to

glial cells. We noted that daptomycin would not have been highly ranked in our initial screen due to the limited maximum effect of the drug as observed in dose-response curves. Unlike sofosbuvir, which likely targets the ZIKV RNA-dependent RNA polymerase (NS5) based on its mechanism against Hepatitis C virus, daptomycin and AZ have unknown mechanisms of action against ZIKV. Nonetheless, the difference in *in vitro* dose response between AZ and daptomycin is intriguing, and suggests different mechanisms of inhibition. Another important factor for a drug candidate for ZIKV treatment is accessibility. Access to sofosbuvir and its derivatives may be limited by its current price, whereas AZ and daptomycin are available as generic forms, although daptomycin is not available in oral formulation due to poor oral bioavailability. Our comparison adds new data to consider alongside other antiviral activity data, safety, cost, and accessibility in moving forward with further exploration of these and related compounds. In parallel with direct comparisons *in vitro*, follow-up studies in animal models can be useful for prioritizing candidates. However, as with *in vitro* studies, there are caveats in interpreting animal models, such as substantial differences between human and mouse immune systems, placental structure, and fetal brain development.

Together, our work identifies cell type specific patterns of ZIKV infection in second trimester human developing brain, provides experimental evidence that AXL is important for ZIKV infection of relevant human brain cell types, and highlights a common antibiotic with inhibitory activity against ZIKV in glial cells. Ongoing studies will be required to determine whether AZ, daptomycin, sofosbuvir, and other inhibitors or combinations are capable of reducing ZIKV infection in the critical cell types identified here *in vivo*. While

preventative measures such as mosquito abatement and a ZIKV vaccine are imperative for long-term control of this pathogen, the study of ZIKV infection of primary human tissues and identification of inhibitors with therapeutic potential remain important components of a global response to this emerging threat.

Materials and Methods

Detailed Materials and Methods are available in SI Materials and Methods.

Cells and Viruses

Cell lines were Vero cells, U87 cells, and human astrocytes derived from human pluripotent stem cells as previously described (Krencik et al. 2015). ZIKV strains were SPH2015 (Brazil 2015, ZIKV-BR), PRVABC59 (Puerto Rico 2015, ZIKV-PR), and FSS13025 (Cambodia 2010, ZIKV-CAM). Focus forming assays and virus yield assays described in SI Materials and Methods.

Brain Samples

De-identified primary tissue samples were collected with previous patient consent in strict observance of the legal and institutional ethical regulations. Protocols were approved by the Human Gamete, Embryo and Stem Cell Research Committee (institutional review board) at UCSF. Organotypic slice culture was performed as described in SI Materials and Methods. Slices were inoculated with ZIKV or mock infected, fixed at 72 hpi or 5 dpi and processed for immunohistochemistry. Quantification was performed using Imaris v. 8.2 (Bitplane) on slices from 13-22 pcw.

AXL

For 1 h prior to infection with ZIKV, cells were treated with AXL blocking antibody (R&D Systems AF154) or goat IgG control (R&D Systems AB108C) at 100 µg/mL, or for AXL kinase inhibition, with 1-3 µM R428 (APExBIO A8329) or vehicle (<0.1% DMSO). For CRISPRi-mediated AXL knockdown, U87 cells stably expressing dCas9-KRAB (S. J. Liu et al. 2016) were transduced with lentiviral particles expressing a pool of sgRNAs targeting AXL or an sgRNA targeting GFP as a control (see SI Materials and Methods).

Drug Screen

A set of 2177 FDA-approved compounds (2016 unique) from the UCSF Small Molecule Discovery Center were tested at a final concentration of 2 µM in Vero cells infected with ZIKV-BR at MOIs of 1, 3, and 10, and in U87 cells at an MOI of 3. Toxicity screens in uninfected Vero and U87 cells were performed in parallel. In each screen, cells were treated with compounds 2 h before addition of ZIKV-BR or media and incubated for 72 h, at which point cell viability was assessed using the CellTiter-Glo 2.0 luminescent assay (Promega). Candidates demonstrating cell viability >2.5 fold that of untreated cells in every Vero cell screen were identified, and compounds of interest as positive controls or for potential use in pregnancy were further validated in U87 cells using ZIKV-PR (see below).

Drug Validation

U87 cells or hPSC-derived astrocytes were treated with azithromycin (Selleck S1835, Sigma 75199 (analytical standard), Sigma PZ0007 (HPLC-purified)), or the clinical preparation Zithromax (Apotex Corp NDC 60505-6076-4), daptomycin (Sigma D2446,

Selleck Chemical S1373), sofosbuvir (Selleck Chemical S2794), or vehicle (PBS, ethanol or DMSO final concentration <0.1%) for at least 1 h, then infected with ZIKV-PR. Cell viability assays were performed using CellTiter-Glo as above. To assess viral envelope production, cells were fixed and stained at 48 hpi using anti-flavivirus envelope (clone D1-4G2-4-15), and then quantified either by plate imaging with automated cell counting in Imaris v. 8.2 (Bitplane), or by flow cytometry.

References for Chapter 1

- Barrows, N.J., Campos, R.K., Powell, S.T., Prasanth, K.R., Schott-Lerner, G., Soto-Acosta, R., Galarza-Muñoz, G., McGrath, E.L., Urrabaz-Garza, R., Gao, J., et al. (2016). A Screen of FDA-Approved Drugs for Inhibitors of Zika Virus Infection. *Cell Host & Microbe* 20, 259–270.
- Brasil, P., Pereira, J.P., Moreira, M.E., Ribeiro Nogueira, R.M., Damasceno, L., Wakimoto, M., Rabello, R.S., Valderramos, S.G., Halai, U.-A., Salles, T.S., et al. (2016). Zika Virus Infection in Pregnant Women in Rio de Janeiro. *N Engl J Med* 375, 2321–2334.
- Brault, J.-B., Khou, C., Basset, J., Coquand, L., Fraissier, V., Frenkiel, M.-P., Goud, B., Manuguerra, J.-C., Pardigon, N., and Baffet, A.D. (2016). Comparative Analysis Between Flaviviruses Reveals Specific Neural Stem Cell Tropism for Zika Virus in the Mouse Developing Neocortex. *EBioMedicine* 10, 71–76.
- Cugola, F.R., Fernandes, I.R., Russo, F.B., Freitas, B.C., Dias, J.L.M., Guimarães, K.P., Benazzato, C., Almeida, N., Pignatari, G.C., Romero, S., et al. (2016). The Brazilian Zika virus strain causes birth defects in experimental models. *Nature* 534, 267–271.
- Hamel, R., Dejarnac, O., Wichit, S., Ekchariyawat, P., Neyret, A., Luplertlop, N., Perera-Lecoin, M., Surasombatpattana, P., Talignani, L., Thomas, F., et al. (2015). Biology of Zika Virus Infection in Human Skin Cells. *J. Virol.* 89, 8880–8896.

- Hanners, N.W., Eitson, J.L., Usui, N., Richardson, R.B., Wexler, E.M., Konopka, G., and Schoggins, J.W. (2016). Western Zika Virus in Human Fetal Neural Progenitors Persists Long Term with Partial Cytopathic and Limited Immunogenic Effects. *Cell Reports* 15, 2315–2322.
- Hazin, A.N., Poretti, A., Di Cavalcanti Souza Cruz, D., Tenorio, M., van der Linden, A., Pena, L.J., Brito, C., Gil, L.H.V., de Barros Miranda-Filho, D., Marques, E.T. de A., et al. (2016). Computed Tomographic Findings in Microcephaly Associated with Zika Virus. *N Engl J Med* 374, 2193–2195.
- Holland, S.J., Pan, A., Franci, C., Hu, Y., Chang, B., Li, W., Duan, M., Torneros, A., Yu, J., Heckrodt, T.J., et al. (2010). R428, a Selective Small Molecule Inhibitor of Axl Kinase, Blocks Tumor Spread and Prolongs Survival in Models of Metastatic Breast Cancer. *Cancer Research* 70, 1544–1554.
- Jaruratanasirikul, S., Hortiwakul, R., Tantisarasart, T., Phuenpathom, N., and Tussanasunthornwong, S. (1996). Distribution of azithromycin into brain tissue, cerebrospinal fluid, and aqueous humor of the eye. *Antimicrob. Agents Chemother.* 40, 825–826.
- Kemp, M.W., Miura, Y., Payne, M.S., Jobe, A.H., Kallapur, S.G., Saito, M., Stock, S.J., Spiller, O.B., Ireland, D.J., Yaegashi, N., et al. (2014). Maternal Intravenous Administration of Azithromycin Results in Significant Fetal Uptake in a Sheep Model of Second Trimester Pregnancy. *Antimicrob. Agents Chemother.* 58, 6581–6591.

- Krencik, R., Weick, J.P., Liu, Y., Zhang, Z.-J., and Zhang, S.-C. (2011). Specification of transplantable astroglial subtypes from human pluripotent stem cells. *Nat Biotechnol* 29, 528–534.
- Krencik, R., Hokanson, K.C., Narayan, A.R., Dvornik, J., Rooney, G.E., Rauen, K.A., Weiss, L.A., Rowitch, D.H., and Ullian, E.M. (2015). Dysregulation of astrocyte extracellular signaling in Costello syndrome. *Sci. Transl. Med.* 7, 286ra66-286ra66.
- Li, C., Xu, D., Ye, Q., Hong, S., Jiang, Y., Liu, X., Zhang, N., Shi, L., Qin, C.-F., and Xu, Z. (2016). Zika Virus Disrupts Neural Progenitor Development and Leads to Microcephaly in Mice. *Cell Stem Cell* 19, 120–126.
- Lin, K.J., Mitchell, A.A., Yau, W.-P., Louik, C., and Hernández-Díaz, S. (2013). Safety of macrolides during pregnancy. *American Journal of Obstetrics and Gynecology* 208, 221.e1-221.e8.
- Liu, S., DeLalio, L.J., Isakson, B.E., and Wang, T.T. (2016a). AXL-Mediated Productive Infection of Human Endothelial Cells by Zika Virus. *Circ Res* 119, 1183–1189.
- Liu, S.J., Nowakowski, T.J., Pollen, A.A., Lui, J.H., Horlbeck, M.A., Attenello, F.J., He, D., Weissman, J.S., Kriegstein, A.R., Diaz, A.A., et al. (2016b). Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol* 17, 67.
- Miner, J.J., Sene, A., Richner, J.M., Smith, A.M., Santeford, A., Ban, N., Weger-Lucarelli, J., Manzella, F., Rückert, C., Govero, J., et al. (2016). Zika Virus Infection in Mice Causes Panuveitis with Shedding of Virus in Tears. *Cell Reports* 16, 3208–3218.

- Mlakar, J., Korva, M., Tul, N., Popović, M., Poljšak-Prijatelj, M., Mraz, J., Kolenc, M., Resman Rus, K., Vesnaver Vipotnik, T., Fabjan Vodusek, V., et al. (2016). Zika Virus Associated with Microcephaly. *N Engl J Med* 374, 951–958.
- National Institutes of Health Guidelines for the Prevention and Treatment of Opportunistic Infections in Adults and Adolescents with HIV.
- Nowakowski, T.J., Pollen, A.A., Di Lullo, E., Sandoval-Espinosa, C., Bershteyn, M., and Kriegstein, A.R. (2016). Expression Analysis Highlights AXL as a Candidate Zika Virus Entry Receptor in Neural Stem Cells. *Cell Stem Cell* 18, 591–596.
- Onorati, M., Li, Z., Liu, F., Sousa, A.M.M., Nakagawa, N., Li, M., Dell’Anno, M.T., Gulden, F.O., Pochareddy, S., Tebbenkamp, A.T.N., et al. (2016). Zika Virus Disrupts Phospho-TBK1 Localization and Mitosis in Human Neuroepithelial Stem Cells and Radial Glia. *Cell Reports* 16, 2576–2592.
- Qian, X., Nguyen, H.N., Song, M.M., Hadiono, C., Ogden, S.C., Hammack, C., Yao, B., Hamersky, G.R., Jacob, F., Zhong, C., et al. (2016). Brain-Region-Specific Organoids Using Mini-bioreactors for Modeling ZIKV Exposure. *Cell* 165, 1238–1254.
- Ramsey, P.S., Vaules, M.B., Vasdev, G.M., Andrews, W.W., and Ramin, K.D. (2003). Maternal and transplacental pharmacokinetics of azithromycin. *American Journal of Obstetrics and Gynecology* 188, 714–718.
- Rothlin, C.V., Ghosh, S., Zuniga, E.I., Oldstone, M.B.A., and Lemke, G. (2007). TAM Receptors Are Pleiotropic Inhibitors of the Innate Immune Response. *Cell* 131, 1124–1136.

Sacramento, C.Q., de Melo, G.R., de Freitas, C.S., Rocha, N., Hoelz, L.V.B., Miranda, M., Fintelman-Rodrigues, N., Marttorelli, A., Ferreira, A.C., Barbosa-Lima, G., et al. (2017). The clinically approved antiviral drug sofosbuvir inhibits Zika virus replication. *Sci Rep* 7, 40920.

Sarkar, M., Woodland C, C., Koren, G., and Einarson, A.R. (2006a). Pregnancy outcome following gestational exposure to azithromycin. *BMC Pregnancy Childbirth* 6, 18.

Sarkar, M., Woodland C, C., Koren, G., and Einarson, A.R. (2006b). Pregnancy outcome following gestational exposure to azithromycin. *BMC Pregnancy Childbirth* 6, 18.

Stenglein, M.D., Jacobson, E.R., Wozniak, E.J., Wellehan, J.F.X., Kincaid, A., Gordon, M., Porter, B.F., Baumgartner, W., Stahl, S., Kelley, K., et al. (2014). Ball Python Nidovirus: a Candidate Etiologic Agent for Severe Respiratory Disease in Python regius. *MBio* 5, e01484-14.

Sutton, A.L., Acosta, E.P., Larson, K.B., Kerstner-Wood, C.D., Tita, A.T., and Biggio, J.R. (2015). Perinatal pharmacokinetics of azithromycin for cesarean prophylaxis. *American Journal of Obstetrics and Gynecology* 212, 812.e1-812.e6.

Tang, H., Hammack, C., Ogden, S.C., Wen, Z., Qian, X., Li, Y., Yao, B., Shin, J., Zhang, F., Lee, E.M., et al. (2016). Zika Virus Infects Human Cortical Neural Progenitors and Attenuates Their Growth. *Cell Stem Cell* 18, 587–590.

Thornton, G.K., and Woods, C.G. (2009). Primary microcephaly: do all roads lead to Rome? *Trends in Genetics* 25, 501–510.

Workowski, K.A., Bolan, G.A., and Centers for Disease Control and Prevention (2015).

Sexually transmitted diseases treatment guidelines, 2015. *MMWR Recomm Rep* 64, 1–137.

Xu, M., Lee, E.M., Wen, Z., Cheng, Y., Huang, W.-K., Qian, X., Tcw, J., Kouznetsova, J.,

Ogden, S.C., Hammack, C., et al. (2016). Identification of small-molecule inhibitors of Zika virus infection and induced neural cell death via a drug repurposing screen. *Nat Med* 22, 1101–1107.

Chapter 2 Innate immune response of the developing brain to viruses

Contributions

Includes contributions from Galina Schmunk, David Shin, Denise Allen (sample acquisition), Tomasz Nowakowski, Arnold Kriegstein (guidance and resources).

The human brain develops via a tightly coordinated series of molecular events that define regions, cell types, and connections between them (Kriegstein and Alvarez-Buylla 2009). This orchestration is highly reproducible. While resilient to some perturbations (such as a multitude of genetic variation), and susceptible in ways we don't yet understand to many others (such as environmental contributors to neuropsychiatric disease), the course of development is very clearly damaged by Zika virus and other infections of the fetus during pregnancy, often with debilitating consequences (Patrícia Brasil et al. 2016; Ostrander and Bale 2019). Damage could take the egregious form of cell death and immune infiltration, or simply delaying and disorganizing the normal course of development.

At the outset of this project, it was unclear how Zika virus might globally affect development. In particular, what mechanisms of damage occurred in a complete tissue representing all the cell types of the developing brain? Here, I explored how the innate immune response contributes to dysregulated development of the human brain. The focus is the changes that occur in **un**infected cells, in response to cytokines and other changes that occur in infected cells. These uninfected, bystander cells may share a large role in the damage that occurs in the setting of viral infection. Importantly, the mechanisms by which they do so that may give us insight into pathophysiology of other developmental abnormalities where similar systems are activated without the initial viral insult.

Primary human brain exhibits a strong response to RNA viruses consistent with type I interferon signaling

As an initial approach I cultured organotypic slices of mid-gestation human cortical tissue, infected in vitro with Zika virus, and then examined the transcriptional response with a method that allowed resolution of the multitude of cell types in the complex tissue: single cell RNA sequencing (scRNASeq) (Figure 2.1). Using the 10X platform, I observed a strong overall response that distinguished ZIKV from mock infection, with upregulation of genes involved in responding to type I interferon signaling such as ISG15 and IFI6 (Figure 2.2, Figure 2.3). While a similar set of genes was upregulated across all cell types, the magnitude of change was greatest in radial glia and astrocytes compared to neurons (Figure 2.4, Figure 2.5, Figure 2.6). In addition, a few genes exhibited some cell-type specificity, such as CCL8 in microglia, IFI27 primarily in glial cells, and LGALS3BP which increased in many cell types but was absent from microglia while its ligand, LGALS3, is strongly expressed (Figure 2.7, Figure 2.8, Figure 2.9).

Sample	bHR00 (GW21)		bHR12 (GW18)		bHR13 (GW15)		bHR9 (GW 23)			bHR40 (GW19)	
condition	mock	ZIKV	mock	ZIKV	mock	ZIKV	mock	RV	ZIKV	untx	IFNB
# cells	7,726	10,852	9,513	7,219	15,452	10,044	23,810	26,674	22,866	17,258	9,815
mean reads/cell	20,197	15,264	11,033	13,964	6,575	11,115	6,752	5,259	7,730	19,335	35,949
median genes/cell	1,689	1,467	1,183	1,460	1,051	1,348	819	693	904	1,559	1,908
median UMI/cell	3,956	3,368	2,506	3,257	2,064	2,846	1,433	1,118	1,622	2,887	3,940

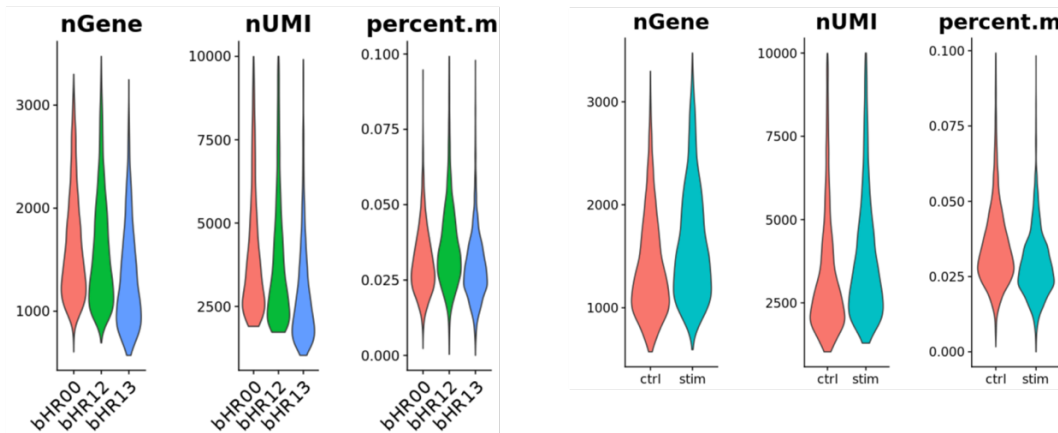
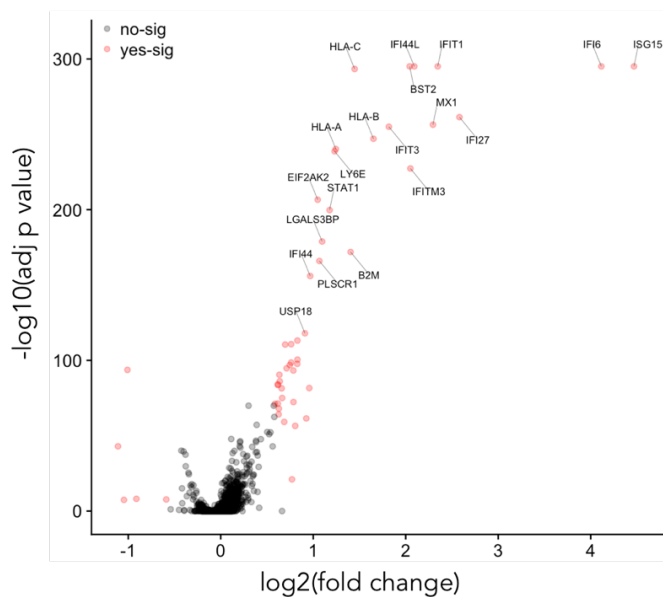


Figure 2.1 Approach: scRNA-Seq on primary human fetal brain tissue after in vitro infection.

Tissue from primary human brain of ages GW15-23 was cultured, infected in vitro with Zika virus (ZIKV) or rubella virus (RV) or treated with recombinant human interferon-beta protein (IFNB). At 72hrs post-treatment, tissue was dissociated and processed on the 10X platform to make single-cell RNA sequencing libraries, which were processed and analyzed using Cell Ranger, Seurat, and custom tools.



GO Term	#	FDR
type I interferon signaling pathway	21	1.9E-34
defense response to virus	19	1.4E-21
negative regulation of viral genome replication	12	6.9E-17
interferon-gamma-mediated signaling pathway	11	5.8E-12
response to virus	12	1.1E-11
antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent	9	9.5E-09
innate immune response	13	2.2E-06
antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-independent	5	9.2E-06
antigen processing and presentation of peptide antigen via MHC class I	6	2.6E-05
response to interferon-beta	4	2.3E-03
response to interferon-alpha	4	3.3E-03
positive regulation of T cell mediated cytotoxicity	4	7.9E-03

Figure 2.2 Overall transcriptional response to ZIKV vs. mock

Volcano plot shows fold change and significance of each gene comparing ZIKV to mock across 3 samples. GO terms for significant genes shown at right, highlighting common pathways that are upregulated in response to type I interferon signaling.

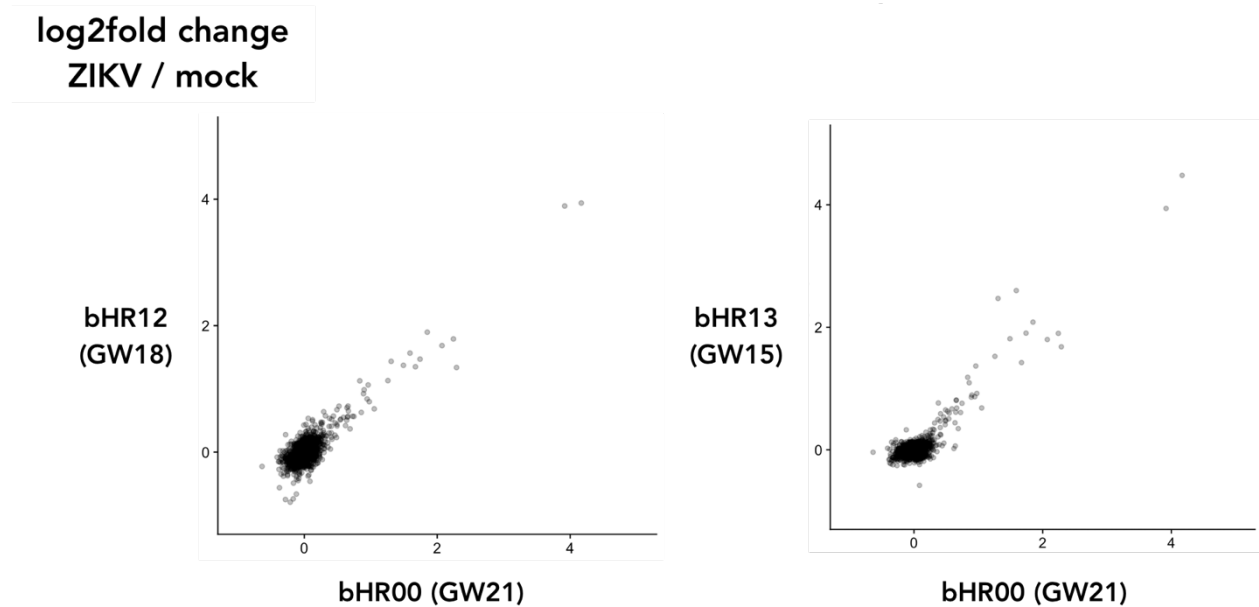


Figure 2.3 Consistency in transcriptional response to ZIKV between replicate samples.

Seurat toolkit

1. Use 1000 highly variable genes within each sample for canonical correlation analysis (CCA)
2. Align subspaces
3. Identify clusters by SNN algorithm using CCA reduction
4. Find conserved markers for each cluster
5. Check expression of canonical cell type markers

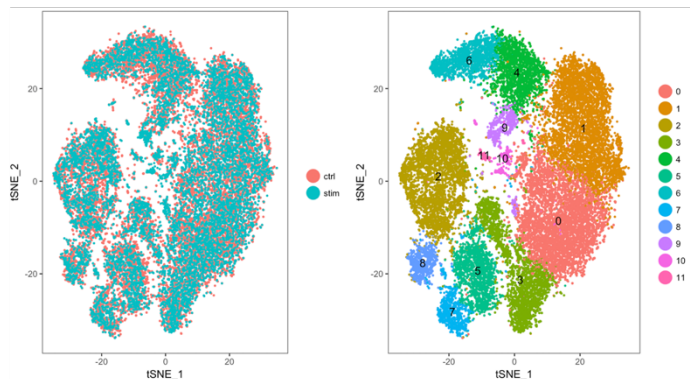


Figure 2.4 Cell type classification approach for scRNAseq

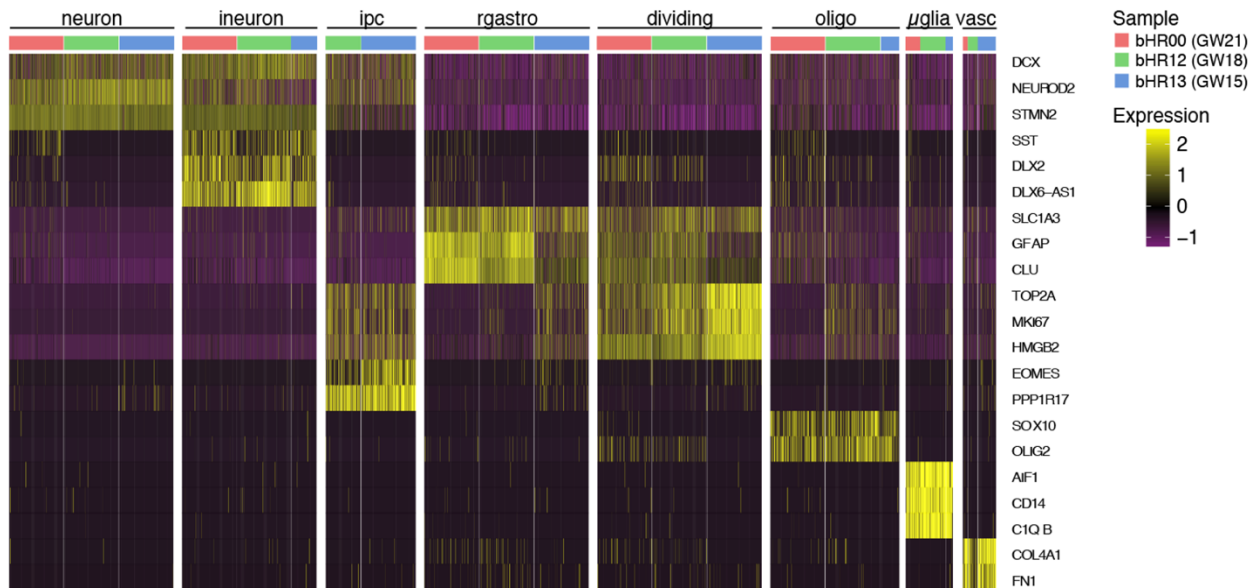


Figure 2.5 Cell type classification in scRNAseq data.

Across all three samples (GW15, 18, 21), expected cell types were identified by canonical markers, including neurons, interneurons, intermediate progenitor cells, radial glia and astrocytes, dividing cells, oligodendrocyte precursor cells, microglia, and vasculature-associated cells.

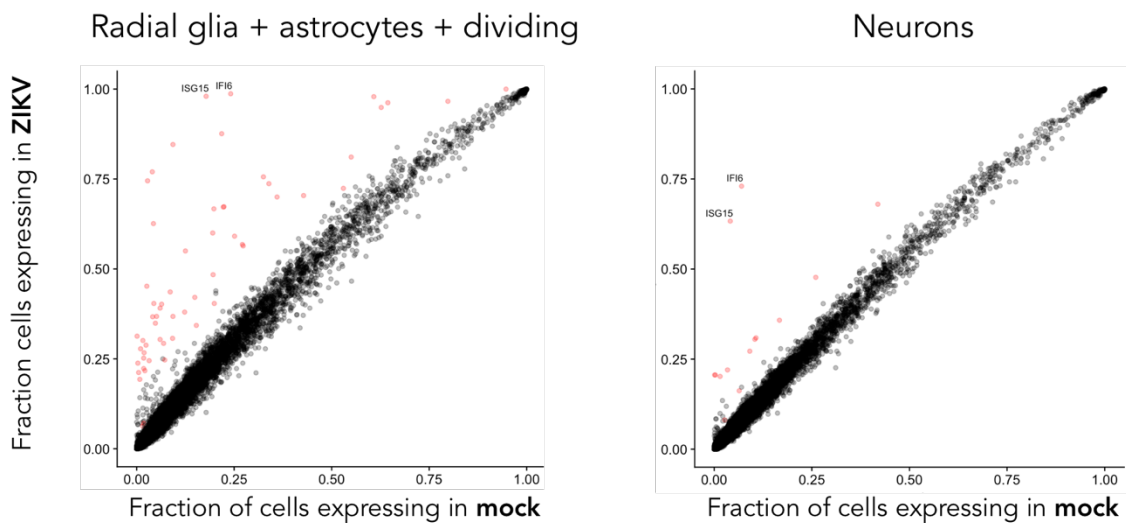


Figure 2.6 Strongest transcriptional response to ZIKV observed in radial glia and astrocytes.

Each gene is plotted as the fraction of cells in which that gene was detected, comparing ZIKV vs mock, in glial cells (left) and in neurons (right).

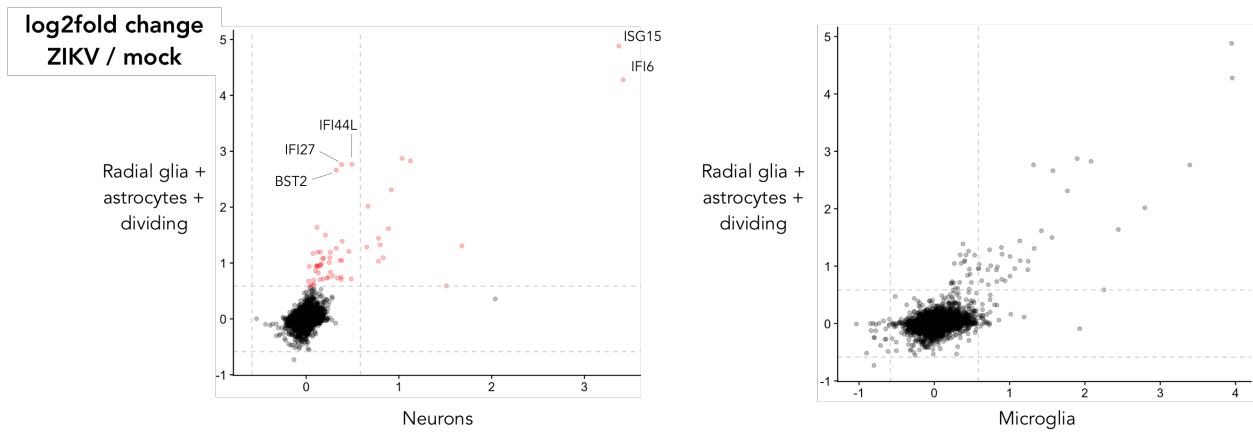


Figure 2.7 Distinguishing whether cell types show fundamentally different responses or different amplitudes of the same response.

Left, comparing radial glia and astrocytes to neurons, many genes are significant only for glia, but the strongest transcriptional changes are seen in both. Right, comparing radial glia and astrocytes to microglia, there may be genes that are specific to the microglia response such as *CCL8*.

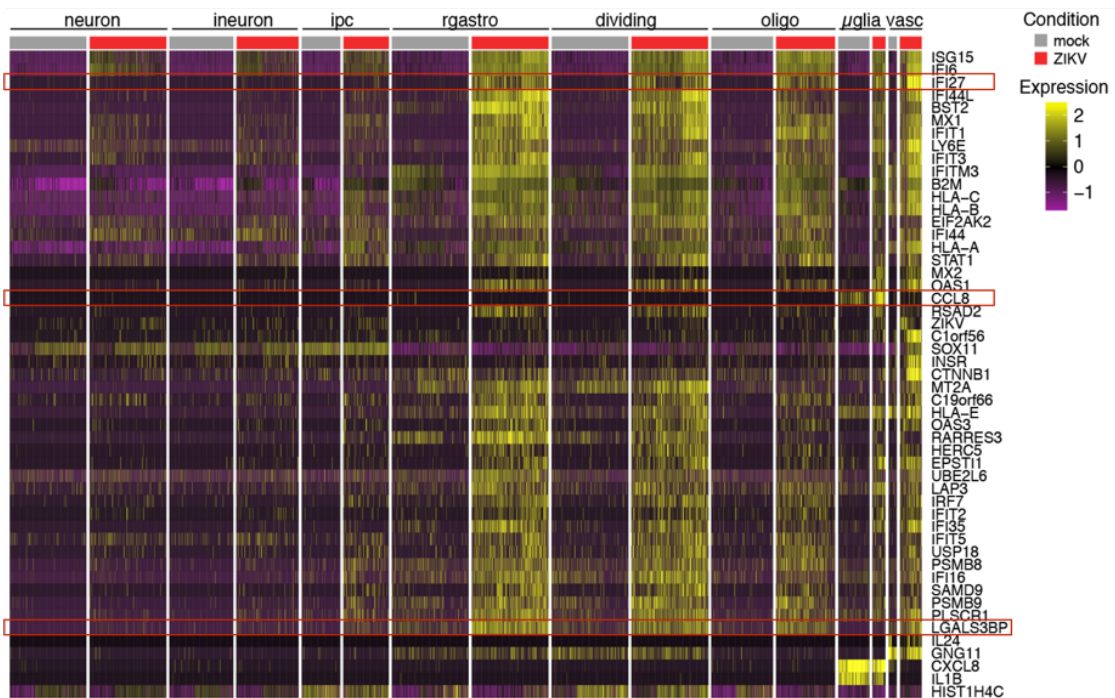


Figure 2.8 Cell type-specific responses to ZIKV

Most of the strongest response genes are observed across multiple cell types. A few highlighted here, may be more cell type specific, e.g. *IFI27* and *CCL8*.

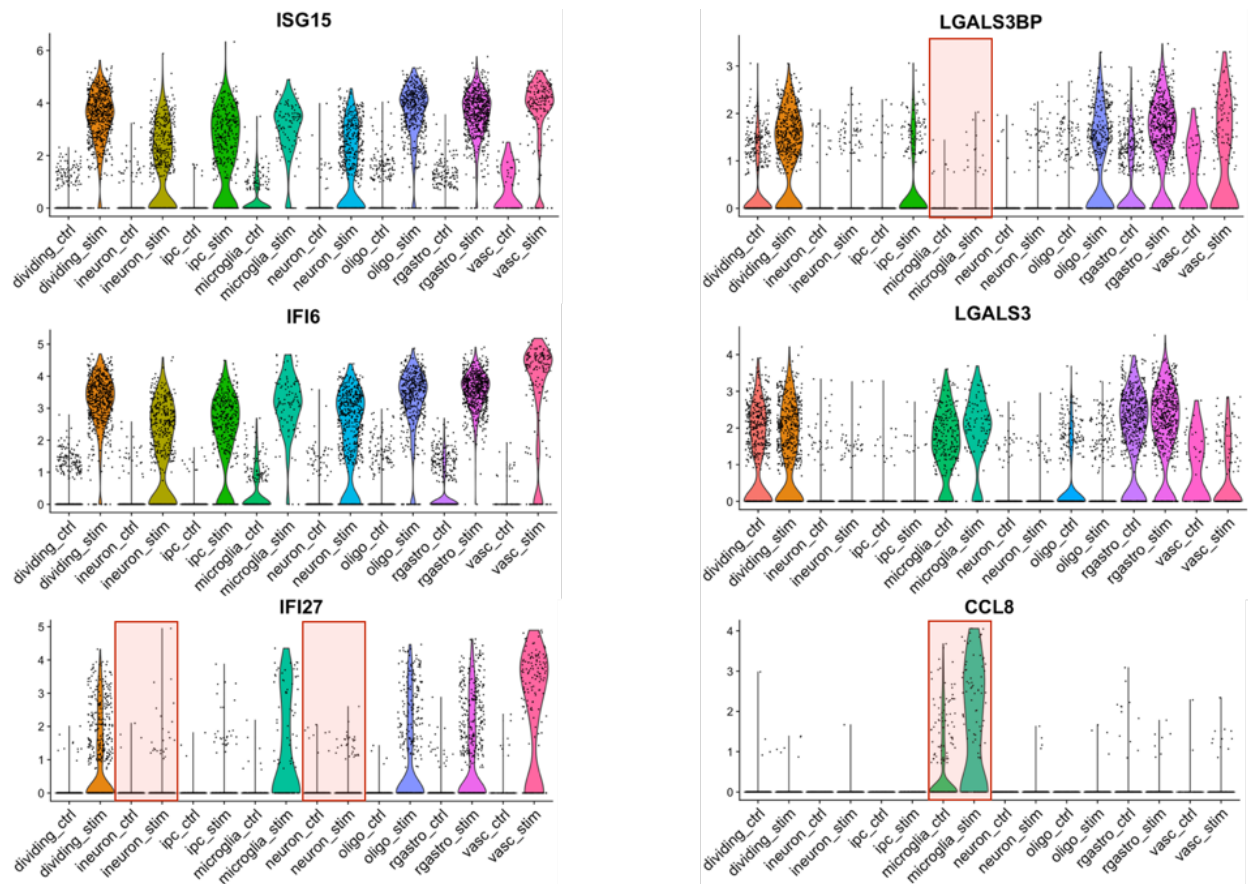


Figure 2.9 Putative cell-type specific responses to ZIKV

For each gene, plotting normalized expression levels in ZIKV (“stim”) and uninfected (“ctrl”). The strongest are upregulated in all cell types, ISG15, IFI6. In contrast, IFI27 was not detectable in neurons at baseline or after stimulation, and LGALS3BP was not detectable in microglia. Both genes increased after infection in other cell types. Notably, CCL8 was observed to increase primarily in microglia.

The set of genes activated were most canonically type I interferon response genes. Of course, the molecular pathways downstream of the various receptors that bind different classes of interferons are overlapping (Schneider, Chevillotte, and Rice 2014). However, for this project I chose to follow the strongest candidate molecule, interferon-beta (IFNB). The scRNA-Seq data showed very low counts of IFNB transcripts but only in the ZIKV condition (as expected for potent cytokines like this), and qPCR verified that IFNB

72 hrs post-ZIKV infection of organotypic slice

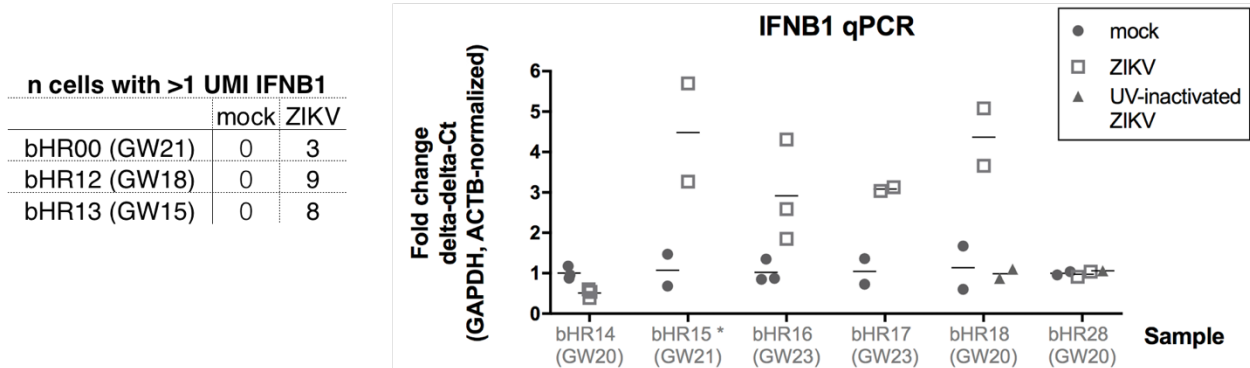


Figure 2.10 Upregulation of interferon-beta (IFNB) after ZIKV infection.

Cytokines are potent, even at low RNA and protein levels. Left, in three samples, a very small number of cells showed >1 count to IFNB transcripts in ZIKV. Right, performing RT-qPCR on additional slice cultures infected with ZIKV (or non-replicating control), clearly showed an increase in IFNB1 transcripts.

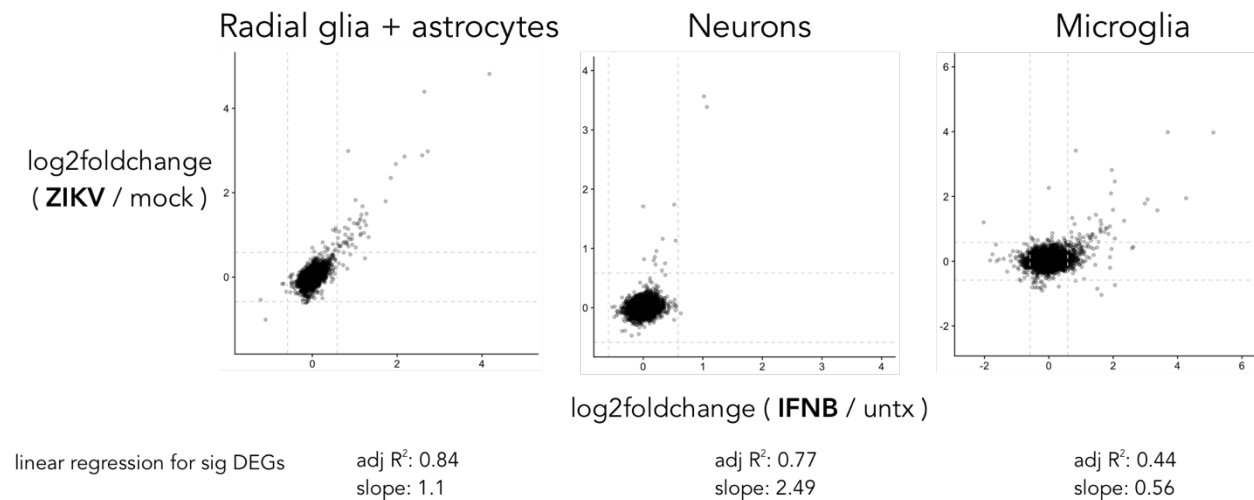


Figure 2.11 Similarity of transcriptional response to ZIKV and to IFNB treatment.

Within each rough cell type grouping, the overall response to ZIKV and to IFNB was similar, with most genes falling along the diagonal. In neurons, IFNB had little transcriptional effect, but the strongest genes were still activated in both (ISG15, IFI6).

was upregulated in response to ZIKV infection, with a requirement for viral replication (Figure 2.10). In addition, the transcriptional response in 3 major cell types was similar in tissues infected with ZIKV or treated with recombinant IFN β (Figure 2.11). In future experiments, this should be distinguished from other perturbations that may cause overall stress to the tissue.

Do ISG15 and USP18 provide negative feedback to terminate T1IFN signaling in primary human brain cells?

In an infected cell, viral sensing leads to the rapid induction and secretion of type I interferons (T1IFN) including IFN- β (Isaacs and Lindenmann 2015; McNab et al. 2015). Through autocrine and paracrine signaling, IFN- β then initiates a cascade of events to amplify the upregulation of over a hundred IFN-stimulated genes (ISGs), which establish an antiviral state in infected and nearby cells (McNab et al. 2015). Type I IFNs have been shown to have a protective role against many neurotropic viruses (Perry et al. 2005; Delhaye et al. 2006; Lindqvist et al. 2016; Daniels et al. 2017; Préhaud et al. 2005). However, genetic conditions such as Aicardi-Goutières and pseudo-TORCH syndromes demonstrate that unmitigated IFN signaling can disrupt neurodevelopment (Aicardi and Goutières 1984; Knoblauch et al. 2003; Meuwissen et al. 2016). In these diseases, mutations in genes involved in nucleic acid clearance or sensing such as three-prime repair exonuclease 1 (TREX1) and interferon induced with helicase C1 (IFIH1, aka MDA-5) provoke constitutive type I IFN production (Crow et al. 2006; Rice et al. 2014), while mutations in ubiquitin-specific peptidase 18 (USP18, aka UBP43) cause unopposed

response (Livingston and Crow 2016). Encephalopathy develops *in utero* or in the first year of life, with features that mimic congenital viral infections such as microcephaly and calcifications (Livingston and Crow 2016). These paradoxical properties for type I IFNs as both antiviral and neurotoxic raise intriguing questions: **what mechanisms, if any, limit immune-mediated damage in the setting of type I IFN production in the developing human brain?** It is tempting to speculate that irreplaceable cell types, such as post-mitotic neurons, might be protected from immune-mediated damage, while support cells (astrocytes) and self-renewing cells (progenitors) assume the burden of viral control for the tissue, complementing the actions of professional immune cells (resident microglia and infiltrating lymphocytes). At the outset of this project, this model had not yet been explored in the developing human brain.

I first verified and characterized the response to type I interferons observed in single cell sequencing via orthogonal assays. RT-qPCR showed a strong (>100 fold) and reproducible increase in ISG15 RNA upon ZIKV activation that depended on viral replication (UV-inactivated virus was non-stimulatory), as seen in Figure 2.12. An increase in ISG15 transcripts was likewise observed within 2 hrs and peaking ~12 hrs after treating primary dissociated cells or slice cultures with recombinant human IFN-B protein (Figure 2.13). Additional genes that were increased after ZIKV infection in the pilot scRNASeq experiment, IFIT1, MX1, ISG15, IFI6, BST2, IFI27, and IFI44L, were also increased after stimulating dissociated cells with IFN-B in a dose-dependent manner and with different time-courses (Figure 2.14). Finally, at the protein level, stimulation of primary cells with recombinant human IFNB resulted in expression of USP18 which was not detectable at

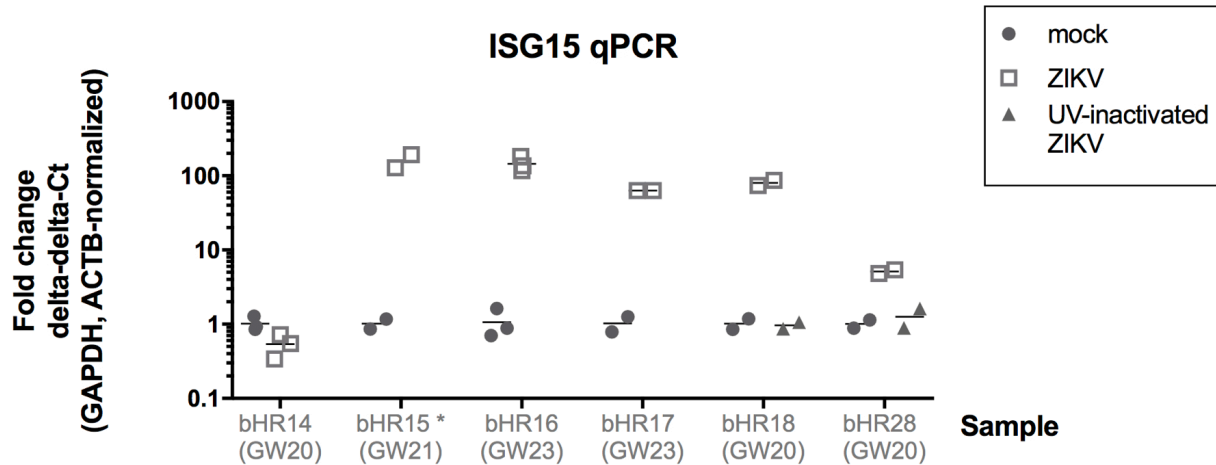


Figure 2.12 Dramatic increase in ISG15 RNA after ZIKV infection. RT-qPCR quantifies ISG15 transcripts after infecting slices with ZIKV or non-replicating control (UV-inactivated).

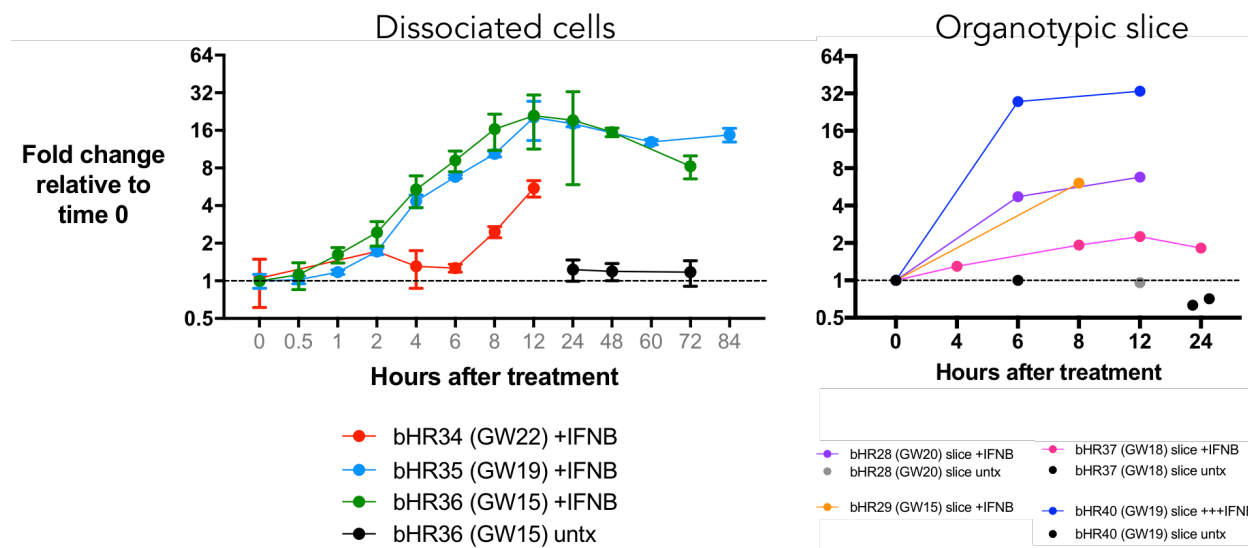


Figure 2.13 Rapid increase in ISG15 RNA after treatment with recombinant IFNB. RT-qPCR after treating dissociated cells (left) or organotypic slice culture (right) with recombinant human IFNB.

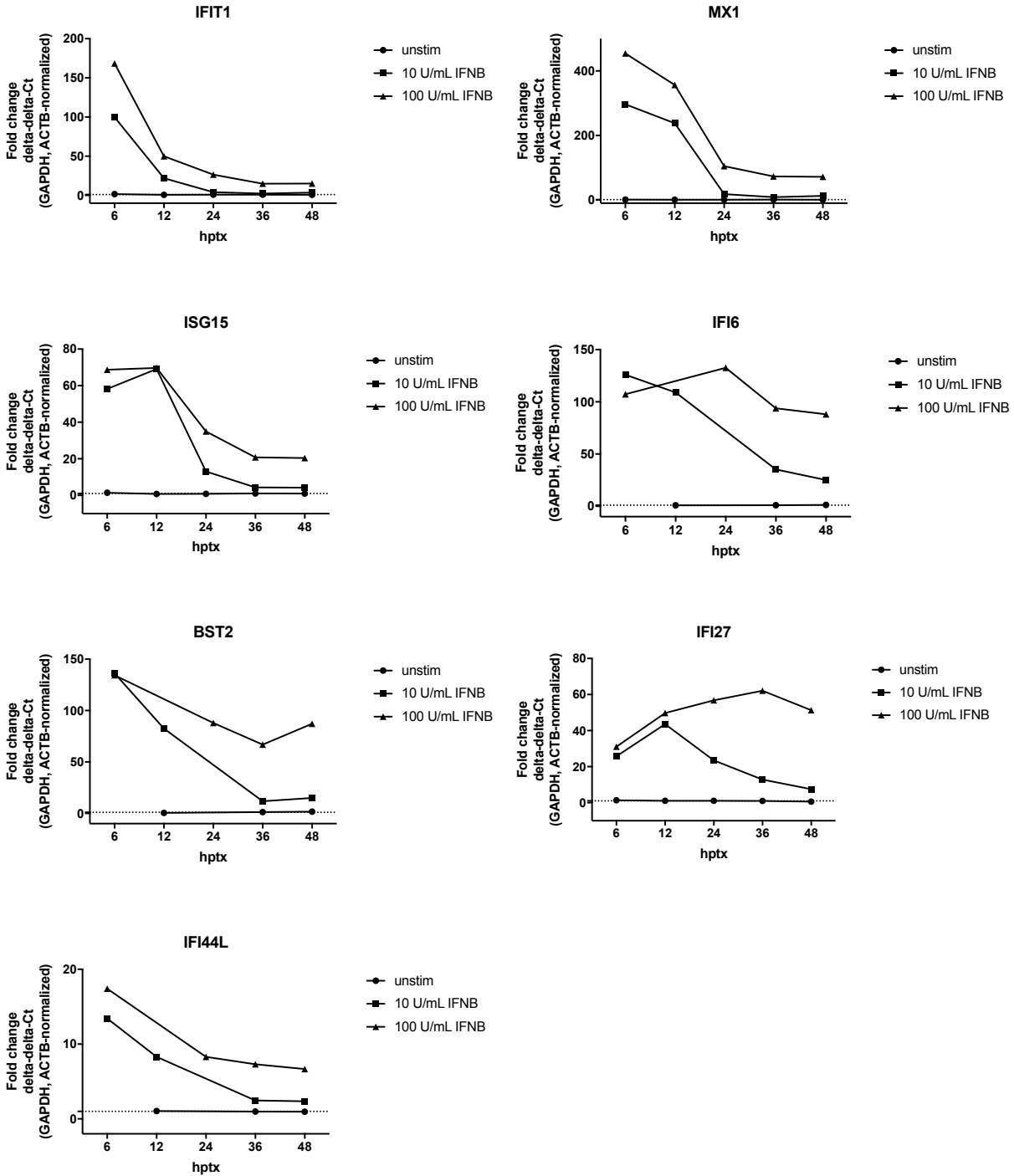


Figure 2.14 Timecourse of type I interferon activation in primary neural cells
 RT-qPCR quantifies the response observed in canonical interferon-stimulated genes after treatment of dissociated primary human brain cells with recombinant human IFNB.

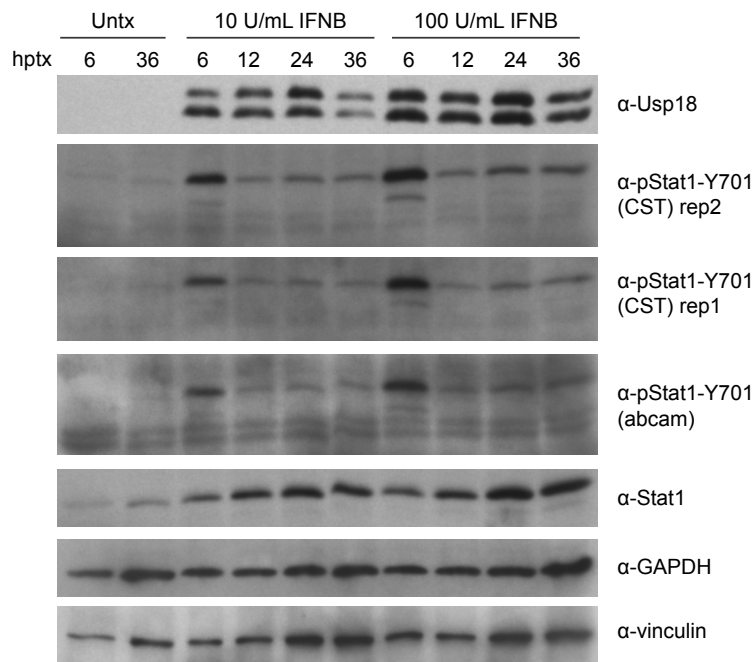


Figure 2.15 Protein upregulation and activation in ISGs after treatment of primary human brain cells with recombinant human IFNB.

Western blots show strong and persistent increase in protein levels of Usp18 and Stat1 after treatment with IFNB. Transient phosphorylation of Stat1 at the Y701 residue primarily observed at 6 hrs post-treatment, and is diminished thereafter.

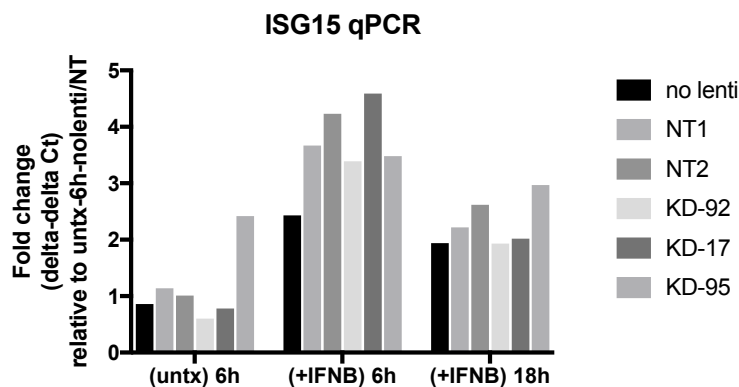


Figure 2.16 Weak response to IFNB at baseline and no significant difference after attempted knockdown of USP18 RNA by shRNAs.

baseline, by 6hrs and lasting out to the final timepoint at 36hrs (Figure 2.15). In addition, I observed upregulation of Stat1 protein with a transient increase in phosphorylation at the Y701 site diminished by 12 hrs post-treatment. Together, these data support the strong response to type I interferon signaling observed in primary cells, dissociated or in the tissue context, comparable when infected with ZIKV or stimulated with recombinant human IFN-B.

For the next step, I attempted to knock down proposed negative regulators of the ISG response, including USP18 and ISG15, which are purported to provide negative feedback to turn off type I

interferon signaling (Zhang et al. 2015; Il Kim and Zhang 2005). Reduced functioning in these proteins is one cause of the interferonopathies that resemble congenital viral infections (Rodero and Crow 2016). In pilot experiments, an attempt to knock down USP18 (Figure 2.16) showed little effect on ISG15 RNA levels. With technical improvements, these experiments may yet help us understand whether similar feedback loops are functional in primary brain cells, which could contribute to mechanisms for damage.

Future directions: functional consequences of T1IFN signaling on progenitor cells

To approach the question of immune-mediated damage from another angle, I asked what are the functional consequences of Type I interferon signaling on development in the human brain. First, I stimulated human cortical slice cultures with recombinant human IFNB and examined the tissue by immunohistochemistry for gross effects on cell death (staining for cleaved caspase 3 - CC3), on proliferation (assessing incorporation of pulsed BrdU to estimate newborn cells and staining for Ki67 to quantify dividing cells by phase of cell cycle), and on changes in the population size of different cell types (staining for Pax6, Satb2, and Iba1 which mark progenitors, neurons for the relevant age of tissue samples, and microglia respectively). In all of these analyses, I observed no strong effects (Figure 2.17, Figure 2.18, Figure 2.19, Figure 2.20, Figure 2.21), despite parallel experiments showing strong transcriptional responses to the same perturbations.

While it is possible that parameters measured are not affected at all, I suspect there are technical limitations that reduced my sensitivity to observe changes in the progression of development. In particular, these pilot experiments were done with a small number of samples, heterogeneity in age, and on cortical tissue but not finer regional dissection. The dose of IFNB and duration of treatment before fixation may also need to be varied. Finally, it may be important to measure other parameters, such as motility in dividing progenitors or microglia, which would be better assessed in live/time-lapse imaging. These data should not be interpreted as harmless effects of IFN-B on the brain, but rather as motivation to study more complete model systems such as organoids or mice, or finer assays with more follow-up to better characterize how an inability to turn off IFN signaling could lead to the abnormalities seen in genetic conditions.

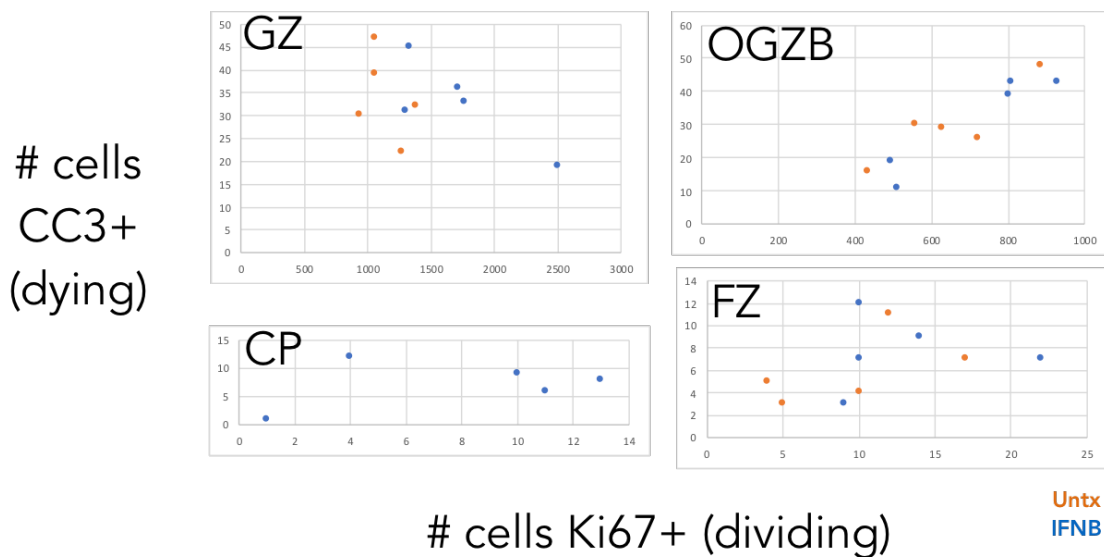


Figure 2.17 Cell death vs. proliferation in primary brain tissue treated with IFNB. Each point represents 1 field of view, fields of view taken from minimum 2 tissue slices for each condition/sample. No strong distinction observed between IFNB-treated and untreated. Regions: GZ (germinal zone), OGZB (outer germinal zone boundary), FZ (fibrous zone), CP (cortical plate)

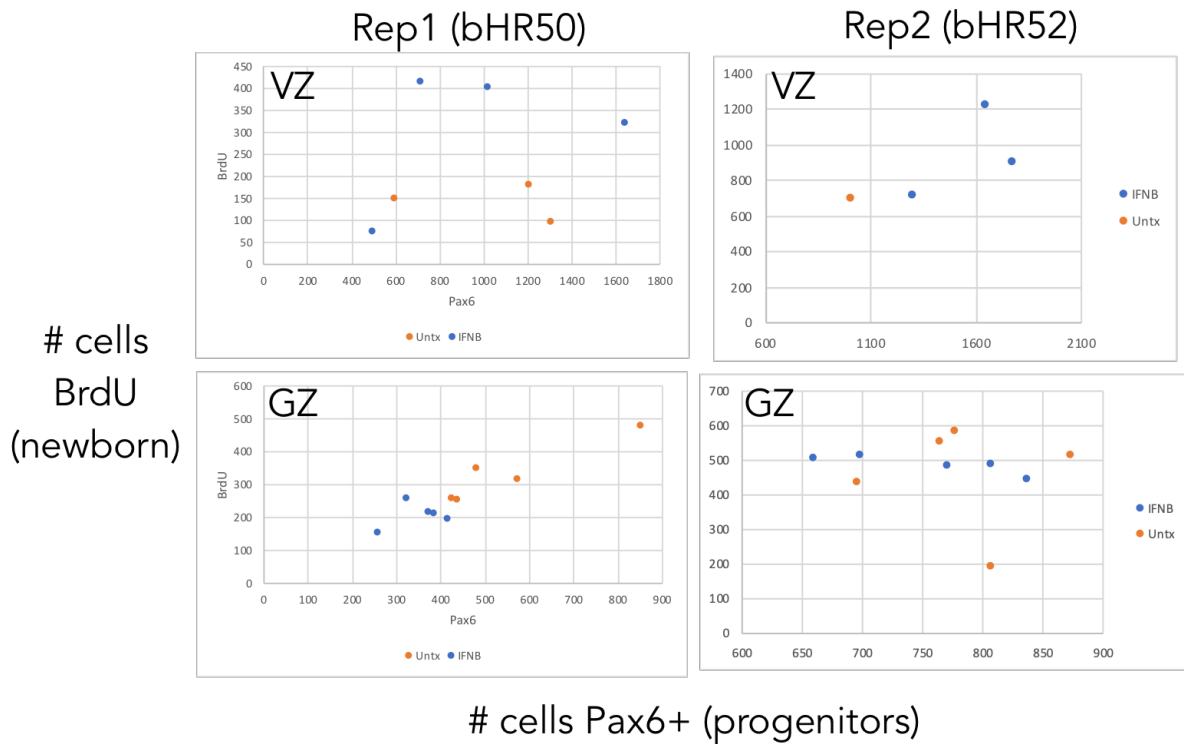


Figure 2.18 Proliferation in germinal zone after IFNB treatment of primary brain tissue.

VZ (ventricular zone); GZ (broader germinal zone). Each point represents 1 field of view, fields of view taken from minimum 2 tissue slices for each condition/sample. No strong distinction observed between IFNB-treated and untreated. VZ (ventricular zone), GZ (broader germinal zone).

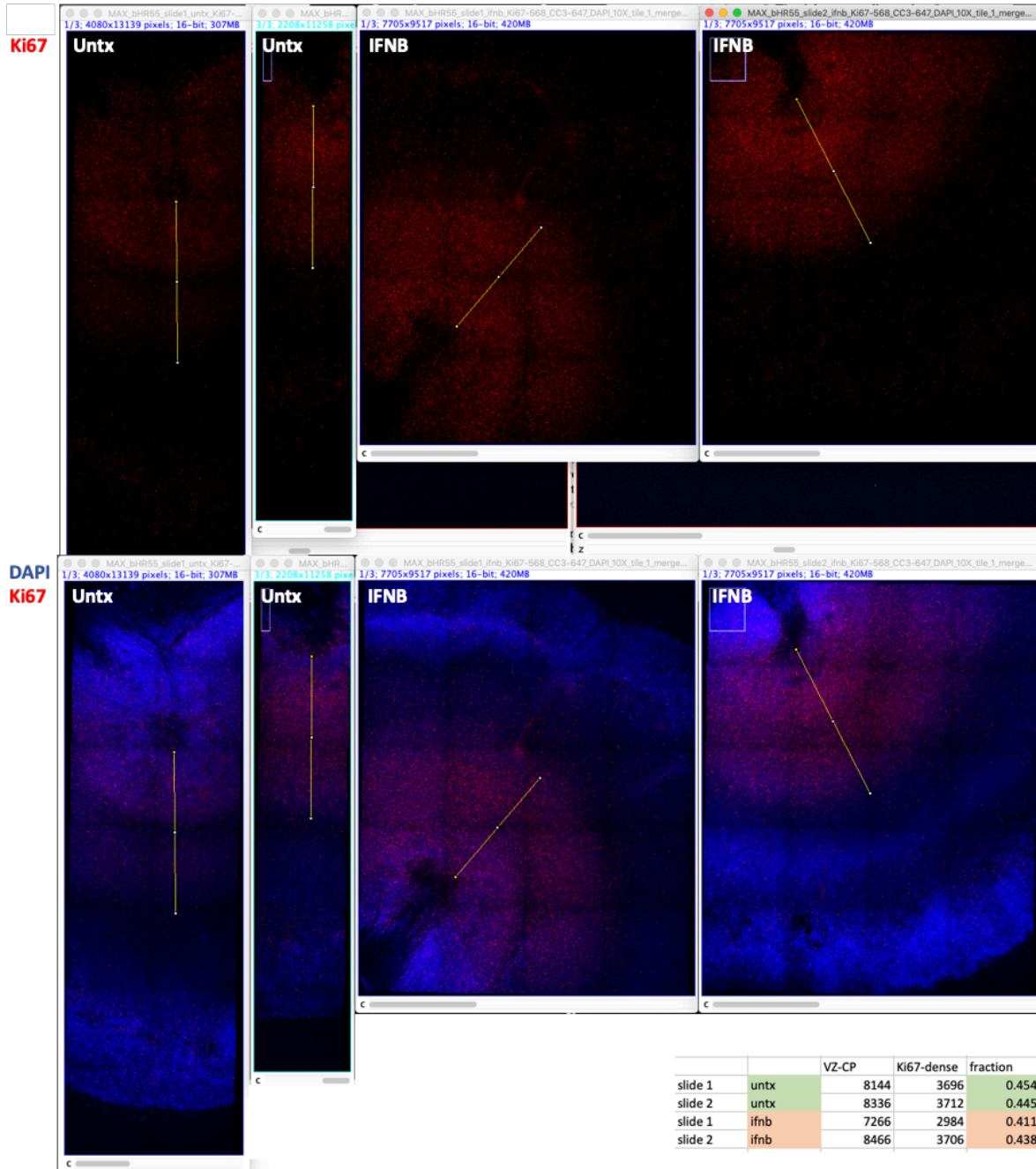
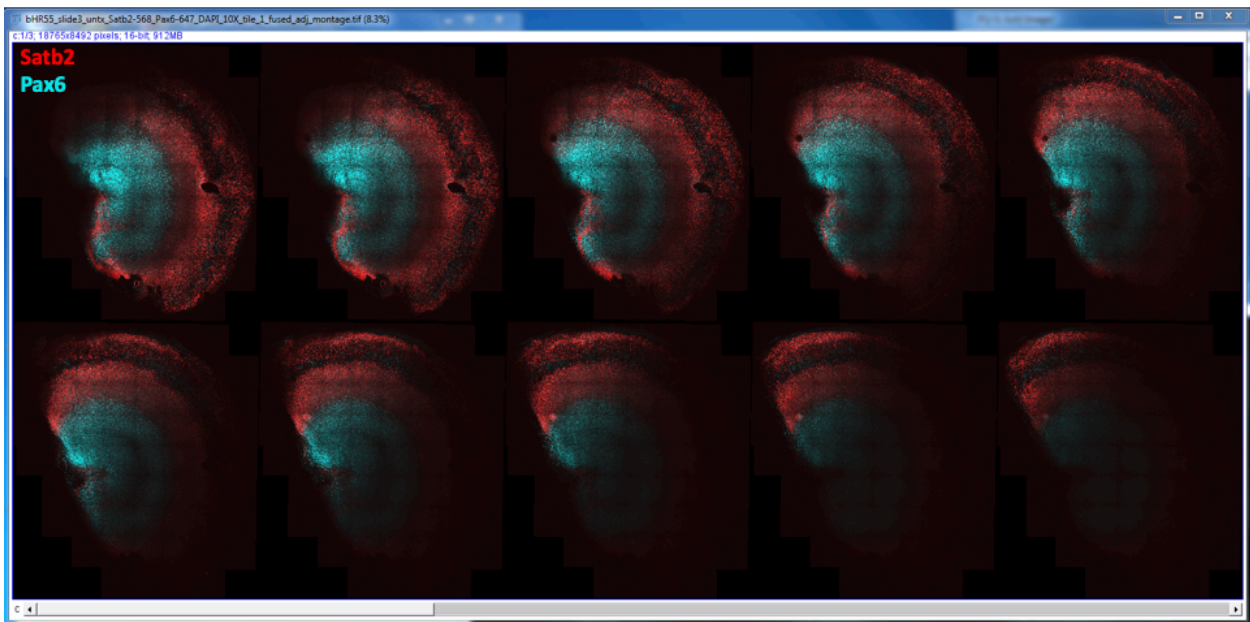


Figure 2.19 Relative size of Ki67-strong germinal zone after IFNB treatment of primary brain tissue.

Measured distance from ventricular edge to cortical plate, and distance of strongest Ki67 staining. No major difference in fraction of slice exhibiting strong Ki67 staining between treated and control.

Untx - zstack



IFNB - zstack

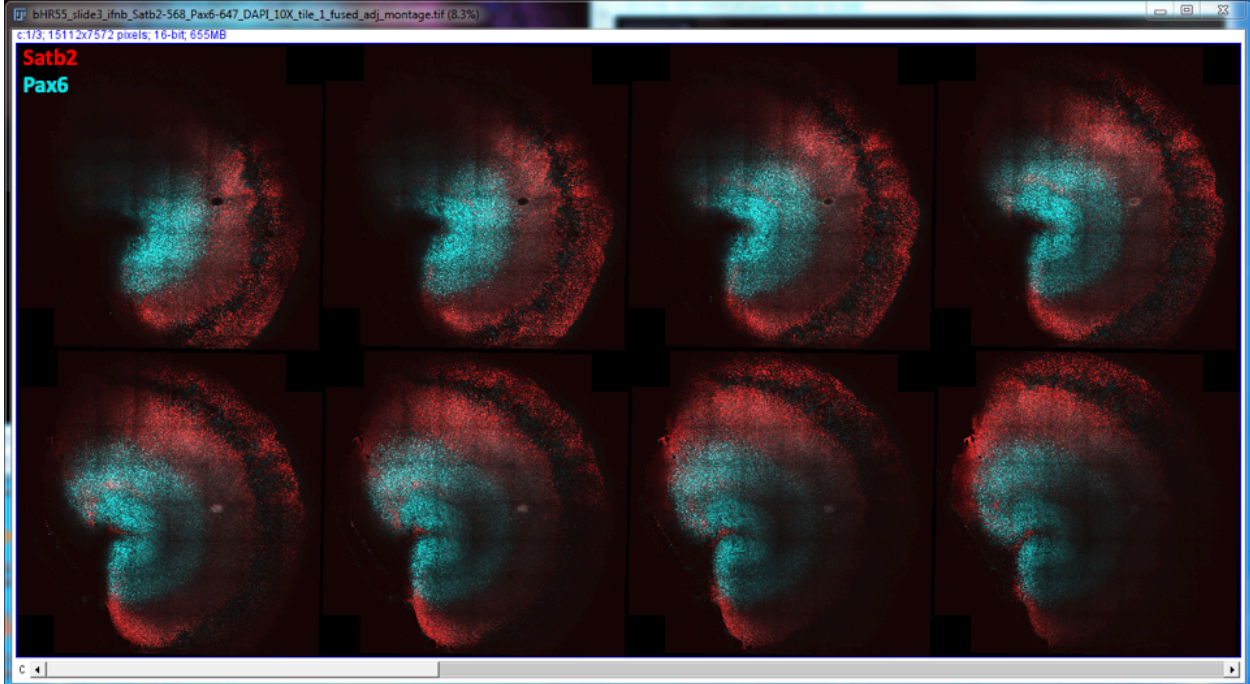


Figure 2.20 Coarse assessment of progenitors and neurons in primary brain tissue treated with IFNB.

Architecture broadly conserved after treatment, with no major difference seen in overall number and location of neurons (Satb2) or progenitors (Pax6).

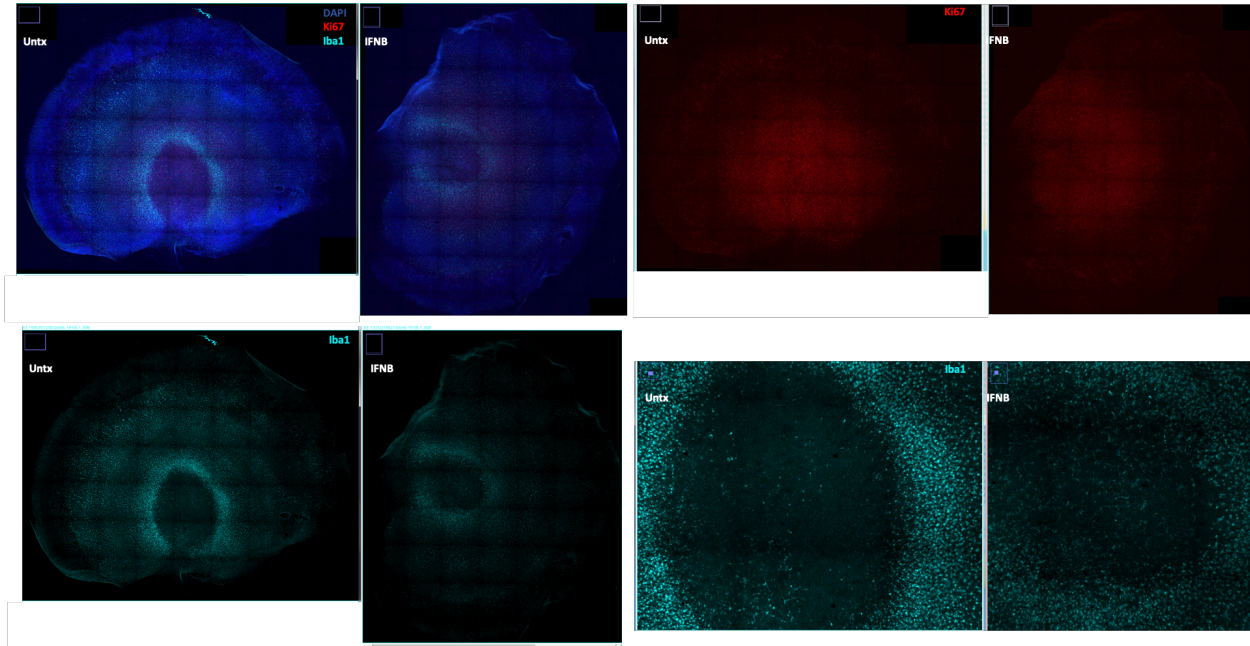


Figure 2.21 Coarse assessment of microglia in primary brain tissue treated with IFNB.

No major differences observed in location or density of microglia (Iba1+ cells). Requires finer resolution to assess morphology, and live imaging to understand behavior.

References for Chapter 2

- Aicardi, J., and Goutières, F. (1984). A Progressive familial encephalopathy in infancy with calcifications of the basal ganglia and chronic cerebrospinal fluid lymphocytosis: Aicardi and Goutieres: Basal Ganglia Calcifications. *Ann Neurol.* 15, 49–54.
- Brasil, P., Pereira, J.P., Moreira, M.E., Ribeiro Nogueira, R.M., Damasceno, L., Wakimoto, M., Rabello, R.S., Valderramos, S.G., Halai, U.-A., Salles, T.S., et al. (2016). Zika Virus Infection in Pregnant Women in Rio de Janeiro. *N Engl J Med* 375, 2321–2334.
- Crow, Y.J., Hayward, B.E., Parmar, R., Robins, P., Leitch, A., Ali, M., Black, D.N., van Bokhoven, H., Brunner, H.G., Hamel, B.C., et al. (2006). Mutations in the gene encoding the 3'-5' DNA exonuclease TREX1 cause Aicardi-Goutières syndrome at the AGS1 locus. *Nat Genet* 38, 917–920.
- Daniels, B.P., Jujjavarapu, H., Durrant, D.M., Williams, J.L., Green, R.R., White, J.P., Lazear, H.M., Gale, M., Diamond, M.S., and Klein, R.S. (2017). Regional astrocyte IFN signaling restricts pathogenesis during neurotropic viral infection. *Journal of Clinical Investigation* 127, 843–856.
- Delhaye, S., Paul, S., Blakqori, G., Minet, M., Weber, F., Staeheli, P., and Michiels, T. (2006). Neurons produce type I interferon during viral encephalitis. *Proceedings of the National Academy of Sciences* 103, 7835–7840.

- Il Kim, K., and Zhang, D. (2005). UBP43, an ISG15-Specific Deconjugating Enzyme: Expression, Purification, and Enzymatic Assays. In *Methods in Enzymology*, (Elsevier), pp. 491–499.
- Isaacs, A., and Lindenmann, J. (2015). Pillars Article: Virus Interference. I. The Interferon. *Proc R Soc Lond B Biol Sci.* 1957. 147: 258-267. *J. Immunol.* 195, 1911–1920.
- Knoblauch, H., Tennstedt, C., Brueck, W., Hammer, H., Vulliamy, T., Dokal, I., Lehmann, R., Hanefeld, F., and Tinschert, S. (2003). Two brothers with findings resembling congenital intrauterine infection-like syndrome (pseudo-TORCH syndrome). *Am. J. Med. Genet.* 120A, 261–265.
- Kriegstein, A., and Alvarez-Buylla, A. (2009). The Glial Nature of Embryonic and Adult Neural Stem Cells. *Annu. Rev. Neurosci.* 32, 149–184.
- Lindqvist, R., Mundt, F., Gilthorpe, J.D., Wölfel, S., Gekara, N.O., Kröger, A., and Överby, A.K. (2016). Fast type I interferon response protects astrocytes from flavivirus infection and virus-induced cytopathic effects. *J Neuroinflammation* 13, 277.
- Livingston, J., and Crow, Y. (2016). Neurologic Phenotypes Associated with Mutations in TREX1, RNASEH2A, RNASEH2B, RNASEH2C, SAMHD1, ADAR1, and IFIH1: Aicardi–Goutières Syndrome and Beyond. *Neuropediatrics* 47, 355–360.
- McNab, F., Mayer-Barber, K., Sher, A., Wack, A., and O’Garra, A. (2015). Type I interferons in infectious disease. *Nat. Rev. Immunol.* 15, 87–103.
- Meuwissen, M.E.C., Schot, R., Buta, S., Oudesluijs, G., Tinschert, S., Speer, S.D., Li, Z., van Unen, L., Heijman, D., Goldmann, T., et al. (2016). Human USP18 deficiency

- underlies type 1 interferonopathy leading to severe pseudo-TORCH syndrome. *Journal of Experimental Medicine* 213, 1163–1174.
- Ostrander, B., and Bale, J.F. (2019). Congenital and perinatal infections. In *Handbook of Clinical Neurology*, (Elsevier), pp. 133–153.
- Perry, A.K., Chen, G., Zheng, D., Tang, H., and Cheng, G. (2005). The host type I interferon response to viral and bacterial infections. *Cell Res* 15, 407–422.
- Préhaud, C., Mégret, F., Lafage, M., and Lafon, M. (2005). Virus Infection Switches TLR-3-Positive Human Neurons To Become Strong Producers of Beta Interferon. *JVI* 79, 12893–12904.
- Rice, G.I., del Toro Duany, Y., Jenkinson, E.M., Forte, G.M.A., Anderson, B.H., Ariaudo, G., Bader-Meunier, B., Baildam, E.M., Battini, R., Beresford, M.W., et al. (2014). Gain-of-function mutations in IFIH1 cause a spectrum of human disease phenotypes associated with upregulated type I interferon signaling. *Nat Genet* 46, 503–509.
- Rodero, M.P., and Crow, Y.J. (2016). Type I interferon–mediated monogenic autoinflammation: The type I interferonopathies, a conceptual overview. *The Journal of Experimental Medicine* 213, 2527–2538.
- Schneider, W.M., Chevillotte, M.D., and Rice, C.M. (2014). Interferon-Stimulated Genes: A Complex Web of Host Defenses. *Annu. Rev. Immunol.* 32, 513–545.
- Zhang, X., Bogunovic, D., Payelle-Brogard, B., Francois-Newton, V., Speer, S.D., Yuan, C., Volpi, S., Li, Z., Sanal, O., Mansouri, D., et al. (2015). Human intracellular

ISG15 prevents interferon- α/β over-amplification and auto-inflammation. Nature
517, 89-93.

Chapter 3 Tropism of rubella virus in the human developing brain

Contributions

Includes contributions from: Galina Schmunk (sample acquisition, assistance with cell culture and experiments with dissociated microglia), Tomasz Nowakowski (guidance), Tom Hobman (infectious clone of rubella strain M33).

Rubella virus was first recognized as an important human pathogen in the 1940s, when it was connected to congenital cataracts (Gregg 1941). During the consequent epidemic that spread worldwide, it was found that rubella virus infection during pregnancy can cause multiple congenital abnormalities in the fetus, including microcephaly, deafness, and various others outside of the central nervous system such as heart defects (Banatvala and Brown 2004). In the late 1960s, highly effective vaccines were developed and deployed to curb the epidemic, such that congenital rubella syndrome is rare in developed countries today, where vaccine coverage is generally high. However, there remain over 100,000 cases annually of congenital rubella infection worldwide, especially in lower resource settings where vaccination is incomplete, resulting in a large burden of disease (Grant et al. 2017).

Several limitations have resulted in major gaps in our understanding of the pathophysiology of congenital rubella syndrome. Firstly, no reliable animal models were well-established that recreated the CNS manifestations in the fetus (Plotkin et al. 2011). Secondly, the success of the vaccine dramatically reduced cases such that human specimens are hard to obtain today – when they were more readily available, the molecular techniques for investigation were more basic. Finally, and related to the vaccine's success, there has simply not been enough interest in this unique virus to maintain research programs. Rubella virus is the only member of its family to affect humans, is fairly distant from other human pathogens of the *Alphavirus* genus such as chikungunya virus, and has little in common at the genomic level with other viruses that cause congenital abnormalities, such as cytomegalovirus, HIV, herpes viruses,

lymphocytic choriomeningitis virus, and zika virus. Nonetheless, approaching the question of viral damage to the developing human brain from multiple complementary angles could help elucidate mechanisms of pathology that are common to many pathogens, and potentially non-infectious diseases as well.

During the early era of Rubella virus research, the manifestations of microcephaly were well-established, but it was not even clear whether the virus replicated in the brain itself. Although viral RNA and/or antigens have been found in the CNS (Lazar et al. 2016; Nguyen, Pham, and Abe 2015; Monif et al. 1965; Korones et al. 1965; Esterly and Oppenheimer 1967) there is no direct evidence for viral replication within particular cells or regions of the developing brain. Furthermore, it is not well understood how the infected cells or nearby **un**infected cells may contribute to developmental abnormalities. Here, I identify cell types within primary human brain tissue that are targeted by rubella virus, and explore the consequences of this infection. These findings and follow-up studies may inform our understanding of the developmental consequences of perturbing such cell types. Such knowledge would extend equally to understanding the pathophysiology of a variety of neurotropic teratogenic viruses, in addition to syndromes of immune activation due to genetic or environmental causes.

Microglia are targeted by rubella virus

The initial approach for exploring tropism of rubella virus (RV) in the human developing brain was to perform organotypic slice culture of primary human mid-gestation cortical tissue, inoculate with rubella virus in culture, and then at 72 hrs post

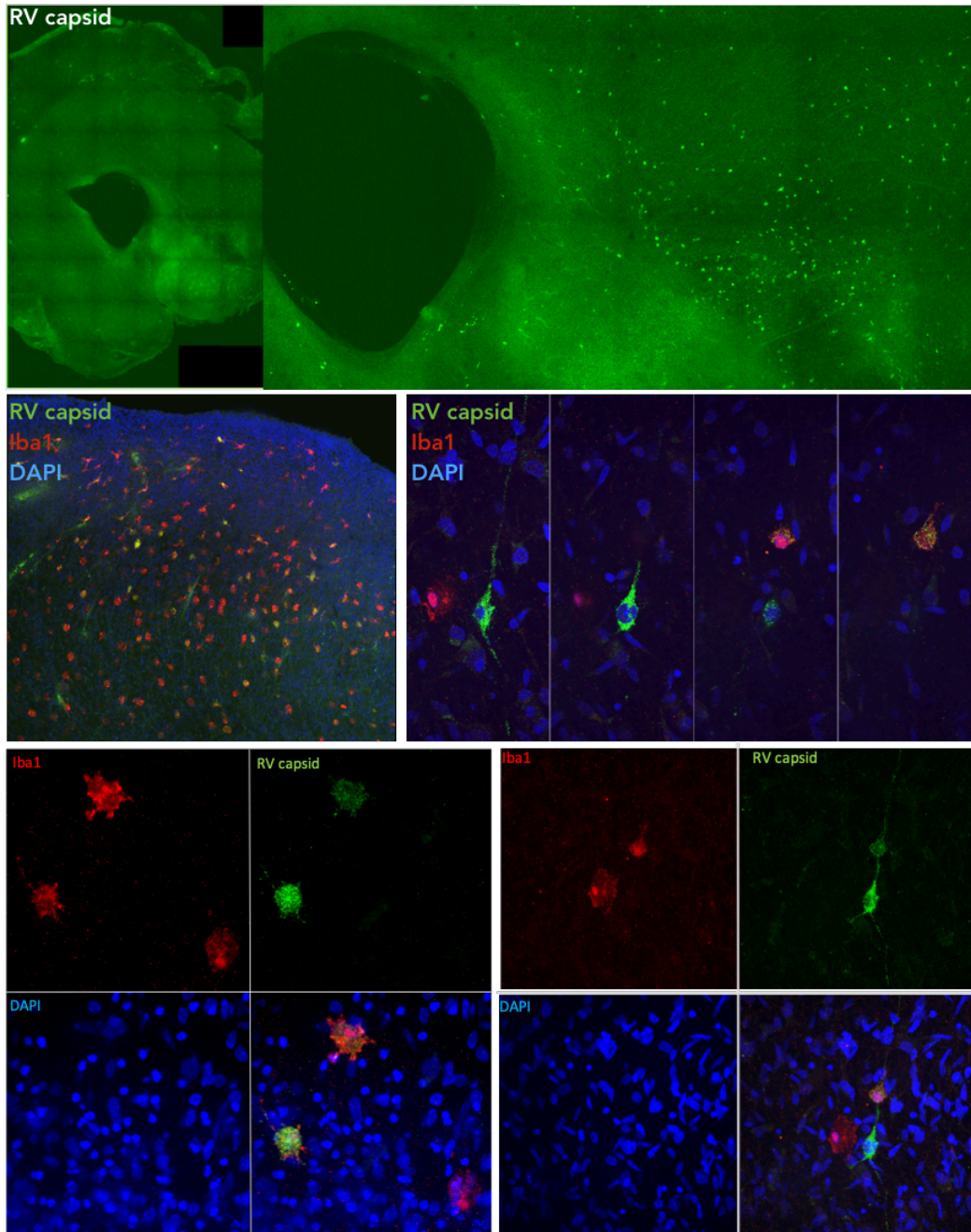


Figure 3.1 Rubella virus capsid protein in microglia after infection of primary human brain tissue

Immunostaining shows viral capsid protein (RV capsid) in cells labeled by Iba1 (microglia marker), 72 hrs after inoculation of primary human brain tissue (gestational week 18-23) in organotypic slice culture with rubella virus (M33 strain). Top panel, low magnification showing near-full slice of ~5mm in diameter, showing cluster of RVcapsid+ cells on the right. Lower panels, increased magnification shows individual cells, many but not all co-labeled with RVcapsid and Iba1.

inoculation, immunostain to identify rubella virus antigens and co-labeling cell type markers. This approach, on tissue from 5 individuals, repeatedly showed immunostaining for rubella capsid protein in cells that stained with Iba1, a microglia marker (**Error! Reference source not found.**). In a pilot experiment, slice cultures from a single individual inoculated with rubella or mock (supernatant from Vero cells, processed comparably to viral stocks) were processed for single cell RNA sequencing on the 10X platform. Although counts of viral RNA were low

Table 3.1 Number of cells from single cell RNA sequencing of Rubella virus or mock-infected organotypic slice culture with (+) or without (-) RV RNA.

Microglia are the predominant cell type where RV RNA was observed.

	Mock-infected			RV-infected		
	-	+	Frac RV+	-	+	Frac RV+
dividing	2471	0	0.000	1349	9	0.007
interneuron	4252	0	0.000	3592	5	0.001
microglia	400	0	0.000	427	67	0.157
neuron	6502	0	0.000	6102	9	0.001
oligo	2033	0	0.000	1292	10	0.008
RG/astrocyte	7583	0	0.000	8300	41	0.005
vascular	77	0	0.000	97	0	0.000

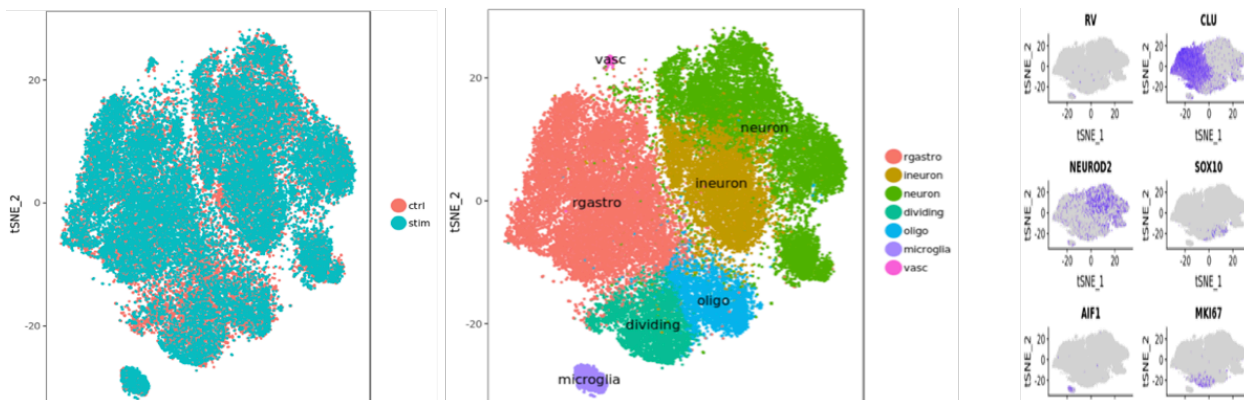


Figure 3.2 Single cell RNA sequencing of Rubella virus (RV)- and mock-infected organotypic slice culture.

RV RNA is predominantly found in microglia (AIF1/Iba1+).

for this poly-adenylated transcript, they were predominantly found in microglia (Figure 3.2, Table 3.1). Relative to microglia without viral RNA in the same inoculated sample, microglia with detectable rubella RNA showed upregulated SOD2 and downregulated SPP1, suggesting putative responses specific to microglia that are actively infected. Major caveats to this pilot experiment include the lack of a heat-killed virus control (only a mock was included), and the absence of spike-in cells (e.g. a different species) to detect ambient viral transcripts recovered in droplets vs. true intracellular RNA. Nonetheless, together with the immunohistochemistry experiments, these provide modest evidence for rubella virus tropism for microglia along with a smaller contribution of other cell types in the human developing brain.

Viral targeting of microglia depends on context of other cell types

Given the apparent targeting of microglia observed in slice culture, where all cell types of the developing brain are present apart from infiltrating immune cells, we next attempted to develop a culturing system based on dissociated primary cells from the same tissue. In developing this system, we observed that rubella virus was very rarely found in a highly enriched microglia population, whereas much greater immunostaining for RV antigens was observed in microglia that were in contact with or shared media with other cell types from the same tissue (Figure 3.3, data from Galina Schmunk). Notably, even in the absence of rubella virus, the presence of other cell types may modify the microglia phenotype, perhaps making microglia more susceptible to virus or more activated and likely to take up the virus particles. Future experiments may define which other cell types

are most important for this phenotype (neurons, astrocytes, radial glia, oligodendrocyte precursors, intermediate progenitor cells, vascular cells etc.), and which components of media conditioned by these cell types are relevant to the increase of RV immunoreactivity in microglia.

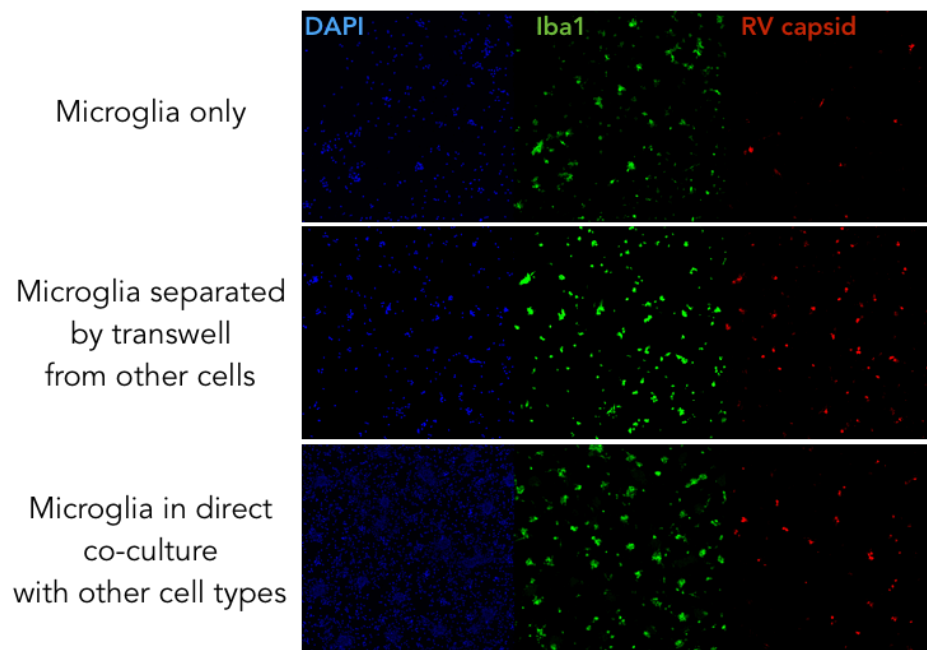


Figure 3.3 Rubella capsid in microglia depends on presence of other cell types

RV capsid immunostaining in microglia (Iba1+ cells) is rare in a purified population of microglia (top row), but highly prevalent when microglia are in direct contact with other cell types including neurons and glia (bottom row), or separated by a transwell from other cell types with free flow of media (middle row). Microglia were isolated from a GW18 cortical tissue using magnetic beads expressing anti-Iba1, then cultured with or without other cell types from the same tissue in direct contact or in a transwell. Cells were inoculated with RV (MOI ~1), then fixed at 72 hrs post-inoculation and immunostained for Iba1 and RVcapsid. Experiments and data analysis performed by Galina Schmunk (Nowakowski Lab, UCSF).

Characterization of infection

Given the observations of RV capsid protein and RNA in microglia after inoculating slice culture or dissociated cells, we next asked whether a productive infection (replication of viral RNA/genomes and proteins, with packaging and release of infectious virions) had

taken place. In the tissue and in the supernatant after inoculating slice culture with RV, no detectable increase in infectious RV could be observed out to 7 days post-inoculation (Figure 3.4, Figure 3.5). Likewise after inoculating microglia in co-culture, compared to the robust viral growth curve in Vero cells, little-to-no production could be measured above the baseline of the inoculum (Figure 3.6).

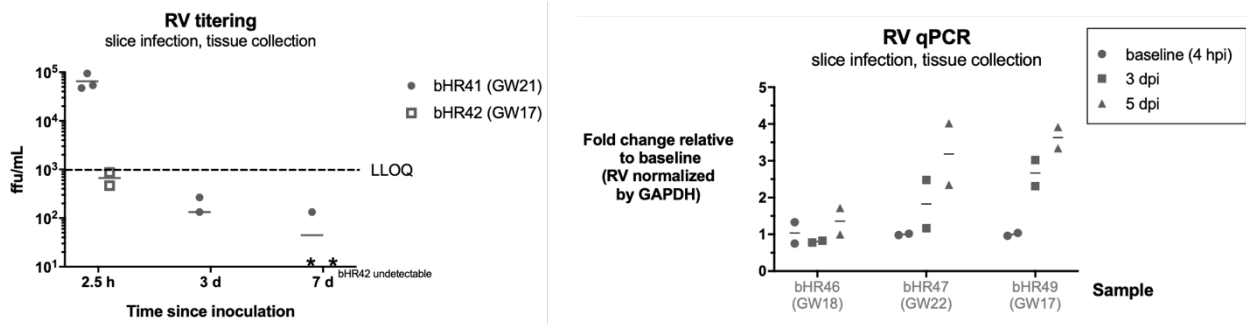


Figure 3.4 Titering and qPCR for RV in tissue at multiple days after inoculation of slice culture.

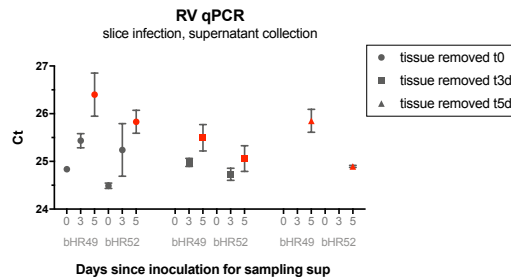


Figure 3.5 qPCR for RV in supernatant at multiple days after inoculation of slice culture

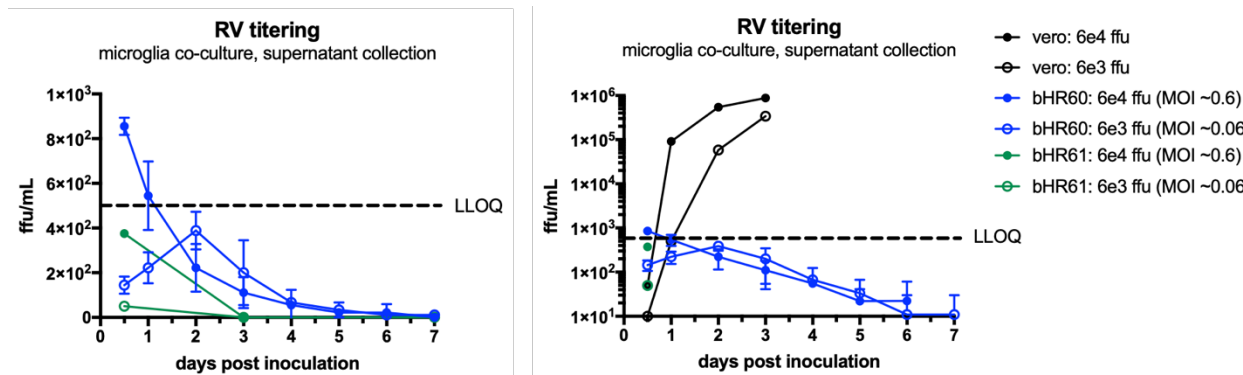


Figure 3.6 Titering for RV in supernatant at multiple days after inoculation of Vero cells or primary microglia in co-culture.

Turning to the viral RNA, we next asked whether negative strand virus was observed, which would indicate viral replication intracellularly. Validation of the FISH probes specific to negative strand (-) RV in Vero cells showed that only very robustly infected Vero cells had an increase in negative strand viral RNA detectable by this method (Figure 3.7, Figure 3.8). In co-cultured microglia, RV (+) strand RNA clearly accumulated intracellularly, but no difference in RV (-) strand was detectable compared to a heat-killed virus control (Figure 3.9, Figure 3.10). Overall, these data from titering, qPCR, and FISH suggest that if viral replication and productive infection occurs in microglia, it is not extensive. Given the technical limitations of sensitivity in these assays, confirmation of viral replication could be achieved using an engineered RV strain that expresses a marker, only when viral replication occurs, such as the fluorophore- p150 fusion strains (Sakata et al. 2014; Matthews and Frey 2012). I have generated a GFP strain of the M33 rubella virus (Figure 3.11), which can be used to test viral replication in microglia and other cell types.

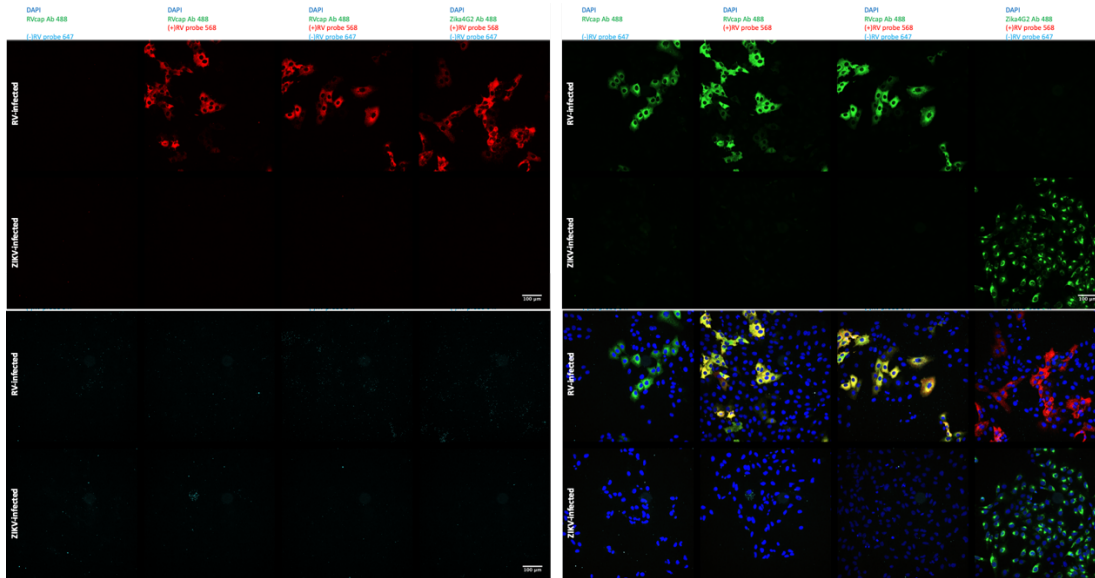


Figure 3.7 Validation of FISH probes and anti-RV capsid antibody for detection of rubella virus.

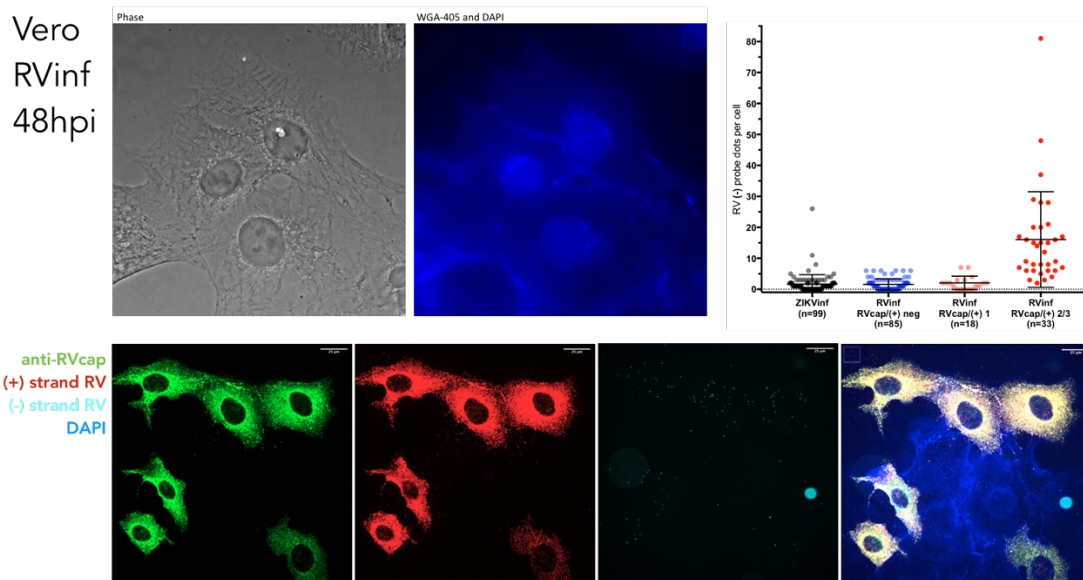


Figure 3.8 Validation of negative strand FISH on RV-infected Vero cells.

Dots for each probe were counted within the cell boundaries, as defined by wheat germ agglutinate-405 fluorescent dye, of RV-infected cells, with Zika virus infection as a control. nCounts displayed separately for cells that were weakly (1) or strongly (2/3) immunostained for RV capsid.

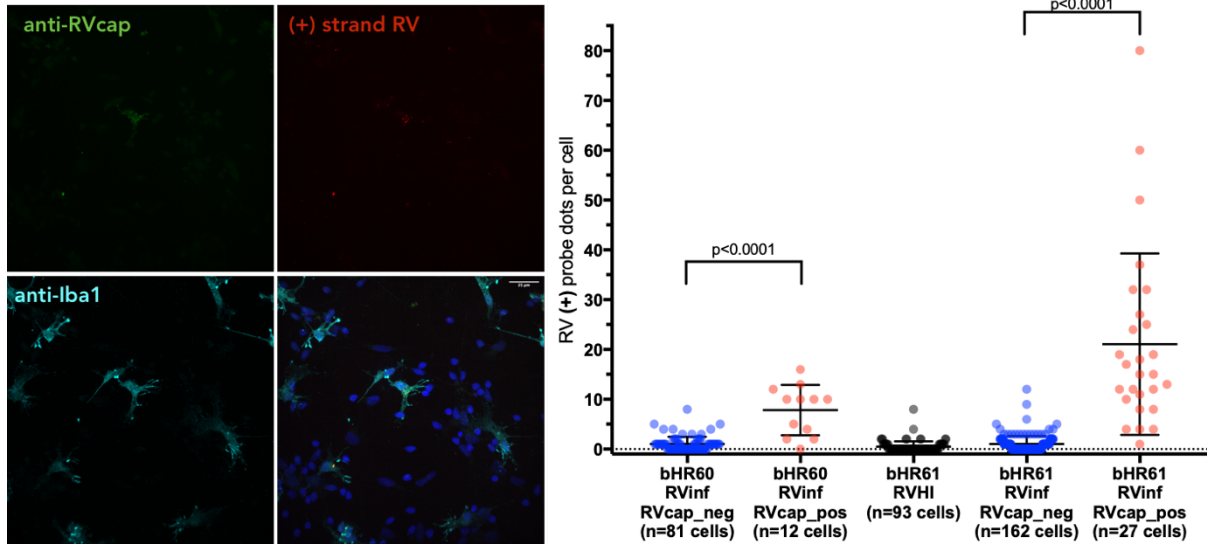


Figure 3.9 RV (+) strand accumulates in cells with anti-RV capsid.

RVHI = heat-inactivated rubella virus. RVinf = infected with rubella virus. RVcap = RV capsid.

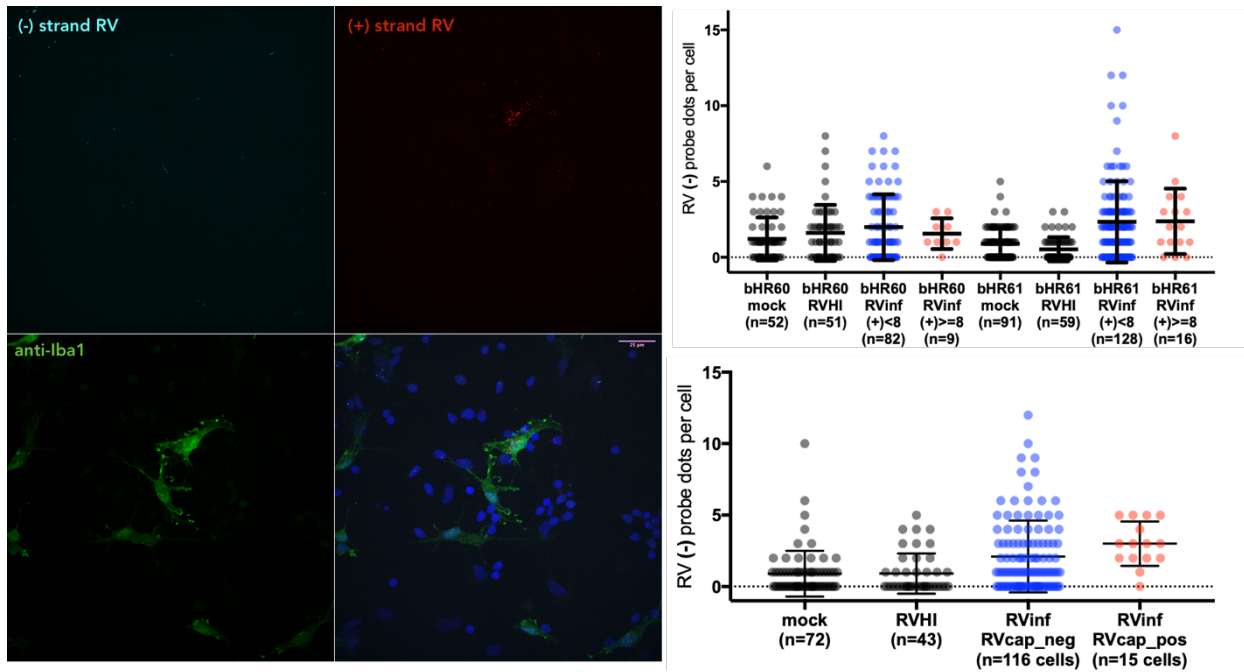


Figure 3.10 Assessment of RV (-) strand RNA in microglia by FISH.

No detectable difference in RV (-) strand RNA in microglia with or without anti-RVcapsid staining, or with or without RV (+) strand RNA. Mock = mock-infected. RVHI = heat-inactivated rubella virus. RVinf = infected with rubella virus. RVcap = RV capsid.

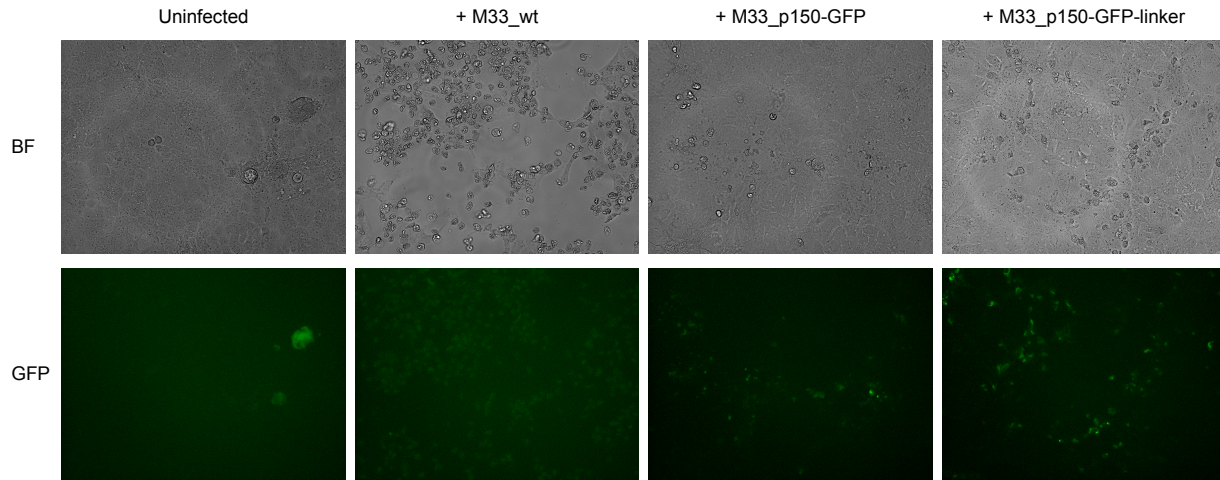


Figure 3.11 RV M33 strain with inserted GFP, expressed as fusion to p150.

Vero cells shown at 72h after inoculation with P1 of IVT-RNA launched viruses. Substantial cytopathic effect apparent in wildtype M33, and modest cytopathic effect in M33_p150-GFP-linker with expression of GFP clearly visible.

Future directions: consequences of viral infection for development

The findings above raise several questions for follow-up. Is microglia activation required for infection? What components of the conditioned media make microglia susceptible? Ideally, the tropism of the virus for microglia could be assessed if tissue were available from natural infections in congenital rubella syndrome cases. Having defined tropism, one might ask why microglia rather than other cell types? Putative entry factors such as sphingomyelin, cholesterol, and MOG (Otsuki et al. 2017; Cong, Jiang, and Tien 2011), are not specific to microglia at the transcriptional level in our single cell sequencing data. One could test whether inhibition of the sphingomyelin pathway affects susceptibility, or more broadly, screen for host dependency factors for rubella in microglia to identify putative cell-type specific factors.

In addition, the consequences of infection are important. I observe a strong transcriptional change in microglia and other cell types that is most similar to response to type I interferons, as discussed in Chapter 2. How do those transcriptional changes affect microglia morphology and behavior in a slice or organoid or animal model? One might activate or infect primary human microglia then transplant into a mouse to model such a scenario. Importantly, what are the lasting consequences of these changes? Perhaps what we learn from microglia targeted by rubella virus could help us understand mechanisms when development goes awry in the setting of other viral infections, or in genetic syndromes that prevent normal regulation of interferon responses.

References for Chapter 3

- Banatvala, J., and Brown, D. (2004). Rubella. *The Lancet* 363, 1127–1137.
- Cong, H., Jiang, Y., and Tien, P. (2011). Identification of the Myelin Oligodendrocyte Glycoprotein as a Cellular Receptor for Rubella Virus. *Journal of Virology* 85, 11038–11047.
- Esterly, J.R., and Oppenheimer, E.H. (1967). Vascular Lesions in Infants with Congenital Rubella. *Circulation* 36, 544–554.
- Grant, G.B., Reef, S.E., Patel, M., Knapp, J.K., and Dabbagh, A. (2017). Progress in Rubella and Congenital Rubella Syndrome Control and Elimination — Worldwide, 2000 – 2016. *MMWR Morb. Mortal. Wkly. Rep.* 66, 1256–1260.
- Gregg, N.M. (1941). Congenital Cataract Following German Measles in the Mother. In *Problems of Birth Defects*, T.V.N. Persaud, ed. (Dordrecht: Springer Netherlands), pp. 170–180.
- Korones, S.B., Ainger, L.E., Monif, G.R.G., Roane, J., Sever, J.L., and Fuste, F. (1965). Congenital rubella syndrome: New clinical aspects with recovery of virus from affected infants. *The Journal of Pediatrics* 67, 166–181.
- Lazar, M., Abernathy, E., Chen, M., Icenogle, J., Janta, D., Stanescu, A., Pistol, A., Santibanez, S., Mankertz, A., Hübschen, J.M., et al. (2016). Epidemiological and molecular investigation of a rubella outbreak, Romania, 2011 to 2012. *Euro Surveill.* 21, 30345.

- Matthews, J.D., and Frey, T.K. (2012). Analysis of subcellular G3BP redistribution during rubella virus infection. *Journal of General Virology* 93, 267–274.
- Monif, Gilles R.G., Avery, Gordon B., Korones, Sheldon B., and Sever, John L. (1965). POSTMORTEM ISOLATION OF RUBELLA VIRUS FROM THREE CHILDREN WITH RUBELLA-SYNDROME DEFECTS. *The Lancet* 285, 723–724.
- Nguyen, T.V., Pham, V.H., and Abe, K. (2015). Pathogenesis of Congenital Rubella Virus Infection in Human Fetuses: Viral Infection in the Ciliary Body Could Play an Important Role in Cataractogenesis. *EBioMedicine* 2, 59–63.
- Otsuki, N., Sakata, M., Saito, K., Okamoto, K., Mori, Y., Hanada, K., and Takeda, M. (2017). Both Sphingomyelin and Cholesterol in the Host Cell Membrane Are Essential for Rubella Virus Entry. *J Virol* 92, e01130-17.
- Plotkin, S.A., Reef, S.E., Cooper, L.Z., and Alford, C.A. (2011). Rubella. In *Infectious Diseases of the Fetus and Newborn*, (Elsevier), pp. 861–898.
- Sakata, M., Otsuki, N., Okamoto, K., Anraku, M., Nagai, M., Takeda, M., and Mori, Y. (2014). Short Self-Interacting N-Terminal Region of Rubella Virus Capsid Protein Is Essential for Cooperative Actions of Capsid and Nonstructural p150 Proteins. *Journal of Virology* 88, 11187–11198.
- Vynnycky, E., Adams, E.J., Cutts, F.T., Reef, S.E., Navar, A.M., Simons, E., Yoshida, L.-M., Brown, D.W.J., Jackson, C., Strebel, P.M., et al. (2016). Using Seroprevalence and Immunisation Coverage Data to Estimate the Global Burden of Congenital Rubella Syndrome, 1996-2010: A Systematic Review. *PLoS ONE* 11, e0149160.

Chapter 4 Exploratory proteomic analysis implicates the alternative complement cascade in Primary CNS Vasculitis

Authors:

Caleigh Mandel-Brehm PhD*¹, Hanna Retallack AB*¹, Giselle M. Knudsen PhD², Alex Yamana², Rula A. Hajj-Ali MD³, Leonard H. Calabrese DO³, Tarik Tihan MD PhD⁴, Hannah A. Sample BS¹, Kelsey C. Zorn MHS¹, Mark P. Gorman MD⁵, Jennifer Madan Cohen MD⁶, Antoine G. Sreih MD⁷, Jacqueline F. Marcus MD⁸, S. Andrew Josephson MD^{9,10}, Vanja C. Douglas MD^{9,10}, Jeffrey M. Gelfand MD^{9,10}, Michael R. Wilson MD^{9,10}, Joseph L. DeRisi PhD^{1,11}

Affiliations:

¹Department of Biochemistry and Biophysics, University of California, San Francisco, CA, USA

²Department of Pharmaceutical Chemistry, University of California, San Francisco, CA, USA

³Department of Rheumatology/Immunology, Cleveland Clinic, Cleveland, OH, USA

⁴Department of Pathology and Laboratory Medicine, University of California, San Francisco, CA, USA

⁵Department of Neurology, Boston Children's Hospital, Boston, MA, USA

⁶Division of Neurology, Connecticut Children's Medical Center, Hartford, CT, USA

⁷Division of Rheumatology, University of Pennsylvania, Philadelphia, PA, USA

⁸Kaiser Permanente, San Francisco Medical Center, San Francisco, CA, USA

⁹UCSF Weill Institute for Neurosciences, San Francisco, CA, USA

¹⁰Department of Neurology, University of California, San Francisco, CA, USA

¹¹Chan Zuckerberg Biohub, San Francisco, CA, USA

Includes material previously published in:

Mandel-Brehm C*, Retallack H*, Knudsen GM, et al. Exploratory proteomic analysis implicates the alternative complement cascade in primary CNS vasculitis. *Neurology*. 2019;93(5):e433-e444. doi:10.1212/WNL.0000000000007850

Abstract

Objective: To identify molecular correlates of primary angiitis of the central nervous system (PACNS) through proteomic analysis of cerebrospinal fluid (CSF) from a biopsy-proven patient cohort.

Methods: Using mass spectrometry, the CSF proteome of biopsy-proven PACNS patients (n=8) was quantitatively compared to CSF from individuals with non-inflammatory conditions (n=11). Significantly enriched molecular pathways were identified using a gene ontology workflow, and high confidence hits within enriched pathways (fold change >1.5 and concordant Benjamini-Hochberg-adjusted p-value <0.05 on DESeq and t-test) were identified as differentially regulated proteins.

Results: Compared to non-inflammatory controls, 283 proteins were differentially expressed in PACNS patient CSF, with significant enrichment of the complement cascade pathway (C4-binding protein, CD55, CD59, properdin, complement C5, complement C8 and complement C9) and neural cell adhesion molecules. A subset of clinically relevant findings was validated by western blot and commercial ELISA.

Conclusions: In this exploratory study we found evidence of deregulation of the alternative complement cascade in CSF from biopsy-proven PACNS as compared to non-inflammatory controls. More specifically, several regulators of the C3 and C5 convertases and components of the terminal cascade were significantly altered. These preliminary findings shed light on a previously unappreciated similarity between PACNS and systemic vasculitides, especially Anti-Neutrophil Cytoplasmic Antibody (ANCA)-associated

vasculitis. The therapeutic implications of this common biology, and the diagnostic and/or therapeutic utility of individual proteomic findings warrant validation in larger cohorts.

Introduction

Primary angiitis of the central nervous system (PACNS) is a severe inflammatory disease affecting the blood vessel walls in the brain, spinal cord and meninges (Hajj-Ali and Calabrese 2013). Without treatment, PACNS is frequently progressive (Salvarani et al. 2015) (Byram, Hajj-Ali, and Calabrese 2018). Although broadly acting immunosuppressants prevent mortality in ~80% of patients, these medications have adverse side effects and nearly half of patients relapse with debilitating neurological symptoms (Hutchinson et al. 2010b). The basis for the variable response to therapy is unknown and cannot be reconciled with clinical data alone (Byram, Hajj-Ali, and Calabrese 2018). The lack of molecular tools to aid in the clinical investigation of PACNS is an obstacle to improving patient outcomes.

Due to the low prevalence of PACNS and lack of available mouse models, the molecular pathogenesis in PACNS is poorly understood (Alba et al. 2011). To elucidate molecular correlates, a proteomic survey of patient CSF using mass spectrometry is a compelling approach. CSF is an accessible biologic fluid that circulates throughout the meninges and parameningeal structures of the brain and spinal cord. Molecular analysis of CSF can provide diagnostic information regarding disease pathologies occurring within these regions (Hajj-Ali and Calabrese 2013; Bastos et al. 2017). CSF abnormalities, including increased protein content, are observed in ~90% of PACNS patients, and

reversal of CSF abnormalities correlate with improved patient outcomes (Oliveira et al. 1994). Previous attempts to characterize molecular abnormalities in PACNS CSF have been made, but without the benefit of biopsy-proven disease (Ruland et al. 2018).

Here, we perform an unbiased proteomic analysis comparing the CSF profiles of biopsy-proven PACNS patients to those of non-inflammatory controls (NIC) and controls with reversible cerebral vasoconstriction syndrome (RCVS). Our goal was to identify candidate proteins and molecular pathways involved in the chronic inflammatory pathophysiology of PACNS and to highlight molecular targets for future therapeutic and diagnostic studies.

Methods

Full details regarding molecular protocols, including mass spectrometry and orthogonal validations, can be found in Appendix e-1.

Patient recruitment and study protocol.

PACNS and NIC patients were recruited as part of a larger study analyzing biological samples from patients with suspected neuroinflammatory disease at UCSF. The UCSF Institutional Review Board (IRB) approved the study protocol, and participants or their surrogates provided written informed consent. RCVS controls were recruited as part of a larger study analyzing biological samples from patients with CNS vascular disorders at Cleveland Clinic. The Cleveland Clinic IRB approved the study protocol, and participants or their surrogates provided written informed consent. PACNS and RCVS patients were

diagnosed according to standard clinical diagnostic criteria, including neuropathology evaluation in all of the patients diagnosed with PACNS.

PACNS Clinical Vignettes

Patient 1

A previously healthy 50-year-old woman developed mild headaches with episodic, migrainous features including visual auras together with worsening "mental fogging" independent of the headaches, all of which worsened over 4 years. At that time, she was discovered to have a thoracic myelopathy on exam and inflammatory CSF of unclear etiology. On magnetic resonance imaging (MRI), she was found to have a thoracic myelitis with nodular leptomeningeal enhancement throughout the spine and the brain. An MR angiogram of the head and neck was unremarkable. A brain biopsy revealed evidence for a small vessel vasculitis and chronic meningitis. An extensive work-up for neoplastic, infectious and other autoimmune etiologies was unrevealing. The CSF profile from a sample from later in the patient's clinical course was used for this study and showed a white blood cell (WBC) count of 2 cells/uL (66% lymphocytes, 34% monocytes) (0-5 cells/uL), red blood cell (RBC) count of 0 cells/uL (0-5 cells/uL), glucose 45 mg/dL (45-80 mg/dL), total protein 192 mg/dL (15-45 mg/dL), IgG index 1.8 (<0.6) and greater than 5 unique oligoclonal bands (OCBs) (≤ 1 band).

Patient 2

A 39-year-old woman with a history of ulcerative colitis well controlled on mesalamine and oral budesonide developed increasing fatigue and increasingly painful, new left-sided headaches and left facial paresthesias over 6 weeks to the point that they

prompted hospitalization. A brain MRI revealed T2 hyperintensities in a gyriform pattern over the left parietal and temporal lobes with associated leptomeningeal enhancement. A CSF examination revealed a WBC count of 15 cells/uL (54% lymphocytes, 39% granulocytes, 7% other), an RBC count of 0 cells/uL, glucose 65 mg/dL, total protein 50 mg/dL and 0 OCBs. A CT angiogram of the head and neck was unremarkable except for the “suggestion of mild smooth narrowing of the left carotid artery terminus and left M1 segment of the middle cerebral artery”. Over the next few days, the patient developed aphasia and apraxia that prompted a brain biopsy which revealed a small vessel vasculitis. Extensive work-up for neoplastic, infectious and other autoimmune etiologies was unrevealing.

Patient 3

A 56-year-old man was hospitalized for rapid cognitive decline and was also found to have bilateral papilledema and a left abducens nerve palsy on exam. He had extensive confluent white matter T2 hyperintensities and multiple small areas of restricted diffusion consistent with acute infarcts on brain MRI. An MR angiogram of the head and neck was unremarkable. A CSF examination showed a WBC count of 26 cells/uL (53% neutrophils, 39% lymphocytes, 8% monocytes), RBC count of 3,275 cells/uL, glucose 60 mg/dL, total protein 141 mg/dL and an IgG index of 0.9. An extra-ventricular drain was placed for elevated intracranial pressure, and a brain biopsy revealed a small vessel vasculitis. Extensive work-up for neoplastic, infectious and other autoimmune etiologies was unrevealing.

Patient 4

A 55-year-old woman with a history of non-insulin dependent diabetes mellitus had progressive difficulty walking over 6 months before she acutely lost sensation in her right leg and developed severe urinary retention. She was found to have a longitudinally extensive transverse myelitis on MRI and a CSF examination that revealed a WBC count of 8 cells/uL, an RBC count of 1 cell/uL, glucose 92 mg/dL, total protein 99 mg/dL, IgG index 0.57, and 1 unique OCB. Despite initial attempts at immunosuppression with glucocorticoids, the patient developed new weakness in her left leg and urinary and fecal incontinence. Serial imaging revealed new inflammatory lesions in the cerebellum and overlying leptomeninges. All vascular imaging including a cerebral and spinal angiogram was unremarkable. A brain biopsy revealed a small vessel vasculitis. Extensive work-up for neoplastic, infectious and other autoimmune etiologies was unrevealing.

Patient 5

A 36-year-old man with a history of possible relapsing polychondritis, with one episode of ear chondritis, presented with a few weeks of new onset daily headaches and fatigue followed by visual hallucinations that prompted a neurologic evaluation. A CSF examination revealed a WBC count of 6 cells/uL (60% lymphocytes, 22% monocytes, 18% neutrophils), an RBC count of 3 cells/uL, glucose 51 mg/dL, total protein 27 mg/dL, and an IgG index of 1.62. A brain MRI revealed patchy leptomeningeal enhancement, multiple areas of T2 hyperintensity with patchy gadolinium enhancement and multifocal areas of restricted diffusion consistent with acute infarcts. An MR angiogram of the head was normal, but a cerebral angiogram showed diffuse vasculopathy bilaterally in the distal

vasculature (i.e., M3, M4, ophthalmic arteries, P3 and P4 vessels). A brain biopsy showed small vessel vasculitis. Extensive work-up for neoplastic, infectious and other autoimmune etiologies was unrevealing.

Patient 6

A previously healthy 57-year-old woman presented with 2 months of new onset headaches with visual aura, 10 days of dizziness and vertigo and an isolated episode of hemi-body sensory symptoms who was found to have a subarachnoid T2 hyperintensity on brain MRI and faint leptomeningeal enhancement. A CT angiogram of the head and neck was normal. CSF examination revealed a WBC count of 31 cells/uL (90% lymphocytes, 6% monocytes, 2% neutrophils, 1% eosinophils and 1% unidentified), an RBC count of 102 cells/uL, glucose 54 mg/dL, total protein 75 mg/dL, IgG index 2.02 and more than 2 unique OCBs. and a brain biopsy revealed a small vessel vasculitis. Extensive work-up for neoplastic, infectious and other autoimmune etiologies was unrevealing.

Patient 7

A previously healthy 13-year-old girl presented with fever, headache, altered mental status and seizure and was found to have uni-hemispheric, subcortical T2 hyperintense lesions on brain MRI, many of which enhanced with gadolinium. An MR angiogram of the head and neck was unremarkable. A CSF exam revealed a WBC count of 11 cells/uL (84% lymphocytes, 16% monocytes), an RBC count of 6 cells/uL, glucose 49 mg/dL, total protein 37 mg/dL and more than 5 unique OCBs. Brain biopsy revealed a small vessel vasculitis. Extensive work-up for neoplastic, infectious and other autoimmune etiologies was unrevealing.

Patient 8

A 51-year-old man with a history of atrial fibrillation, hypertension and seizure presented with bony aches and migratory joint pains that went away and were followed months later with bilateral episcleritis, fever, numbness in his feet and confusion. A brain MRI showed leptomeningeal enhancement and overlying multifocal areas of swollen and T2 hyperintense cortex and possible subcortical U-fiber enhancement. An MR angiogram of the head and neck was unremarkable. CSF examination showed revealed a WBC count of 3 cells/uL (82% lymphocytes, 12% monocytes, 6% neutrophils), an RBC count of 1 cell/uL, glucose 59 mg/dL, total protein 42 mg/dL, IgG index 0.7 and 5 unique OCBs. The clinical impression by the treating neurologist was that the systemic symptoms were unrelated to the patient's neuroinflammatory disease. A brain biopsy revealed a small vessel vasculitis. Extensive work-up for neoplastic, infectious and other autoimmune etiologies was unrevealing.

RCVS Clinical Vignettes**Patient 9**

A 54-year-old woman with a history of hepatitis C virus infection and epilepsy, presented with altered mental status, dysarthria, expressive aphasia, and left greater than right-sided weakness. An MR angiogram of the head revealed irregularities of the right anterior cerebral artery and left internal carotid artery. A cerebral angiogram showed narrowing and beading in multiple vessels including the basilar artery, bilateral posterior cerebral arteries, and the left M1 and A1 segments. There was also focal beading in the distal left anterior and middle cerebral arteries and an aneurysm at the origin of the right

temporal artery. Her CSF profile revealed a WBC count of 11 cells/uL (93% lymphocytes, 3% monocytes, 4% other), an RBC count of 340 cells/uL, glucose 57 mg/dL, total protein 22 mg/dL and no OCBs. She started on calcium channel blockers and repeated MRA of intracranial vessel 10 days later with marked improvement in the intracranial vessel abnormalities.

Patient 10

A 33-year-old woman presented with a sudden-onset, thunderclap headache that was clearly different from her typical migraine headaches. A CT angiogram of the head a non-contrast head CT revealed diffuse beading of the vasculature throughout the anterior and posterior circulation, and a parietal subarachnoid hemorrhage, respectively. A cerebral angiogram similarly found segmental irregularities of the intracranial vessels of the distal left anterior circulation. Her CSF profiled revealed a WBC count of 1 cell/uL (82% lymphocytes, 8% monocytes, 10% other), an RBC count of 1,150 cells/uL, glucose 65 mg/dL, total protein 115 mg/dL and an IgG index of 1.0. The patient was started on calcium channel blockers and had a rapid resolution of her symptoms and no disease recurrence.

Patient 11

A 57-year-old woman with a history of depression treated with citalopram and bupropion hydrochloride presented with 4 days of dizziness, lightheadedness, left greater than right leg weakness and falls followed by a rapid decline in mental status was found to have large areas of restricted diffusion in the bilateral parietal lobes on brain MRI. She became unresponsive and was intubated and transferred to the intensive care unit. A CT

angiogram of the head revealed irregularities in the distal portions of the anterior cerebral arteries, and a cerebral angiogram was similarly consistent with vasospasm. A CSF exam revealed a WBC count of 1 cell/uL, an RBC count of 29 cells/uL, glucose 95 mg/dL, and total protein 22 mg/dL. The patient improved clinically and radiologically after administration of intra-arterial nicardipine and verapamil. She was started on a calcium channel blocker, and citalopram and bupropion hydrochloride were discontinued.

Patient 12

A previously healthy 30-year-old woman was admitted with new left-sided weakness and severe hypothermia after experiencing new onset, recurrent thunderclap headaches for 2 weeks. A CT angiogram of the head and neck showed multiple foci of intracranial vascular narrowing in the bilateral anterior cerebral arteries, middle cerebral arteries and posterior cerebral arteries which was corroborated by a cerebral angiogram. A brain MRI showed acute bilateral subcortical infarcts. A high resolution brain MRI found no evidence of abnormal vessel wall enhancement. Her CSF exam revealed a WBC count of 0 cells/uL, an RBC count of 133 cells/uL, glucose 78 mg/dL, and a total protein of 33 mg/dL. She was started on calcium channel blockers and improved clinically and radiographically.

Mass Spectrometry

Total protein concentration in patient CSF was determined to be 0.1–0.6 mg/ml by Bradford assay (Sigma, B6916). A total of five micrograms of protein was used from each patient CSF sample for LC-MS/MS analysis. See Supplementary Appendix e-1 (data available from Dryad) for full details regarding sample processing and data acquisition.

Statistical Analyses

The following statistical analyses were performed in R v.3.4.1. For comparative analyses of the individual CSF proteomes, spectral counts were aggregated by protein (i.e., protein abundance) for each sample. The protein abundances for the unique 1,043 proteins identified in CSF were compared between PACNS (n= 8) and NIC (n= 11) cohorts using two statistical approaches commonly used for mass spectrometry datasets, DESeq2 v.3.7 and *t*-test^{9,10}. DESeq2: this package utilizes a method based on the negative binomial distribution to assess differential expression in count data. Spectral counts for all 1,043 unique proteins were used as the input for this package, which was then run with default settings. *T*-test: Spectral count values of zero were first replaced with counts of 0.16, a value empirically determined to best approximate normal distributions for each protein within the NIC samples. Spectral counts were then divided by the sum of spectral counts for each sample and multiplied by 10,000, generating “normalized” spectral counts. Only proteins at sufficient abundance were considered in the *t*-test, defined as having a sum across all NIC and PACNS samples of ≥ 10 normalized spectral counts, and being observed in at least five of the combined group of NIC and PACNS samples. Normalized spectral counts for these 713 abundant proteins were then transformed by natural logarithm to avoid large differences in variances for different proteins, and then analyzed using two-sided *t*-tests assuming unequal variance to compare abundances between PACNS and NIC cohorts for each protein. The resulting *p*-values were then adjusted for multiple comparisons using the Benjamini-Hochberg (BH) method.

Table 4.1 Demographics and clinic features of the PACNS, RCVS, and NIC cohorts

	PACNS (n=8) ID 1-8	RCVS (n=4) ID 9-12	NIC (n=11) ID 13-23
Age, median (range)	51 (13-57)	44 (30-57)	48 (31-62)
Female, n (%)	5 (63)	4 (100)	8 (73)
Immunosuppression at time of CSF sampling ^a , n (%)	5 (63)	1 (25)	0
Brain imaging, n (%)			
Angiographic abnormality ^b	2 (25)	4 (100)	–
Abnormal leptomeningeal enhancement on MRI	7 (88)	0	–
CSF parameters, median (range)			
White blood cell count, cells/mm ³	10 (2-31)	1 (0-11)	–
Protein level, mg/dL	63 (27-192)	28 (22-115)	–
Glucose, mg/dL	57 (45-92)	72 (57-95)	–
IgG index ^c	1.26 (0.57-2.02)	–	–
≥2 CNS-specific OCBs ^c , n (%)	4 (67)	–	–
Biopsy of CNS blood vessels, n (%)			
Perivascular and intramural inflammation with vessel wall damage	8 (100)	–	–
Lymphocytic	8 (100)	–	–
Granulomatous	1 (13)	–	–
Small to medium vessels	8 (100)	–	–
Clinical course on follow-up, n (%)			
Monophasic, resolved	0	4 (100)	–
Active disease	3 (38)	0	–
In remission with immunosuppression	4 (50)	0	–
In remission, off immunosuppression	1 (13)	0	–

Abbreviations: IgG = immunoglobulin G; NIC = noninflammatory control; OCB = oligoclonal bands; PACNS = primary angiitis of the CNS; RCVS = reversible cerebral vasoconstriction; WBC = white blood cell.

^a Receiving immunosuppressive therapy at time of CSF sampling, including prednisone, budesonide, mesalamine, or cyclophosphamide.

^b On imaging modalities, including magnetic resonance angiography, CT angiography, or cerebral angiogram.

^c Calculated for 6 of 8 patients with PACNS in whom assay was performed. In patients with RCVS, the IgG index either was not assayed or was normal, and the OCB pattern either was not assessed or showed no CNS-specific OCBs.

Fold changes were calculated as the ratio between the mean of PACNS and NIC samples for each protein. In our conservative approach, only proteins that were significantly different (fold change > 1.5 and BH-adjusted p -value < 0.05) between PACNS and NIC in both tests were considered differentially regulated proteins in PACNS. Due to the low sample number in the RCVS cohort ($n = 4$) this comparative statistical analysis was restricted to the PACNS and NIC cohorts only.

Unbiased hierarchical clustering was performed using numpy (v1.15.0) and seaborn (v0.9.0) packages in python (v3.7.0). For this analysis, abundant proteins across PACNS, NIC, and RCVS samples were grouped to combine counts for isoforms of the same protein with shared peptides (peptides mapping to multiple isoforms). Proteins and samples were then clustered by applying an unweighted pair group method with arithmetic mean (UPGMA) to the correlation distance matrix.

Molecular Pathway Enrichment Analysis

Pathway Enrichment analysis was performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) Bioinformatics resource (<https://david.ncifcrf.gov/>). For enrichment analysis, the 222 downregulated proteins and 61 upregulated proteins in PACNS were analyzed against the CSF background of all 1,043 observed proteins. For 50 of the proteins most highly downregulated in PACNS, annotations for transmembrane and topological domains were retrieved from UniProt. The location of these domains, as well as the location of peptides observed in the NIC samples by mass spectrometry were then mapped onto the protein chains.

Data Availability

Raw mass spectrometry data files and peak list files have been deposited at ProteoSAFE (<http://massive.ucsd.edu>) with accession number MSV000082129.

Results

Summary of clinical characteristics

In addition to the individual patient vignettes above, an aggregated summary of patient demographics, clinical features, imaging abnormalities and clinical CSF parameters for the PACNS, NIC and RCVS cohorts is provided for comparison in

Table 4.1. The majority of our biopsy-proven PACNS patients were refractory to initial treatment strategies or notably, all PACNS patients had a chronic disease course, as compared to the monophasic nature of RCVS.

All eight PACNS patients had a brain biopsy with documented transmural inflammation and vessel wall damage of small to medium-sized vessels in the brain parenchyma and/or meninges (

Figure 4.1). A single patient's biopsy was described as having granulomatous vasculitis; the remainder were lymphocytic. Biopsies were not performed in RCVS or NIC individuals. No PACNS patient had evidence of amyloid angiitis or systemic vasculitis. All required continual immunosuppression to achieve remission. patients tested negative for anti-Neutrophil Cytoplasmic Antibody (ANCA)-associated vasculitis.

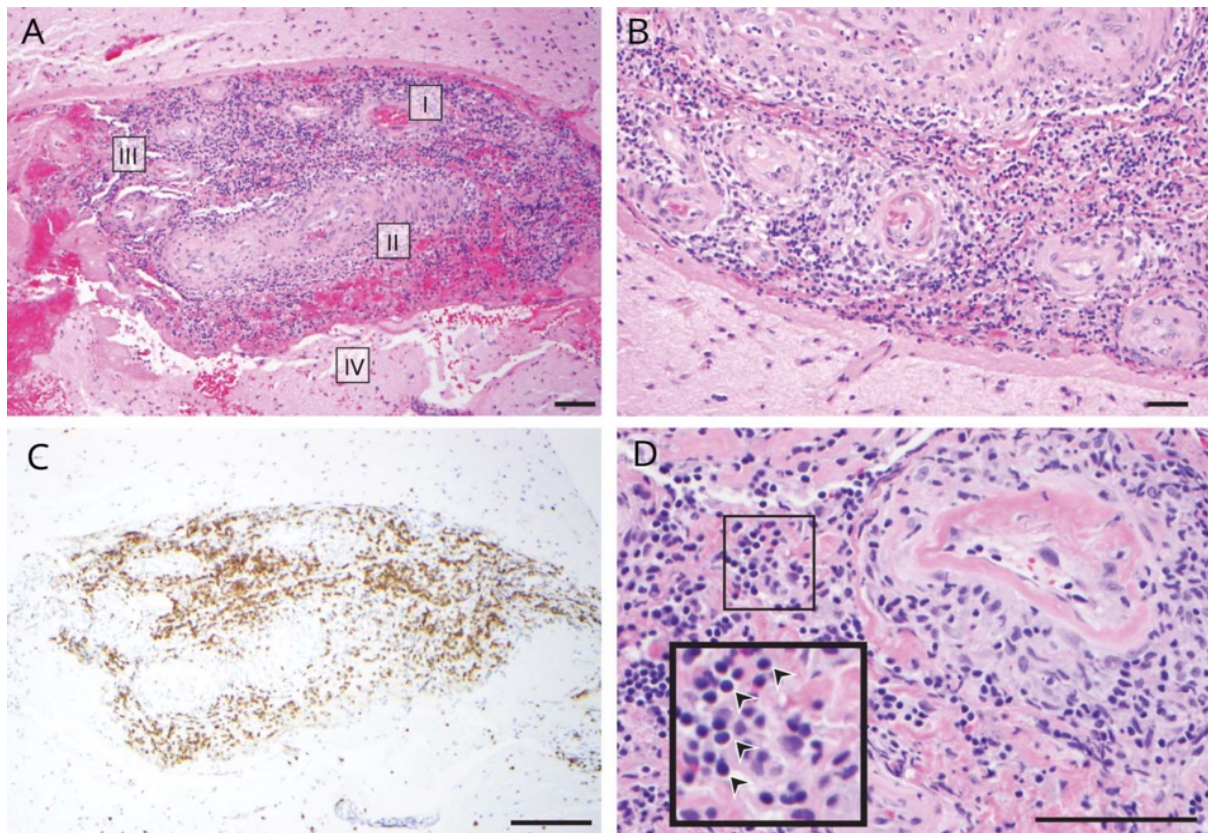


Figure 4.1 Demonstrative biopsies from patients with primary angiitis of the CNS (PACNS).

Biopsies from all patients in our primary angiitis of the CNS (PACNS) cohort ($n = 8$) showed inflammation of small to medium-sized CNS blood vessels. (A) Representative images from a patient biopsy (patient 3) demonstrating several hallmark features of PACNS histopathology, including (I) perivascular inflammation, (II) intra-mural inflammation with thickening of blood vessel wall, (III) leptomeningeal inflammation, and (IV) rupture of blood vessel wall. (B) Additional section of meningeal vessels (brain parenchyma, bottom left), showing transmural inflammation. (C) Immunohistochemistry with anti-CD3 shows enrichment of T cells among immune cell infiltrates corresponding to panel A. (D) High magnification of perivascular and intramural inflammation. Arrowheads in the inset indicate mononuclear lymphocytes. Tissue sections were stained with hematoxylin and eosin unless otherwise noted. Scale bar, $100 \mu m$.

Unbiased discovery of a putative molecular phenotype in PACNS

CSF samples from PACNS (n=8), RCVS (n=4), and NIC (n=11) individuals were analyzed by mass spectrometry. A total of 1,043 proteins were identified across all cohorts (Appendix e-2). Unbiased clustering of individuals based on the normalized protein counts showed that patients with PACNS were more similar to each other than to individuals in the RCVS and NIC cohorts (Figure 4.2).

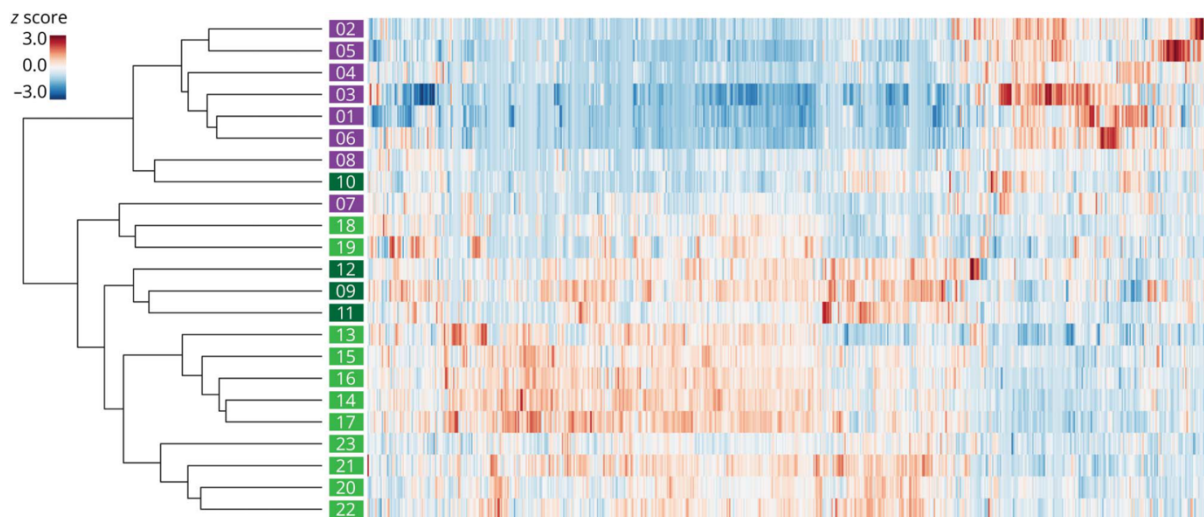


Figure 4.2 Unbiased clustering of patients by CSF proteome.

Dendrogram (left) depicts the distance between samples (sample identifiers shown at branch tips). Two major clusters emerge from an unbiased clustering analysis: top, containing mostly primary angiitis of the CNS (PACNS) samples (purple); and bottom, containing noninflammatory control (light green) and reversible cerebral vasoconstriction (dark green) samples. Heat map displays levels of 651 proteins for every sample as a by-protein zscore. Hierarchical clustering was performed by applying the unweighted pair group method with arithmetic mean method to the correlation distance matrix. Length of horizontal lines on the dendrogram represents the distance between samples.

Defining differentially expressed proteins and molecular pathways in PACNS

To identify specific proteins that distinguish PACNS from NIC, we compared protein abundances in PACNS versus NIC. We identified 283 proteins that had statistically significant differential regulation in PACNS, with 61 up-regulated and 222 down-regulated proteins compared to NIC (Appendix e-3). Pathway enrichment analysis showed significant enrichment for KEGG pathways “Complement and Coagulation Cascades” (adjusted p -value < 0.05) and “Cell Adhesion Molecules” (adjusted p -value < 0.01). Differentially expressed proteins in PACNS relative to NIC are shown in Figure 4.3. Although statistical analyses were not performed on the RCVS cohort due to the small number of patients, the relative protein abundances are displayed for qualitative comparison. Note that RCVS patients have elevated levels of serum-derived proteins (Hemoglobin, Serum Amyloid 1, Serum Amyloid 2, carbonic anhydrase) similar to PACNS, but show minimal evidence of complement protein dysregulation. There are many significantly altered proteins within the complement pathway in PACNS, including Decay Accelerating Factor (CD55), MAC-Inhibitory Protein (CD59), properdin (CFP), Complement C4 Binding Protein A (C4BPA), Complement C4 Binding Protein B (C4BPB), Ficolin-3 (FCN3), Carboxypeptidase N catalytic chain (CPN1), Carboxypeptidase N subunit 2 (CPN2), Complement C5 (C5), C8, and C9.

The enrichment for “Cell Adhesion Molecules” was driven by 22 proteins down-regulated in PACNS compared to NIC. More specifically, these proteins are transmembrane cell adhesion molecules expressed in neural tissue. Manual inspection of

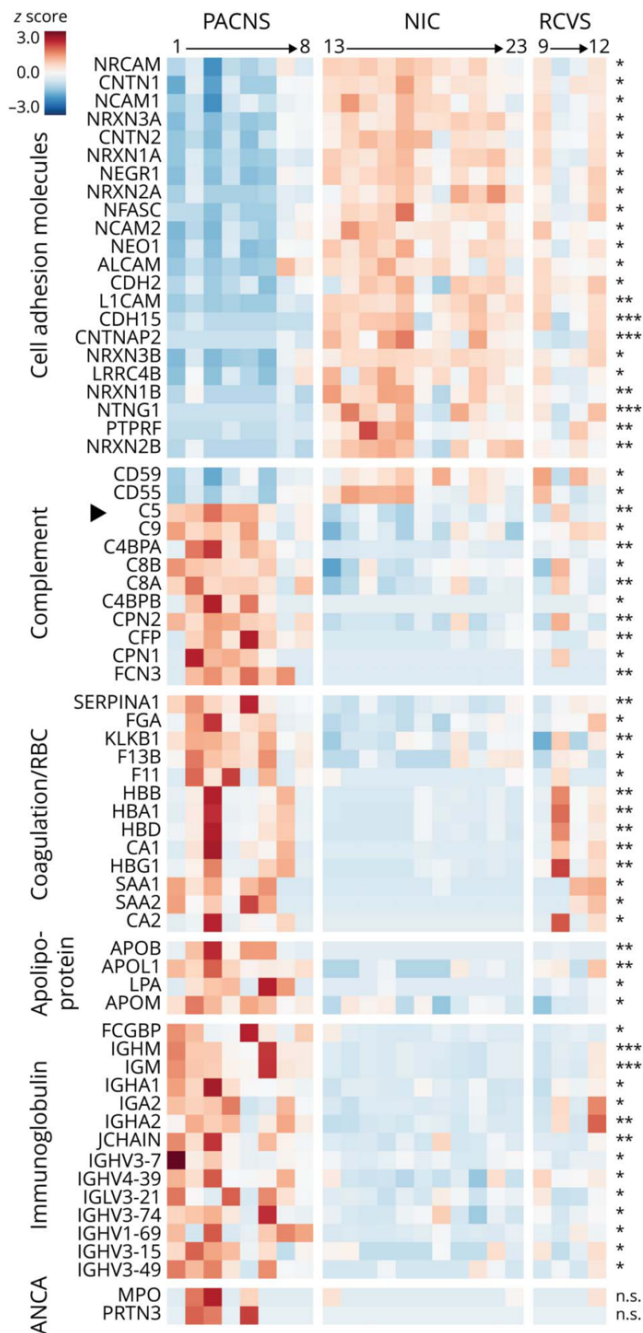


Figure 4.3 Comparison of discriminating proteins across cohorts.

Relative protein abundances for a subset of differentially regulated proteins in primary angiitis of the CNS (PACNS) vs noninflammatory control (NIC) are reported. Proteins representing the statistically significantly enriched pathways cell adhesion molecules and complement and coagulation cascades are included, as well as a subset of proteins manually curated to reflect findings of significant clinical interest, with functional classifiers informed by Database for Annotation, Visualization and Integrated Discovery and KEGG annotations. Heat map displays relative protein abundance across individual samples (PACNS $n = 8$, NIC $n = 11$, and reversible cerebral vasoconstriction [RCVS], $n = 4$), plotted as the zscore of the normalized spectral counts. Differential expression was evaluated with DESeq2 and the ttest (see Methods), with significance defined as having fold change > 1.5 and Benjamini-Hochberg-adjusted $p < 0.05$ for both tests (* $p < 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$). n.s. = not significant.

the remaining 200 down-regulated proteins revealed many more proteins that have roles in neural cell adhesion and contain transmembrane domains, including Amyloid Beta (APP), despite the absence of these proteins in the core set of the KEGG “Cell Adhesion Molecules” pathway. We localized the recovered peptides from mass spectrometry according to protein domain annotations assigned by UniProt and found that NIC CSF contained peptides specifically from the extracellular domains of transmembrane proteins which are down-regulated in PACNS CSF (Figure 4.4,

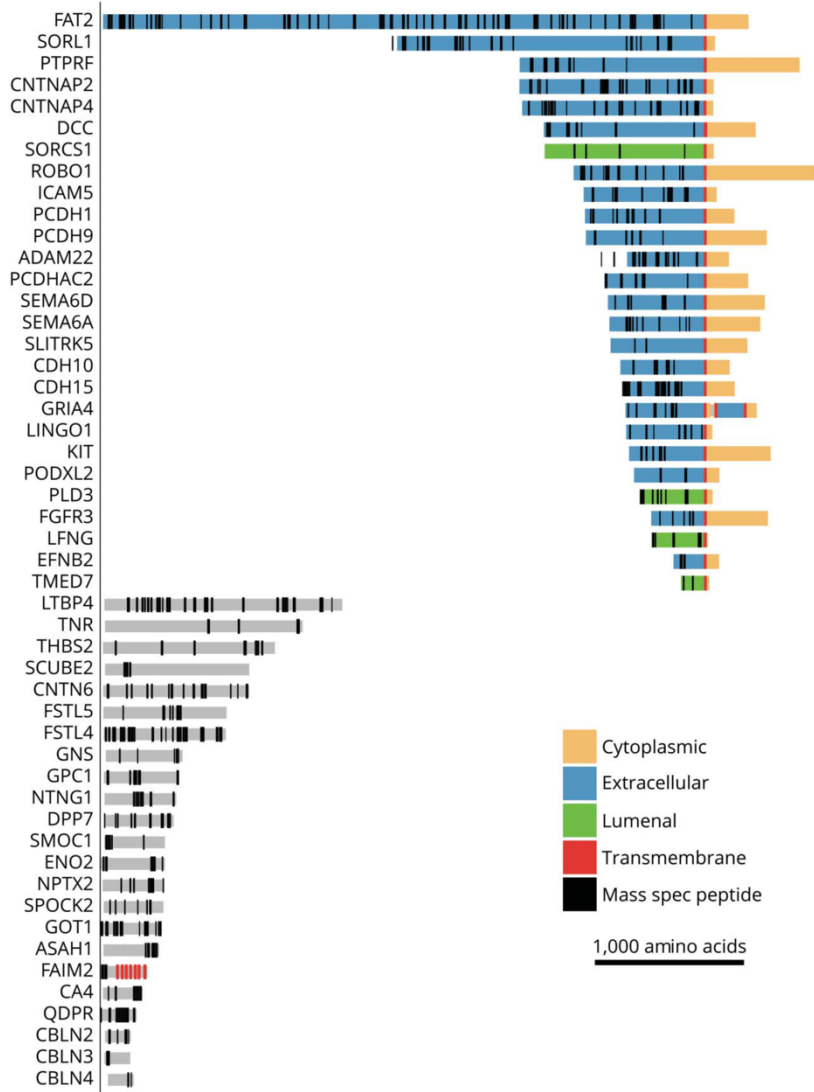


Figure 4.4 Enrichment of peptides in extracellular domains in proteins downregulated in PACNS. A subset of proteins that are strongly downregulated in primary angiitis of the CNS (PACNS) compared to noninflammatory control (NIC) are shown with UniProt annotations for transmembrane and topologic domains. Locations of peptides from mass spectrometry of NIC samples are plotted. Note the absence of peptides from cytoplasmic regions of proteins with transmembrane domains

Figure 4.5). In addition, we noted several significant findings that do not correspond to a specific pathway but pique clinical interest. These findings include elevated levels of immunoglobulins (IgM and IgA) and apolipoproteins (including Apolipoprotein B100 (ApoB100) and lipoprotein(a) (LPA)) in PACNS.

Orthogonal Confirmation

To validate our technical approach, we reproduced a subset of findings by western blot and commercial ELISA. Due to the potential clinical and therapeutic implications of identifying a role for the alternative complement cascade in PACNS, specifically complement C5, we validated elevated C5 levels in PACNS CSF through western blotting with commercial antibody and commercial ELISA (Supplementary Figure e-1, Supplementary Figure e-2, data available from Dryad). Notably, the relative C5 levels by mass spectrometry analysis and by ELISA are correlated, suggesting that the variation in these data across patients is reproducible across technical approaches. The substantial changes in IgM and IgA levels identified by mass spectrometry were also reproduced orthogonally through western blotting.

Discussion

In this exploratory study, a mass spectrometry-based approach was used to characterize the CSF proteome associated with ongoing PACNS pathology relative to non-inflammatory disease, with the intention to discover new diagnostic and/or therapeutic candidates. Currently, the diagnosis of PACNS remains challenging (Byram, Hajj-Ali, and Calabrese 2018). The diagnostics criteria for PACNS, including CSF cytology, imaging

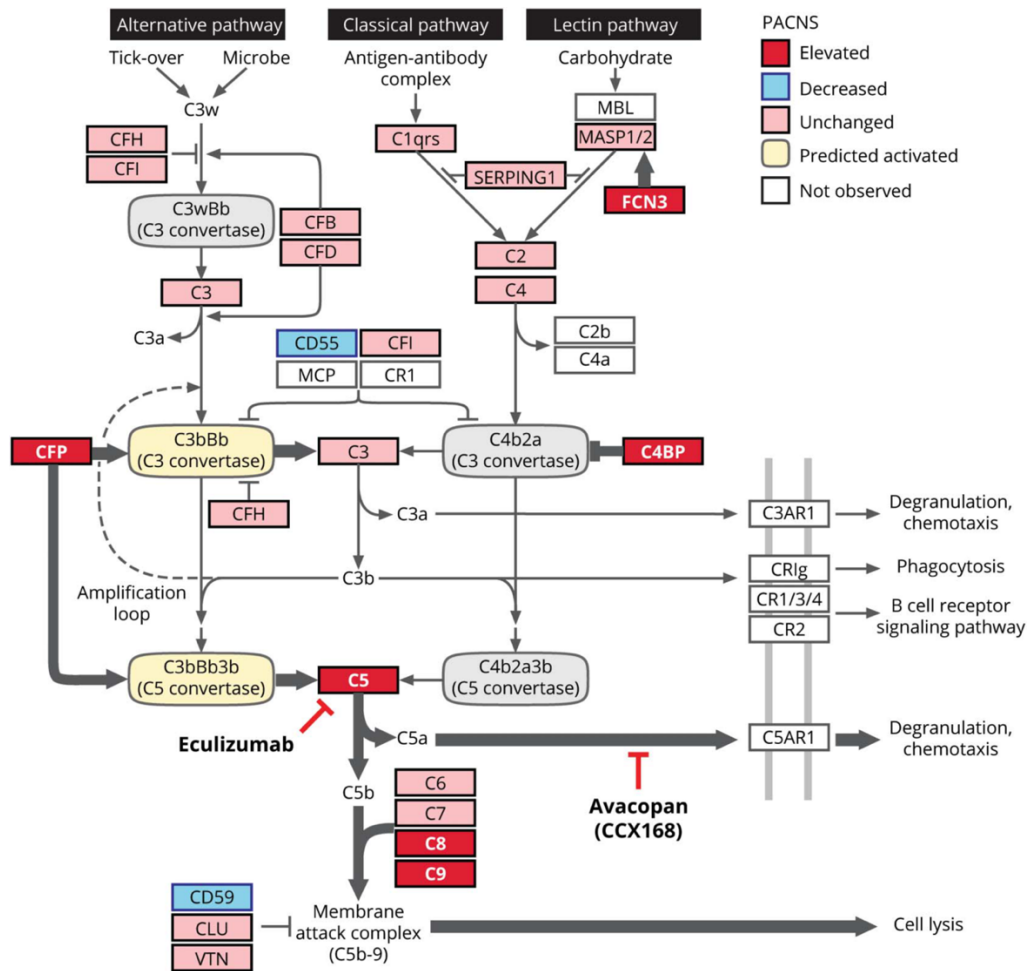


Figure 4.5 Molecular phenotype in primary angiitis of the CNS (PACNS) CSF is informed by proteomic comparison of CSF between PACNS and noninflammatory control (NIC) cohorts.

Molecular data are overlaid onto the complement cascade pathway (adapted from KEGG, hsa04610). Fold change in protein abundance between PACNS and NIC was evaluated for all proteins annotated in the pathway. Proteins are reported as elevated (red, >1.5 fold significant increase in PACNS), unchanged (pink, not significantly changed in PACNS), decreased (blue, >1.5 fold significant reduction in PACNS), or not observed (white, no abundance in NIC or PACNS). The prediction for C3 and C5 convertase activity (activated, yellow; inhibited, gray) is informed specifically by molecular changes observed in the complement regulatory proteins CFP, complement C4 binding protein (C4BP), CD59, and CD55. The proposed model predicts a shift toward activation of the alternative pathway (CFP, CD55), active inhibition of classical pathway (C4BPA and C4BPB), and elevated signaling from the terminal cascade (CD59).

abnormalities, and neurological manifestations, are largely non-specific and long-term outcomes are variable (Hajj-Ali and Calabrese 2013),(Salvarani et al. 2007). Several of these features, including elevated protein and imaging abnormalities, are also observed in individuals diagnosed with an early mimic, RCVS, as detailed in the clinical summary of our patient cohorts (Table 1) and individualized clinical vignettes (Methods) for PACNS and RCVS patients (Ducros et al. 2007)(Rocha et al. 2019). We restricted our analyses to patients with biopsy-proven PACNS to enhance the rigor of this exploratory study and to ensure that our molecular data was most closely correlated to the key features that associate with PACNS pathology (D. V Miller et al. 2009).

Our findings highlight the complement cascade as a significant feature of PACNS CSF. We find up to 12 significantly, dysregulated proteins in PACNS CSF that function within the complement cascade pathway and these changes are highly reproducible across the PACNS patient cohort. In addition, the proteomic changes within the complement pathway are specific, affecting the alternative and terminal cascade only, and include changes in transient, fluid phase regulators, whose presence/alteration are suggestive of an ongoing pathological process (Thurman and Holers 2006). Lastly, the robust changes in complement are not observed in patients with RCVS that, early in the disease course, can mimic PACNS clinically and radiologically but does not manifest with chronic inflammation (Ducros et al. 2007), (de Boysson et al. 2018). Thus, changes in the complement pathway in PACNS are unlikely due to secondary processes from acute vascular injury alone. Taken together, these data suggest that differential expression of the

complement pathway is a potential correlate of the ongoing inflammatory processes in PACNS.

The specific proteomic findings within the complement cascade have predictable consequences on complement pathway function in PACNS. The complement cascade is one of the primary effector systems of an immunologically induced inflammatory reaction, that can be triggered via two separate proteolytic pathways, the classical pathway (homologous to the lectin pathway) and the alternative pathway (Gigli 1976) (Thurman and Holers 2006). Signaling events from the classical pathway and alternative pathway converge onto a common effector pathway, known as the terminal cascade, which enables lysis and phagocytosis of foreign/inflammatory material. The alternative pathway and classical pathways are molecularly unique at the level of the Complement C3 (C3) and C5 convertases, that drive cleavage of C3 and C5, respectively. In PACNS CSF, elevated levels of the two major regulators of the C3 and C5 complement cascade convertases— CFP and C4BPA/C4BPB – were observed. CFP is the only known positive regulator of the alternative pathway, stabilizing the alternative pathway convertases, while C4BPA and C4BPB are well-studied inactivators of the C3 convertases specific to the classical pathway (Blatt, Pathan, and Ferreira 2016)(Gigli, I, Fujita, T, AND Nussenzweig 1979). These two features are consistent with an alternative pathway activation state. Downstream, among the terminal complement components, an increase in C5 and members of the membrane attack complex (MAC) (C8A, C8B, C9), and a reduction in CD59, an inhibitor of the MAC were also observed (Thurman and Holers 2006). Taken

together, we speculate that abnormal signaling of the terminal cascade is occurring in PACNS CSF, and this is due to sustained activation of the alternative pathway.

While the role of a dysregulated alternative complement cascade with respect to the pathogenesis of PACNS remains unclear, these findings implicate a critical pathway that has been exploited for therapeutic intervention in similar diseases¹⁵(Xiao et al. 2007)(Jayne et al. 2017). Specifically, activation of the alternative complement pathway has been implicated in ANCA-associated vasculitis. Blockade of the alternative pathway reduced disease activity in an ANCA-mediated mouse model of peripheral vasculitis and clinical trials have demonstrated the effectiveness of anti-C5a receptor therapies in ANCA-associated vasculitis in humans (Bekker et al. 2016)(Xiao et al. 2007)(Jayne et al. 2017). Our results reveal a previously unappreciated overlap in molecular targets between PACNS and ANCA-associated vasculitis, suggesting that similar therapeutic interventions should be considered for future PACNS trials, assuming these findings extend to larger cohorts of PACNS patients and controls.

Additionally, several observations in this cohort, beyond complement pathway components, warrant further investigation, either as diagnostic biomarkers, or as therapeutic candidates themselves. These include elevated IgM, IgA, and ApoB100. While an elevated CSF IgG index is commonly found in PACNS, our finding of elevated IgA and IgM in CSF from PACNS patients has not been previously described (Byram, Hajj-Ali, and Calabrese 2018)(D. V Miller et al. 2009)(Salvarani et al. 2007). While the functional significance of elevated IgA and IgM in the context of PACNS is unclear, elevated IgA and IgM levels may be explored as a separate diagnostic differentiator.

In contrast, there are several anecdotal pieces of evidence that implicate ApoB100 with a subset of PACNS pathological features. For one, an immune response to ApoB100, mediated by T-cells and IgM antibodies, is commonly observed in the progressive development of atherosclerotic lesions (Jan Nilsson, Björkbacka, and Fredrikson 2012) (J. Nilsson and Hansson 2008). Furthermore, elevated LDL and antibodies to apoB100 have been identified in alternative non-PACNS vasculitides, including ANCA-associated vasculitis. In the case of both MPO-ANCA and PR3-ANCA, elevated levels of anti-apoB100 antibodies are thought to be an indirect result of the chronic inflammatory disease (Slot et al. 2007). Given the robust elevation of ApoB100 in this PACNS cohort, further investigation into the direct or indirect role of ApoB100 in PACNS pathology is warranted.

Finally, an unexpected finding of this study was the loss of neural cell adhesion molecules in PACNS CSF. For non-inflammatory CSF, we observe peptides almost exclusively from the extracellular domains (ectodomains) of transmembrane proteins. Changes in these proteins in PACNS may be the result of transcriptional, translational, or post-translational regulatory differences (Tsumagari et al. 2017) (Shirakabe et al. 2017) (Lichtenthaler, Lemberg, and Fluhrer 2018). The latter may include abnormal regulation of the normal process of ectodomain shedding (Tsumagari et al. 2017) (Waldera-Lupa et al. 2017) (M. A. Miller, Sullivan, and Lauffenburger 2017). The clear absence of ectodomain peptides in PACNS CSF suggests a loss of these proteins or a loss of proteolytic homeostasis associated with shedding. While the roles of ectodomain

shedding are diverse, the mechanism and impact of dysregulated shedding in CSF is unknown.

Overall, these exploratory findings suggest potential new biomarkers of PACNS, subject to validation in larger cohorts. These results also underline the importance of future mechanistic studies around the role of complement pathways in PACNS disease pathobiology, with the ultimate goal of creating targeted therapeutic interventions for this devastating and poorly understood disease.

References for Chapter 4

- A. Alba, M., Espigol-Frigole, G., Prieto-Gonzalez, S., Tavera-Bahillo, I., Garcia-Martinez, A., Butjosa, M., Hernandez-Rodriguez, J., and C. Cid, M. (2011). Central Nervous System Vasculitis: Still More Questions than Answers. *CN* 9, 437–448.
- Bastos, P., Ferreira, R., Manadas, B., Moreira, P.I., and Vitorino, R. (2017). Insights into the human brain proteome: Disclosing the biological meaning of protein networks in cerebrospinal fluid. *Critical Reviews in Clinical Laboratory Sciences* 54, 185–204.
- Bekker, P., Dairaghi, D., Seitz, L., Leleti, M., Wang, Y., Ertl, L., Baumgart, T., Shugarts, S., Lohr, L., Dang, T., et al. (2016). Characterization of Pharmacologic and Pharmacokinetic Properties of CCX168, a Potent and Selective Orally Administered Complement 5a Receptor Inhibitor, Based on Preclinical Evaluation and Randomized Phase 1 Clinical Study. *PLoS ONE* 11, e0164646.
- Blatt, A.Z., Pathan, S., and Ferreira, V.P. (2016). Properdin: a tightly regulated critical inflammatory modulator. *Immunol Rev* 274, 172–190.
- de Boysson, H., Parienti, J.-J., Mawet, J., Arquizan, C., Boulouis, G., Burcin, C., Naggara, O., Zuber, M., Touzé, E., Aouba, A., et al. (2018). Primary angiitis of the CNS and reversible cerebral vasoconstriction syndrome: A comparative study. *Neurology* 91, e1468–e1478.
- Byram, K., Hajj-Ali, R.A., and Calabrese, L. (2018). CNS Vasculitis: an Approach to Differential Diagnosis and Management. *Curr Rheumatol Rep* 20, 37.

- Ducros, A., Boukobza, M., Porcher, R., Sarov, M., Valade, D., and Bousser, M.-G. (2007). The clinical and radiological spectrum of reversible cerebral vasoconstriction syndrome. A prospective series of 67 patients. *Brain* 130, 3091–3101.
- Gigli, I. (1976). Immunochemistry And Immunobiology Of The Complement System. *Journal of Investigative Dermatology* 67, 346–353.
- Gigli, I., Fujita, T., and Nussenzweig, V. (1979). Modulation of the classical pathway C3 convertase by plasma proteins C4 binding protein and C3b inactivator. *Proceedings of the National Academy of Sciences* 76, 6596–6600.
- Hajj-Ali, R.A., and Calabrese, L.H. (2013). Primary angiitis of the central nervous system. *Autoimmunity Reviews* 12, 463–466.
- Hutchinson, C., Elbers, J., Halliday, W., Branson, H., Laughlin, S., Armstrong, D., Hawkins, C., Westmacott, R., and Benseler, S.M. (2010). Treatment of small vessel primary CNS vasculitis in children: an open-label cohort study. *The Lancet Neurology* 9, 1078–1084.
- Jayne, D.R.W., Bruchfeld, A.N., Harper, L., Schaier, M., Venning, M.C., Hamilton, P., Burst, V., Grundmann, F., Jadoul, M., Szombati, I., et al. (2017). Randomized Trial of C5a Receptor Inhibitor Avacopan in ANCA-Associated Vasculitis. *JASN* 28, 2756–2767.
- Langley, S.R., and Mayr, M. (2015). Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics. *Journal of Proteomics* 129, 83–92.

- Lichtenthaler, S.F., Lemberg, M.K., and Fluhrer, R. (2018). Proteolytic ectodomain shedding of membrane proteins in mammals—hardware, concepts, and recent developments. *EMBO J* 37.
- Miller, D.V., Salvarani, C., Hunder, G.G., Brown, R.D., Parisi, J.E., Christianson, T.J., and Giannini, C. (2009). Biopsy Findings in Primary Angiitis of the Central Nervous System: The American Journal of Surgical Pathology 33, 35–43.
- Miller, M.A., Sullivan, R.J., and Lauffenburger, D.A. (2017). Molecular Pathways: Receptor Ectodomain Shedding in Treatment, Resistance, and Monitoring of Cancer. *Clin Cancer Res* 23, 623–629.
- Nilsson, J., and Hansson, G.K. (2008). Autoimmunity in atherosclerosis: a protective response losing control? *J Intern Med* 263, 464–478.
- Nilsson, J., Björkbacka, H., and Fredrikson, G.N. (2012). Apolipoprotein B100 autoimmunity and atherosclerosis – disease mechanisms and therapeutic potential: *Current Opinion in Lipidology* 23, 422–428.
- Oliveira, V., Póvoa, P., Costa, A., and Ducla-Soares, J. (1994). Cerebrospinal fluid and therapy of isolated angiitis of the central nervous system. *Stroke* 25, 1693–1695.
- Rocha, E.A., Topcuoglu, M.A., Silva, G.S., and Singhal, A.B. (2019). RCVS 2 score and diagnostic approach for reversible cerebral vasoconstriction syndrome. *Neurology* 92, e639–e647.
- Ruland, T., Wolbert, J., Gottschalk, M.G., König, S., Schulte-Mecklenbeck, A., Minnerup, J., Meuth, S.G., Groß, C.C., Wiendl, H., and Meyer zu Hörste, G. (2018).

Cerebrospinal Fluid Concentrations of Neuronal Proteins Are Reduced in Primary Angiitis of the Central Nervous System. *Front. Neurol.* 9, 407.

Salvarani, C., Brown, R.D., Calamia, K.T., Christianson, T.J.H., Weigand, S.D., Miller, D.V., Giannini, C., Meschia, J.F., Huston, J., and Hunder, G.G. (2007). Primary central nervous system vasculitis: analysis of 101 patients. *Ann Neurol.* 62, 442–451.

Salvarani, C., Brown, R.D., Christianson, T.J.H., Huston, J., Giannini, C., Miller, D.V., and Hunder, G.G. (2015). Adult Primary Central Nervous System Vasculitis Treatment and Course: Analysis of One Hundred Sixty-Three Patients: THERAPY AND OUTCOMES IN PRIMARY CNS VASCULITIS. *Arthritis & Rheumatology* 67, 1637–1645.

Shirakabe, K., Omura, T., Shibagaki, Y., Mihara, E., Homma, K., Kato, Y., Yoshimura, A., Murakami, Y., Takagi, J., Hattori, S., et al. (2017). Mechanistic insights into ectodomain shedding: susceptibility of CADM1 adhesion molecule is determined by alternative splicing and O-glycosylation. *Sci Rep* 7, 46174.

Slot, M.C., Theunissen, R., van Paassen, P., Damoiseaux, J.G.M.C., Cohen Tervaert, J.W., and the Limburg Nephrology Working Group (2007). Anti-oxidized low-density lipoprotein antibodies in myeloperoxidase-positive vasculitis patients preferentially recognize hypochlorite-modified low density lipoproteins. *Clinical & Experimental Immunology* 149, 257–264.

Thurman, J.M., and Holers, V.M. (2006). The Central Role of the Alternative Complement Pathway in Human Disease. *J Immunol* 176, 1305–1310.

- Tsumagari, K., Shirakabe, K., Ogura, M., Sato, F., Ishihama, Y., and Sehara-Fujisawa, A. (2017). Secretome analysis to elucidate metalloprotease-dependent ectodomain shedding of glycoproteins during neuronal differentiation. *Genes Cells* 22, 237–244.
- Waldera-Lupa, D.M., Etemad-Parishanzadeh, O., Brocksieper, M., Kirchgaessler, N., Seidel, S., Kowalski, T., Montesinos-Rongen, M., Deckert, M., Schlegel, U., and Stühler, K. (2017). Proteomic changes in cerebrospinal fluid from primary central nervous system lymphoma patients are associated with protein ectodomain shedding. *Oncotarget* 8, 110118–110132.
- Xiao, H., Schreiber, A., Heeringa, P., Falk, R.J., and Jennette, J.C. (2007). Alternative Complement Pathway in the Pathogenesis of Disease Mediated by Anti-Neutrophil Cytoplasmic Autoantibodies. *The American Journal of Pathology* 170, 52–64.
- Zybailov, B., Mosley, A.L., Sardi, M.E., Coleman, M.K., Florens, L., and Washburn, M.P. (2006). Statistical Analysis of Membrane Proteome Expression Changes in *Saccharomyces cerevisiae*. *J. Proteome Res.* 5, 2339–2347.

Chapter 5 Metagenomic Next-Generation Sequencing Reveals

Miamiensis Avidus (Ciliophora: Scuticociliatida) in the 2017

Epizootic of Leopard Sharks (*Triakis Semifasciata*) In San Francisco Bay, California, USA

Authors:

Hanna Retallack,¹ Mark S. Okihiro,^{2,6} Elliot Britton,³ Sean Van Sommeran,⁴ Joseph L. DeRisi^{1,5}

Affiliations:

¹ Department of Biochemistry and Biophysics, University of California San Francisco, 1700 4th St., San Francisco, California 94158, USA

² Fisheries Branch, Wildlife and Fisheries Division, California Department of Fish and Wildlife, 1880 Timber Trail, Vista, California 92081, USA

³ San Francisco University High School, 3065 Jackson St., San Francisco California 94115, USA

⁴ Pelagic Shark Research Foundation, 750 Bay Ave. #2108, Capitola California 95010, USA

⁵ Chan-Zuckerberg Biohub, 499 Illinois St., San Francisco, California 94158, USA

⁶ Corresponding author: (email: Mark.Okihiro@wildlife.ca.gov)

Includes material previously published in:

Retallack H, Okihiro MS, Britton E, Sommeran SV, DeRisi JL. Metagenomic next-generation sequencing reveals *miamiensis avidus* (ciliophora: scuticociliatida) in the 2017 epizootic of leopard sharks (*Triakis Semifasciata*) in San Francisco bay, California, USA. J Wildl Dis. 2019;55(2):375-386. doi:10.7589/2018-04-097

Introduction

The investigation of mass mortality events among wildlife populations can provide insight into ecosystem health and human impact. However, identifying an etiology is often challenging. Metagenomic next-generation sequencing (mNGS) provides an unbiased approach, which has been used successfully in human and animal infections (Wilson et al. 2014; Zylberberg et al. 2016; Dervas et al. 2017). Through the analysis of all nucleic acids in a sample, mNGS can simultaneously test for all known organisms, and can also identify novel pathogens including distantly related species. Furthermore, the cost of NGS technologies continues to decrease, making these methods an increasingly viable option for routine wildlife surveillance and disease investigations.

In the past 50 years, several mass mortality events of unknown etiology have affected leopard sharks (*Triakis semifasciata*) in San Francisco (SF) Bay, California. In 1967, over 1,000 dead sharks, mainly leopard sharks, were collected in 1 mo in Alameda (Russo and HERALD 1968; Russo 2015). More recently, unusual shark deaths were noted in the spring of 2006, and mass mortality again afflicted SF Bay leopard sharks in the spring and early summer of 2011 involving likely hundreds of leopard sharks though the event was not systematically documented. Moribund sharks were often described as confused and disoriented, with erratic behaviors and swimming patterns.

Scuticociliates are free-living marine protozoa that belong to the subclass Scuticociliatida of the phylum Ciliophora (Gao et al. 2016). As opportunistic pathogens, several species of scuticociliates have been reported to cause disease in diverse marine

teleost fish species (Munday et al. 1997; Ramos et al. 2007; Garza et al. 2017), and recently, in the subclass of cartilaginous fish known as elasmobranchs (Stidworthy et al. 2014; W. Li et al. 2017). Scuticociliatosis is an economically important problem for commercial marine fish culture (R. Iglesias et al. 2001) but has not been observed in wild fish populations.

We sought to identify a cause for mass mortality of leopard sharks in SF Bay in the spring of 2017. Using mNGS and confirmatory molecular and histologic assays, we identified the scuticociliate, *Miamiensis avidus*, in the central nervous system of stranded sharks, suggesting that this pathogen could contribute to significant disease in wild elasmobranchs.

Materials and Methods

Shark stranding surveillance

The majority of shark and ray strandings were reported to the California Department of Fish and Wildlife (CDFW) by members of the public, often via The Marine Mammal Center (Sausalito, California) and the Pelagic Shark Research Foundation (Santa Cruz, California). Additional stranding data were provided by East Bay Regional Park District rangers (Oakland, California), the National Parks Service, and CDFW wardens working in and around San Francisco Bay. Stranding data were also acquired during three brief foot surveys of the Foster City shoreline conducted by CDFW in April, June, and August of 2017. Stranding data included date, location, species, approximate size, condition (live, dead, autolyzed), and presence of abnormal behavior (e.g., swimming in

circles). Photos were often submitted, with occasional videos. Stranding data were recorded and sorted on the basis of species and date.

Sample collection

Stranded sharks were chosen for necropsy by a CDFW pathologist based on condition, with preference given to live moribund and fresh dead sharks (non-autolyzed with red gills). Sharks were either necropsied in the field or iced and necropsied at CDFW (Vista, California) within 72 h. Heads of some sharks were removed and frozen at –10 C until necropsy. Two captive sharks were necropsied: one Pacific angelshark (*Squatina californica*) on display at the Aquarium of the Bay (San Francisco, California), and one moribund leopard shark on display at the Marine Science Institute (Redwood City, California). As controls, grossly normal leopard sharks were collected via gill net from Newport Bay in southern California. A great white shark (*Carcharodon carcharias*) and soupfin shark (*Galeorhinus galeus*) were collected from outside SF Bay.

Necropsy

Sampled sharks were cleaned of external mud and debris via freshwater spray. Species and sex were determined via examination of fins and dentition. Sharks were weighed and total and fork length measured. The dorsum of the head was cleaned with multiple passes using disposable disinfecting wipes (Clorox, Oakland, California, USA). When possible, endolymphatic pores were identified. The endolymphatic fossa (oval concave depression in the chondrocranium) was located by digital palpation. Using sterilized instruments, a 3x5 cm incision was made centered on the endolymphatic fossa and pores. Subcutaneous tissues overlying the fossa were sampled with a sterile cotton

swab for microbiologic assessment. Subcutaneous fluid was aspirated with a sterile 1 mL pipette for cytological assessment. The skin sample containing the endolymphatic pores and ducts was fixed in 10% formalin. The calvarium, including the endolymphatic fossa, was removed with a sterile scalpel and new blade, then fixed in formalin. Removal of the calvarium exposed both inner ears and the cerebellum. Cerebrospinal fluid (CSF) overlying the cerebellum was sampled with a sterile cotton swab. Two 1 mL CSF samples were taken by sterile pipette and frozen at -10°C in cryovials. A third CSF sample was taken for cytological assessment. Perilymph from one inner ear was sampled with a sterile cotton swab. A second perilymph sample was taken for cytological assessment. The brain and olfactory lamellae were exposed via sharp dissection. The meninges, CSF, brain, inner ears, and olfactory lamellae were examined for evidence of inflammation and hemorrhage. Brains were separated from the chondrocranium via inversion of the skull and severing cranial nerves. Olfactory lamellae and associated olfactory bulbs were removed via sharp dissection. The brain and olfactory lamellae were fixed in 10% formalin. In some sharks, one otic capsule was also taken and fixed in formalin. Gills, heart, kidneys, and abdominal organs were also examined at necropsy. Selected organs were sampled and fixed in formalin from some sharks.

Cytology and histology

Samples of subcutaneous fluid surrounding the endolymphatic ducts, inner ear perilymph, and CSF were examined on glass slides under darkfield light microscopy at 200 and 400X using a binocular microscope. Red blood cells, inflammatory cells, and microbial pathogens were identified. Histology samples (primarily brain and nasal

olfactory lamellae) were immersion fixed in 10% formalin for 2 wk to 3 mo and then routinely paraffin processed. Paraffin blocks were sectioned at 5-7 μm and sections stained with hematoxylin and eosin, then examined with light microscopy. Degree of inflammation and necrosis, as well as numbers of protozoa in tissue sections, were semi-quantitatively scored as: not present (0), mild (1+), moderate (2+), or severe (3+).

Microbiology

Samples of subcutaneous fluid surrounding the endolymphatic ducts, inner ear perilymph, and CSF were plated onto blood agar and Sabouraud-Dextrose agar. Cultures were incubated aerobically at room temperature (15-20 C) for 4 wk and checked daily for growth. Selected isolates were sent to the University of Florida (Gainesville, Florida) for biochemical and PCR identification.

Nucleic acid extraction and sequencing

For RNA, 250 μL of CSF was placed in TRI-Reagent (Zymo Research, Irvine, California, USA) and homogenized with 2.8 mm ceramic beads (Omni, Kennesaw, Georgia, USA) on a TissueLyser II (Qiagen, Germantown, Maryland, USA) at 15 Hz for two 30 sec pulses, separated by 1 min on ice. Total RNA was then extracted using the Direct-Zol RNA MicroPrep Kit with DNase treatment (Zymo Research), eluted in 12 μL and stored at -80 C until use. For DNA, 250 μL of CSF was placed in 750 μL Lysis Solution of the Fungal/Bacterial DNA kit (Zymo Research) and homogenized as above with a single 2 min homogenization pulse. Total DNA was then extracted using the Fungal/Bacterial DNA kit (Zymo Research), eluted in 25 μL , and stored at -80 C until use. RNA samples were processed using 5 μL total RNA as input into the NEBNext Ultra II

RNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, Massachusetts, USA). Samples were sequenced on an Illumina MiSeq instrument using 150 nucleotide (nt) paired-end sequencing. A no-template control (nucleic acid-free water) was included in each batch of nucleic acid extractions and library preparation. Raw sequencing reads were deposited at the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under BioProject PRJNA438541, SRA accession SRP136047.

Primers used to amplify ciliate and shark genomic sequences are listed in Table S1. Purified PCR products were sequenced via the Sanger method by QuintaraBio (Albany, California, USA). See Supplemental Methods for details.

Bioinformatics

Next-generation sequencing data were analyzed using a computational pipeline originally developed to identify potential pathogens in human samples (Wilson et al. 2014). Briefly, host sequences were identified using publicly available shark genomes and transcriptomes, and the remaining non-host sequences were compared to the NCBI nucleotide and protein databases. Potential pathogens were identified based on a minimum read abundance, likelihood of pathogenicity, and absence in negative control samples. For species determination, reads mapping to the ciliate 18S small subunit (SSU) and 28S large subunit (LSU) of the nuclear ribosomal RNA locus (rDNA) were assembled and compared to the NCBI database using BLASTn (S. F. Altschul et al. 1990). See Supplemental Methods for details.

New sequences in this study include partial sequences of the mitochondrial (mt) cytochrome c oxidase I (*cox1*) gene, and SSU and LSU rDNA of the ciliate identified in the

shark samples, deposited in GenBank (Accession numbers MH078243-MH078249, MH062876, MH064355). See Supplemental Methods for details.

Results

Beginning in March 2017, members of the public reported sharks swimming with unusual behaviors and stranding on beaches along the SF Bay shoreline, with the majority of strandings occurring in the Foster City area (Figure 5.1). Leopard sharks were observed swimming unusually close to shore, appearing uncoordinated and disoriented, suggestive of an inner ear or central nervous system issue. At the height of the epizootic in April and May, 20-30 dead leopard sharks were being found daily along the shoreline in Foster City. We estimated that over 1,000 leopard sharks died between March and August 2017 in SF Bay. Necropsies were performed on 11 fresh dead or live moribund leopard sharks, and on the heads of five frozen sharks. Gross and cytological lesions were consistent with meningoencephalitis, and characterized by hemorrhage, cloudy CSF, and thickened meninges (Figure 5.2). Lesions were especially prominent in the olfactory bulbs and lobes. Olfactory lamellae, adjacent to olfactory bulbs, were often markedly hemorrhagic and inflamed. There was no gross evidence of inflammation in the subcutaneous tissues surrounding the endolymphatic ducts or inner ears, which are target organs for a common bacterial pathogen (*Carnobacterium maltaromaticum*) of sharks (Schaffer et al. 2013). (No lesions were observed in gills, heart, or abdominal organs. Cytologic exam of CSF revealed dense mixed inflammation (mononuclear inflammatory cells and polymorphonuclear cells). No pathogens were observed. Conventional microbiology was

uninformative: blood and Sabouraud-Dextrose agar cultures of CSF, inner ear perilymph, and subcutaneous tissues and surrounding endolymphatic ducts yielded no growth or fungal and bacterial contaminants associated with field sampling or postmortem colonization of tissues.

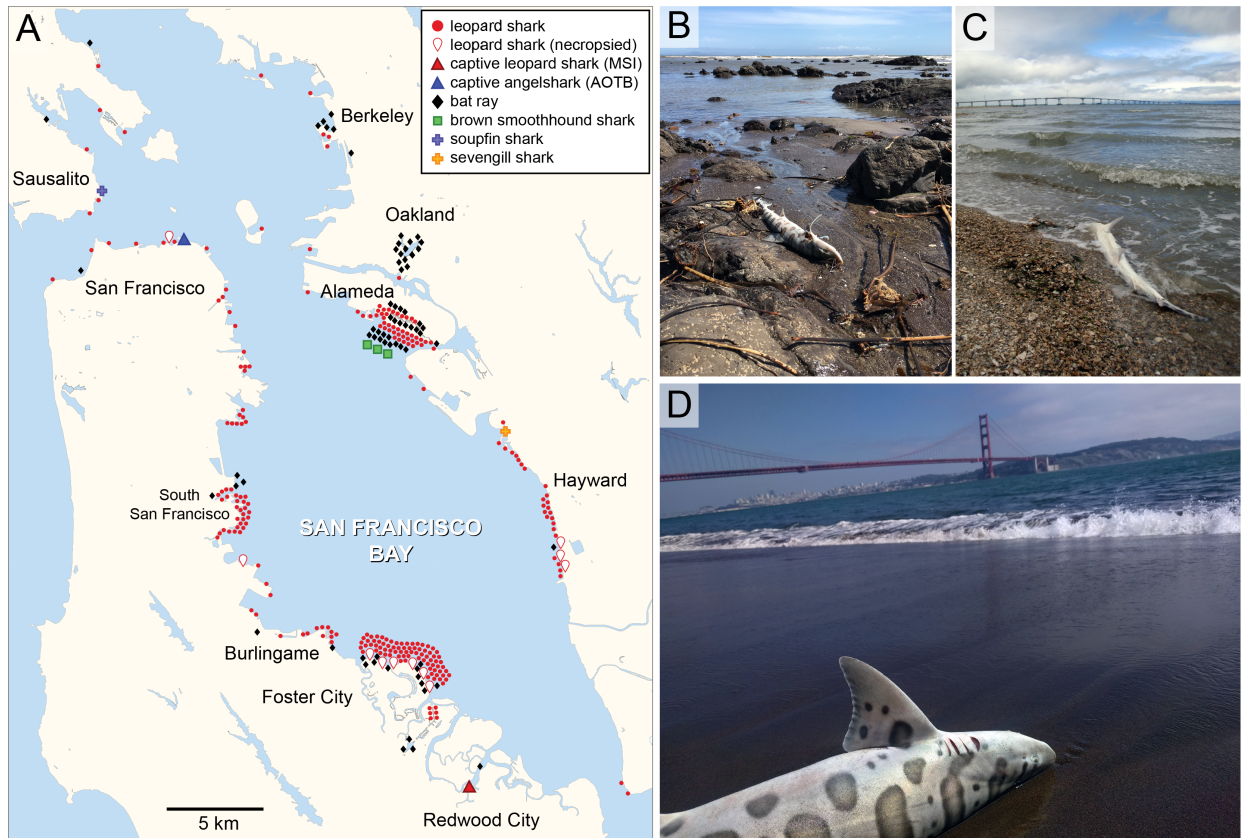


Figure 5.1 Map and photographs of shark strandings occurring in San Francisco Bay (SF Bay) in spring of 2017.

A) Map of SF Bay showing locations of stranded sharks and bat rays, including leopard sharks (*Triakis semifasciata*), a Pacific angelshark (*Squatina californica*), bat rays (*Myliobatis californica*), brown smoothhound sharks (*Mustelus henlei*), a soupfin shark (*Galeorhinus galeus*), and a sevengill shark (*Notorynchus cepedianus*). (MSI) Marine Science Institute. (AOTB) Aquarium of the Bay. B-D) Representative photographs of stranded leopard sharks around SF Bay taken between March and August 2017.

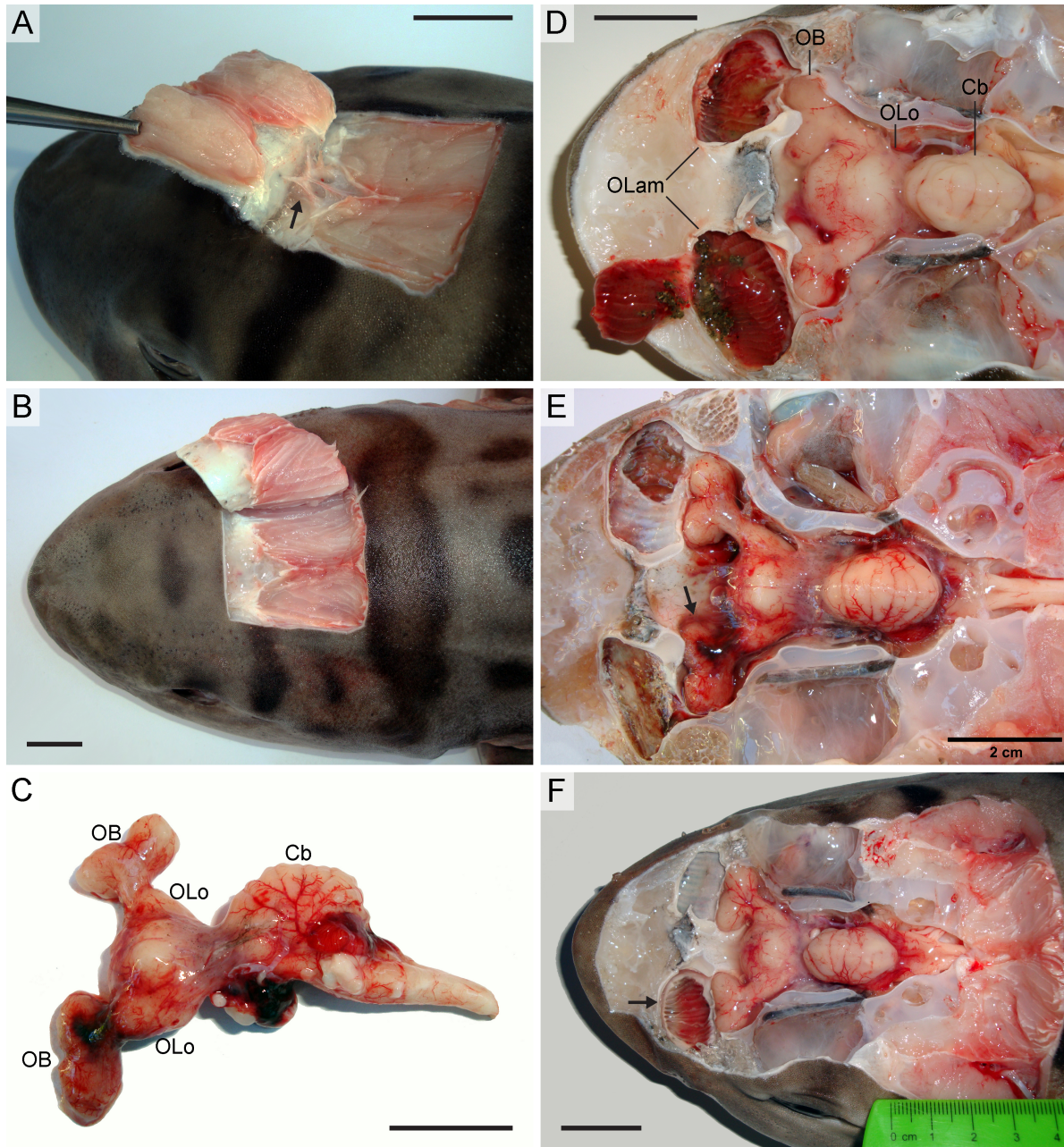


Figure 5.2 Gross observations of lesions in brains of stranded leopard sharks (*Triakis semifasciata*) in San Francisco Bay in spring of 2017 in which the scuticociliate *Miamiensis avidus* was involved.

A-B) Dissection from the superior aspect of the head exposing the endolymphatic ducts (arrow). C) Brain with hemorrhagic lesions removed from cranial cavity of (B) and depicted in situ in (E). D-F) Dissection of cranial vault revealing superior surface of brain, hemorrhagic lesions and congested olfactory lamellae (arrows). (OB) olfactory bulb; (OLO) olfactory lobe; (Cb) cerebellum; (OLam) olfactory lamellae. Scale bars, 2cm.

Guided by the signs of meningoencephalitis, samples of CSF were taken for molecular analysis from 15 stranded or ill-appearing sharks from SF Bay including 11 leopard sharks (*Triakis semifasciata*), one sevengill shark (*Notorynchus cepedianus*), and one soupfin shark (*Galeorhinus galeus*), one captive leopard shark from the Marine Science Institute (Redwood City), and one captive Pacific angelshark (*Squatina californica*) from the Aquarium of the Bay (San Francisco; Table 5.1, Table S2). Control CSF samples were collected from four grossly normal leopard sharks captured by gill net from Newport Bay in southern California, and from two sharks with meningoencephalitis that had died in southern or central California: one great white shark (*Carcharodon carcharias*) and one soupfin shark.

To identify potential pathogens associated with leopard shark mortality, we performed mNGS on CSF samples from five sharks exposed to SF Bay water and two sharks from elsewhere on the California coast. Reads aligning to species in the Ciliophora phylum (taxonomy ID 5878) were identified in all five SF Bay sharks but were absent from the no-template control and non-SF Bay sharks (Table S3). No other credible pathogens were identified. Sixty-four percent (64%) of identified Ciliophora reads aligned to the ciliate rDNA locus. High-confidence contigs were assembled for the partial SSU (943nt with a gap of 310nt) and LSU (1916nt with gaps of 39nt and 56nt) rDNA, with coverage between four and 2545 unique reads, and nucleotide identity >99%. The SSU contig aligned with greater confidence to *M. avidus* than other scuticociliate species by BLASTn.

Table 5.1 Histopathologic lesions in stranded and captive sharks from SF Bay.

Sharks were sampled as part of an investigation of a large-scale mortality event of leopard sharks (*Triakis semifasciata*) in March–August 2017. Species other than leopard sharks include Pacific angelshark (*Squatina californica*), and sevengill shark (*Notorynchus cepedianus*). Tissues were examined for inflammation (I), necrosis (N), and abundance of protozoa (P), and scored on a four-point scale as not present (0), mild (1+), moderate (2+), or severe (3+).

Fish ID	Shark Species	Collection		Histopathology		mNGS Ciliate ^b	PCR <i>M. avidus</i> ^c
		Date	Location	Meningo-encephalitis ^a	Protozoa		
LS01	<i>T. semifasciata</i>	09 April 2017	SF	Severe	Brain	n/a	+
LS02	<i>T. semifasciata</i>	15 April 2017	Foster City	Severe	Brain	n/a	+
LS03	<i>T. semifasciata</i>	25 April 2017	Foster City	Severe	-	n/a	+
LS04	<i>T. semifasciata</i>	25 April 2017	Foster City	n/a	n/a	n/a	+
LS05	<i>T. semifasciata</i>	25 April 2017	Hayward	n/a	n/a	n/a	+
LS06	<i>T. semifasciata</i>	25 April 2017	Hayward	n/a	n/a	n/a	+
LS07	<i>T. semifasciata</i>	25 April 2017	Hayward	n/a	n/a	n/a	+
LS08	<i>T. semifasciata</i>	26 April 2017	Foster City	Severe	Brain/OL	+	+
LS09	<i>T. semifasciata</i>	26 April 2017	Foster City	Severe	Brain	n/a	+
LS10	<i>T. semifasciata</i>	02 May 2017	SF	Severe	Brain/OL	+	+
LS11	<i>T. semifasciata</i>	13 May 2017	Foster City	n/a	n/a	n/a	+
LS12	<i>T. semifasciata</i> ^d	23 May 2017	MSI	Severe	Brain	+	+
S1	<i>S. californica</i> ^d	17 May 2017	AOTB	Mild	-	+	+
S2	<i>N. cepedianus</i>	17 May 2017	San Leandro	Severe	-	+	+
S3	<i>G. galeus</i>	05 July 2017	Sausalito	Moderate	-	n/a	-
LS13	<i>T. semifasciata</i>	18 July 2017	Newport Harbor	n/a	n/a	n/a	-
LS14	<i>T. semifasciata</i>	18 July 2017	Newport Harbor	n/a	n/a	n/a	-
LS15	<i>T. semifasciata</i>	18 July 2017	Newport Harbor	n/a	n/a	n/a	-
LS16	<i>T. semifasciata</i>	18 July 2017	Newport Harbor	n/a	n/a	n/a	-
S4	<i>C. carcharias</i>	08 April 2017	Santa Cruz	Severe	-	-	-
S5	<i>G. galeus</i>	24 May 2017	La Jolla	Severe	-	-	-

^aMeningoencephalitis in olfactory lamellae or brain (olfactory bulbs/lobes). ^bCiliate identified by mNGS as *Miamiensis avidus*. ^c*Miamiensis avidus* identified by PCR. ^dCaptive shark on display in aquarium. - = absent/negative; + = present/positive; n/a = assay not performed (not applicable); OL = olfactory lamellae; MSI = Marine Science Institute (Redwood City, California); AOTB = Aquarium of the Bay (San Francisco, California).

To confirm mNGS results with an orthogonal molecular approach, we amplified a variable region of the ciliate *cox1* gene from DNA extracted from CSF of these sharks. Amplification was detected in a nested PCR for *M. avidus* in five of five stranded sharks that were positive for *M. avidus* by mNGS (Figure 5.3A, Fig. S1), and no amplification was detected in two of two sharks negative by mNGS or in the no-template control (Fig. S1). The sequence of the *cox1* amplicons (via the Sanger method) was most similar to other *M. avidus* sequences (Figure 5.3B).

Given these findings, brain and nasal tissues from nine affected sharks were more closely examined for histopathologic evidence of ciliated protozoa (Table 5.2). The majority of sharks histologically examined had moderate to severe inflammation and mild to severe necrosis in the olfactory lamellae of the nose (Figure

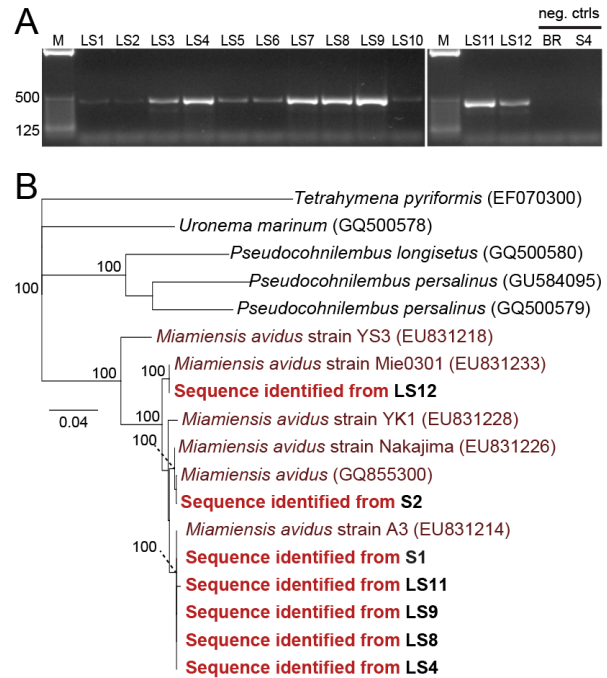


Figure 5.3 Molecular identification of the scuticociliate parasite *Miamiensis avidus* in cerebrospinal fluid from leopard sharks (*Triakis semifasciata*) in San Francisco Bay in spring of 2017.

A) DNA samples from SF Bay leopard sharks (LS1-12) and negative control animals (bat ray (BR) and S4) were tested by nested PCR using primers specific to the *cox1* gene of *M. avidus* (expected size 422bp). M: 25bp ladder. B) Neighbor-joining phylogenetic tree constructed from mt *cox1* nucleotide sequences. *Tetrahymena pyriformis* served as the outgroup. New sequences in this study are in bold, labeled according to Fish ID (see Table 1). Nodes are labeled with bootstrap values based on 1,000 resamplings (for values >80). GenBank accession numbers provided for reference sequences. Scale bar, nucleotide substitutions per site.

5.4). Inflammation was variable in composition but was largely a mixed infiltrate of mononuclear cells (macrophages and lymphocytes) and polymorphonuclear cells (primarily eosinophils and heterophils).

Table 5.2 Histopathologic lesions in stranded and captive sharks from San Francisco Bay.

Sharks were sampled as part of an investigation of a large-scale mortality event of leopard sharks (*Triakis semifasciata*) in the San Francisco Bay area in March-August 2017. Species other than leopard sharks include Pacific angelshark (*Squatina californica*), and sevengill shark (*Notorynchus cepedianus*). Tissues were examined for inflammation (I), necrosis (N), and abundance of protozoa (P), and scored on a four-point scale as not present (0), mild (1+), moderate (2+), or severe (3+).

Fish		Olfactory lamellae			Olfactory bulbs			Olfactory lobes			Optic lobes			Cerebellum		
ID	Shark species	I	N	P	I	N	P	I	N	P	I	N	P	I	N	P
LS01	<i>T. semifasciata</i>	3+	1+	0	3+	3+	1+	2+	1+	0	0	0	0	1+	0	0
LS02	<i>T. semifasciata</i>	3+	0	0	3+	3+	0	3+	1+	0	3+	1+	1+	2+	0	0
LS03	<i>T. semifasciata</i>	n/a	n/a	n/a	3+	3+	0	3+	1+	0	3+	2+	0	1+	0	0
LS08	<i>T. semifasciata</i>	3+	3+	1+	3+	3+	2+	3+	3+	3+	1+	0	0	1+	0	0
LS09	<i>T. semifasciata</i>	3+	2+	0	3+	3+	0	3+	3+	2+	1+	0	0	n/a	n/a	n/a
LS10	<i>T. semifasciata</i>	2+	0	1+	3+	3+	3+	3+	3+	3+	2+	3+	3+	n/a	n/a	n/a
LS12	<i>T. semifasciata</i>	2+	1+	0	3+	3+	3+	3+	3+	3+	n/a	n/a	n/a	n/a	n/a	n/a
S1	<i>S. californica</i>	1+	0	0	1+	1+	0	0	0	0	0	0	0	0	0	0
S2	<i>N. cepedianus</i>	3+	1+	0	3+	3+	0	3+	2+	0	3+	2+	0	3+	2+	0

n/a = not examined

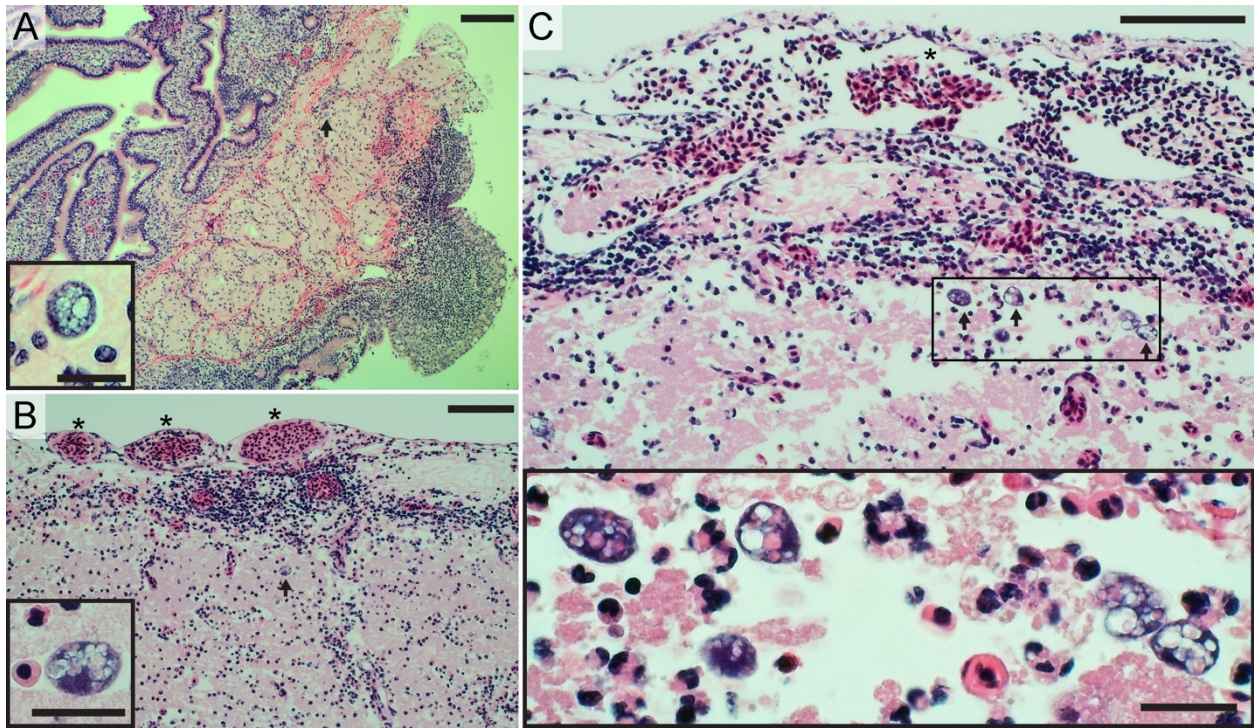


Figure 5.4 Histology showing protozoa morphologically consistent with the scuticociliate *Miamiensis avidus* in brain tissues of stranded leopard sharks in San Francisco Bay in spring of 2017.

A) Olfactory lamellae (top left) and filament (center), with submucosal inflammation (bottom right) and scattered protozoa (arrow, magnified in inset). Scale bars, 200 μm and 25 μm (inset). B-C) Olfactory bulb of the brain with congested vessels in overlying meninges (asterisks), inflammatory infiltrate, and protozoa (arrows, magnified in insets). Scale bars, 100 μm and 25 μm (insets). Representative sections stained with hematoxylin and eosin are shown.

Lamellar inflammation was present within the hyperplastic mucosal epithelium, expanded submucosal connective tissues, and branches of the olfactory nerve. Ciliated protozoan parasites, morphologically consistent with *M. avidus*, were present in the olfactory lamellae of two sharks in small numbers. Protozoa were oblong, with irregular eccentric nuclei and characteristic vacuolated basophilic cytoplasm. Cytoplasmic vacuoles were clear or filled with pale eosinophilic material. Parasites were 20-30 μm in length by 10-20 μm wide.

In the brain, inflammatory and necrotizing lesions were concentrated rostrally in the meninges and parenchyma of the olfactory bulbs and olfactory lobes. Cellular composition of inflammatory lesions was comparable to that of lesions in the olfactory lamellae. Inflammatory lesions were consistently associated with moderate to marked congestion of capillaries and veins in the meninges and parenchyma of the brain. The majority of sharks also had severe congestion of blood vessels associated with the lateral ventricles of the olfactory bulbs and lobes. Many lateral ventricles were filled with mixed inflammatory cells. Necrotizing lesions were characterized by finely granular, pale eosinophilic cellular debris mixed with degenerating and necrotic inflammatory cells, degenerating and necrotic neurons, and variable numbers of protozoa. Small to large numbers of protozoa were found in olfactory bulb and/or lobe sections of five of nine sharks. In many necrotic sections of brain, protozoa were present in large numbers but were often difficult to detect because of loss of membrane integrity, and loss of cytoplasmic and nuclear basophilia. Necrotic protozoa could, however, be usually identified because of their relative large size and because of characteristic cytoplasmic vacuoles. Intact protozoa were similar to those observed in the olfactory lamellae. Caudally, inflammatory and necrotizing lesions were less common and less severe in the optic lobes and cerebellum, corresponding to fewer protozoa. Of the nine sharks examined, the captive angelshark (S1) had the fewest and mildest histologic lesions and no protozoa.

With evidence for *M. avidus* as a candidate pathogen, we used PCR to screen nine additional leopard sharks that stranded in SF Bay in spring of 2017, compared to four

grossly normal leopard sharks that were caught in in southern California. A soupfin shark (*Galeorhinus galeus*) that stranded in SF Bay in July 2017 was also tested. The PCR was targeted to the ciliate SSU and mt *cox1* regions, initially using universal ciliate primers (Jung et al. 2005; Whang et al. 2013), followed by Sanger sequencing or species-specific nested PCR for the *cox1* gene to test for the related pathogenic ciliate species, *Uronema marinum*, *Pseudocohnilembus longisetus*, and *Pseudocohnilembus persalinus* (Whang et al. 2013). For *cox1*, amplification specific to *M. avidus* was detected in nine of nine SF Bay leopard sharks (Figure 5.3A, Fig. S2), and was not detected in the southern California leopard sharks (Fig. S3) or the SF Bay soupfin shark (Fig. S1). Amplicon sequencing revealed 99.1% pairwise identity. The *cox1* sequences clustered together with reference *M. avidus* sequences on a neighbor-joining (NJ) tree (Figure 5.3B). For the SSU, an amplicon of the expected size was detected in three of 12 SF Bay leopard sharks and in the two SF Bay non-leopard sharks that were positive by mNGS and was absent from all four of the leopard sharks and both of the non-leopard sharks from southern California (Fig. S4). The sequences of the ciliate SSU amplicon from five sharks (LS3, LS4, LS11, S1, and S2) were 100% identical, were concordant with the regions of overlap from mNGS, and clustered together with reference *M. avidus* sequences on a NJ tree (Fig. S5).

Discussion

In this study, we described an epizootic of wild leopard sharks characterized by stranding behavior and meningoencephalitis and we provided strong molecular and histological evidence that implicated the ciliated protozoan, *M. avidus*, as the candidate

pathogen associated with the 2017 SF Bay mass mortality event. We identified *M. avidus* through an unbiased, NGS-based approach, which has been used previously in investigations of a wide range of human and non-human infectious diseases.

The lack of a leopard shark genome presented a technical challenge, and thus we utilized sequences from related species. Despite being unable to identify all host sequences using our proxy-metagenome host sequences, we were still able to identify a plausible pathogen embedded in a large amount of unrelated and unidentified host sequence. Future contributions of shotgun sequencing data will improve our ability to identify sequences of unusual hosts, such as sharks, thereby improving our ability to detect novel pathogens. Among the species-specific regions flanked by conserved sequences, such as the commonly used ribosomal RNA (SSU and LSU) and *cox1* genes, we found that *cox1* was similar to SSU and better than LSU in discriminating between *M. avidus* and related pathogenic scuticociliates. This finding is consistent with reports of higher intraspecific variation of the *cox1* gene (Budiño et al. 2011; Jung et al. 2011). Nonetheless, using multiple genes for molecular phenotyping can add confidence, as there remain discrepancies in the field about highly-similar taxa such as *M. avidus* and *Philasterides dicentrarchi* (Jung et al. 2011; De Felipe et al. 2017), and there are likely sub-species divisions yet to be realized (Gao, Katz, and Song 2012).

We observed *M. avidus* only in sharks exposed to SF Bay water, including two wild-caught animals in captivity. The associated phenotype was consistent with other reports of ciliate infection of elasmobranchs, notable for necrotizing meningoencephalitis (Stidworthy et al. 2014; W. Li et al. 2017). Given the distribution of protozoa observed on

histopathology, pathogenesis in leopard sharks likely involves a nasal route as suggested for other host species (Moustafa et al. 2010), with initial protozoal invasion of olfactory lamellae, followed by extension into the olfactory bulbs and lobes of the brain. Massive inflammation and severe encephalomalacia, associated with *M. avidus* infection, could account for the disorientation and abnormal behavior of sharks prior to stranding. Further support implicating *M. avidus* in these repeated mortality events comes from the necropsy of a single leopard shark that stranded in the 2011 SF Bay epizootic, which showed extensive inflammation and an abundance of unicellular ciliated protozoa throughout the brain (Kubinski et al., n.d.). We found no evidence of other pathogens that have been reported in elasmobranchs near the Pacific coast of North America.

We cannot exclude the possibility that *M. avidus* was not the primary or sole driver of disease and mortality. Factors such as water temperature, salinity, toxins, or other pathogens could increase susceptibility to an opportunistic infection by *M. avidus* (Hopkins and Cech 2003; Carlisle and Starr 2009). In SF Bay, leopard sharks may be especially vulnerable each spring when they aggregate in large numbers in the warm shallow waters of bays and estuaries (Hight and Lowe 2007; Nosal et al. 2013), with greater exposure to runoff that may contain toxins or decreased salinity. Lower salinity has been associated with increased fish mortality in the context of scuticociliate infection (Takagishi, Yoshinaga, and Ogawa 2009), and with more rapid growth of *Philasterides dicentrarchi* (synonymous with *M. avidus*) *in vitro* (Iglesias et al. 2003). Notably, heavy seasonal rainfall with runoff into the bay preceded each of the spring epizootics in 2006,

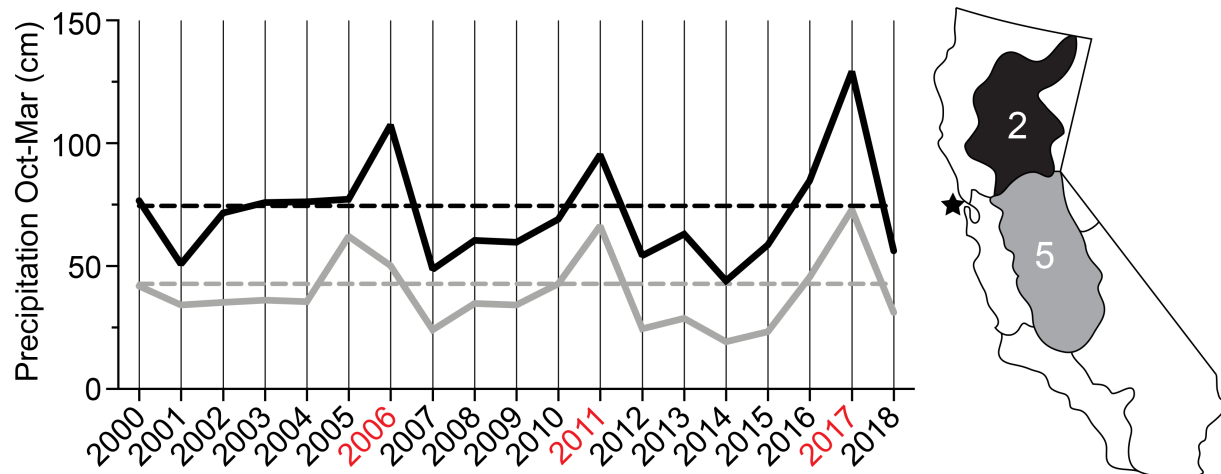


Figure 5.5 Precipitation in regions draining to San Francisco Bay (SF Bay) by year, including 2017 when strandings of leopard sharks (*Triakis semifasciata*) were associated with infection by *Miamiensis avidus*.

Six-month cumulative precipitation (solid lines) ending in March of given year, in California Climate District 2 (Sacramento Drainage, black) and Climate District 5 (San Joaquin Drainage, grey). Mean October-March precipitation values for 1901-2000 plotted as baseline (dashed lines). Map of California (right) shows climate districts, respective to their outflow into SF Bay (star). Red highlights years with abnormal shark deaths in the spring. Precipitation data from NOAA National Centers for Environmental Information.

2011, and 2017 (Figure 5.5; National Oceanographic and Atmospheric Administration National Centers for Environmental Information 2018). Although the specific conditions that led to the 2017 episode of scuticociliatosis are not fully known, it is worth noting that scuticociliates have frequently been observed in marine fish hatcheries in addition to wild fish populations. It would be prudent to better understand interactions between farmed and wild fish in order to identify risks to both populations. Further studies are needed to clarify susceptibility factors and exposures, especially in the context of major urban centers where planned human development could prevent or mitigate negative impacts of human activity on wild marine fish.

Future investigations of mass mortality events should include *M. avidus* as a potential pathogen. We anticipate that the episode of scuticociliatosis in wild elasmobranchs described here is not an isolated event. As similar epizootics are uncovered through seasonal monitoring, future research is needed to describe the host-pathogen relationship and potential implications for nearby human populations. Although the only known ciliate parasite of humans, *Balantidium coli*, is far distantly related to scuticociliates (Schuster and Ramirez-Avila 2008), and no scuticociliates have been proven to cause pathogenic disease in mammals, sport fishing and consumption of leopard sharks is common in SF Bay and the consequences of *M. avidus* ingestion are unknown. Finally, this study demonstrates the ability of mNGS to rapidly identify potential pathogens in an unbiased manner. Surveillance and disease investigations in wildlife populations will likely benefit from the incorporation of mNGS-based techniques.

Acknowledgements

We would like to acknowledge the Marine Science Institute (Redwood City, California), the Aquarium of the Bay (San Francisco, California), The Marine Mammal Center (Sausalito, California), East Bay Regional Park District (Oakland, California), the National Parks Service, Jennifer Kampe, and Paige Coluccio for their assistance in this work. We would also like to acknowledge Eric Chow and Derek Bogdanoff at the Center for Advanced Technology (University of California San Francisco) for assistance with sequencing. This work was supported by the Chan Zuckerberg Biohub (J.D.), UCSF Medical Scientist Training Program (H.R.), and the State of California (M.O.).

Addendum

Figure 5.6 Field specimen collection, 2017.

Left, collecting freshly-dead leopard shark carcasses south of the San Mateo Bridge, California. Right, performing field necropsies with Mark Okihiro to collect samples for the lab.



Further exploration of *Miamiensis avidus* biology

The investigation of *Miamiensis avidus* as a pathogen in sharks in the wild led me to several further questions. Most importantly, what route did this unicellular organism take to invade the shark brain? Was this entirely opportunistic, or driven by certain chemoattractants?

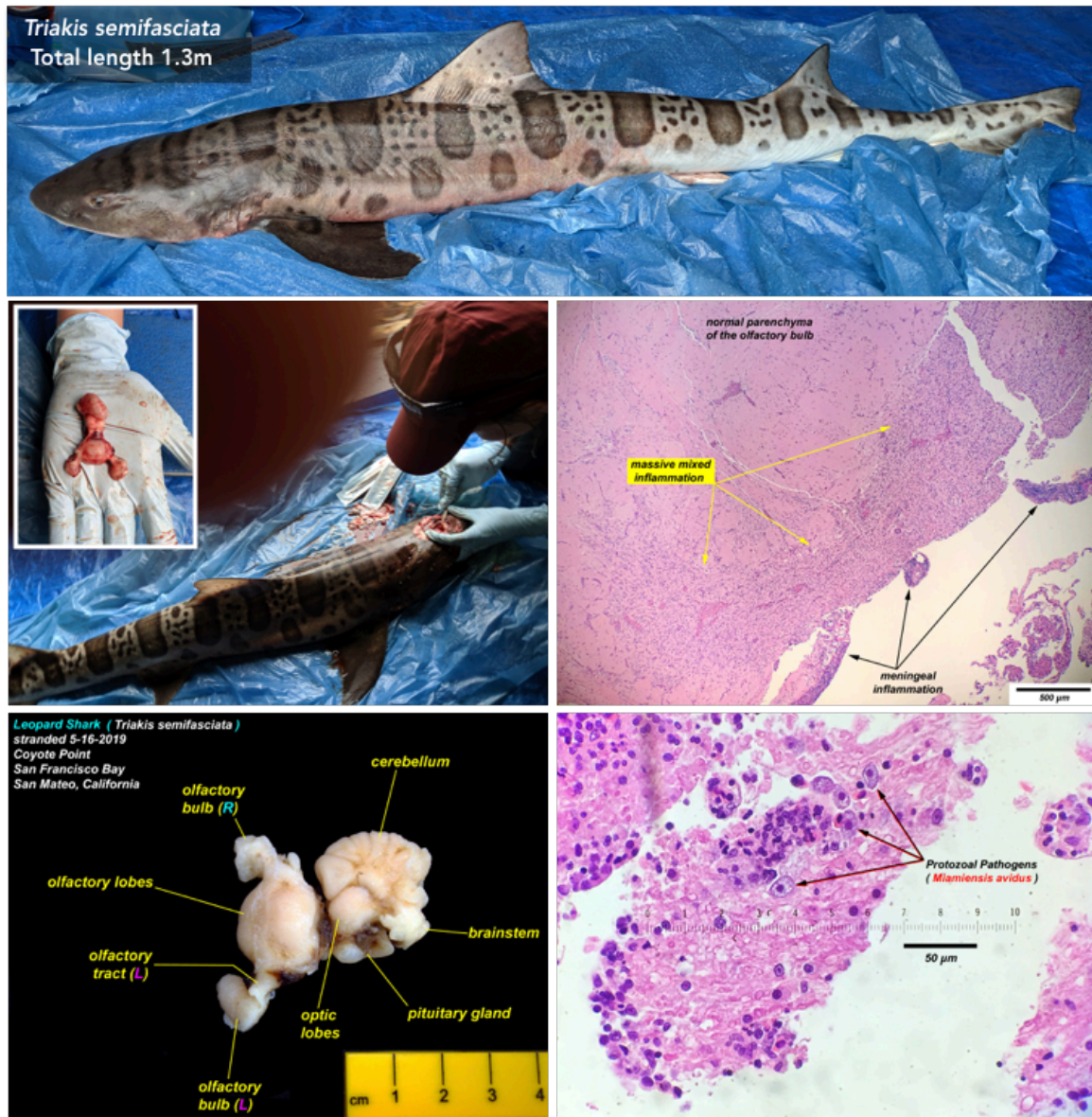


Figure 5.7 Moribund leopard shark (*Triakis semifasciata*) stranded in San Mateo, CA, 2019.

In the spring of 2019 (2 years following the initial sampling), many leopard sharks again stranded on the shores of San Francisco Bay. The individual shown here was found moribund at Coyote Point Marina, San Mateo, California. Upon dissection the brain was removed with intact olfactory bulbs and fixed in formalin. Histology (images and analysis courtesy of Mark Okihiro) showed substantial inflammation of the meninges and olfactory bulb, with a unicellular organism similar in appearance to the previously identified *Miamiensis avidus*.

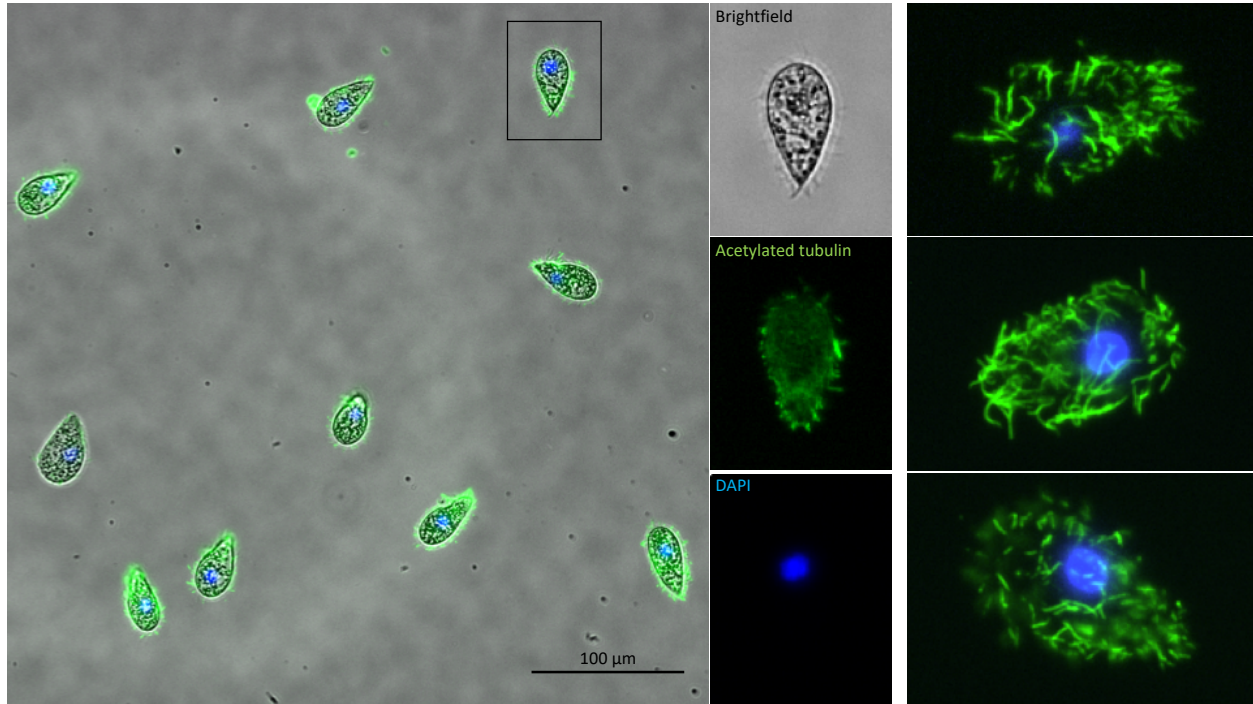


Figure 5.8 Culture of *Miamiensis avidus*, fixed and stained for acetylated tubulin and DAPI

To begin to answer these questions we cultured a lab strain of *Miamiensis avidus* (ATCC 50180) axenically, at room temperature in modified Leibovitz's L-15 medium (salinity 10%, pH 7.2, supplemented with 10% heat-inactivated FBS, containing 90 mg/L adenosine, cytidine and uridine, 150 mg/L guanosine, 5g/L glucose, and 1X Sigma Antibiotic Antimycotic Solution). We found the ciliates to be quite robust, permissive of starvation which induced different morphology (more banana-shaped with elongated and narrower cells), and tolerant of freezing in 5-10% DMSO with thawing to reinitiate culture. See Figure 5.8 for morphology of cultured *Miamiensis avidus*, where cilia are apparent by acetylated tubulin staining. We also visualized the macro- and micro-nucleus (Figure 5.9). This genomic organization, common to ciliates, poses a unique challenge to genome sequencing of this organism.

In a project completed together with Catherine Kawaja, we labeled live *Miamiensis avidus* by feeding them *E. coli* expressing an inducible GFP, and were able to visualize and track the ciliates' movements by the bright GFP in their food vacuoles. Working toward the goal of tracking live *Miamiensis avidus* during infection of a host fish to understand the route of invasion, we further assessed chemotaxis of *Miamiensis avidus* to various possible components of the brain, and found that complete tissue of zebrafish was a potent attractant, with yeast extract peptone dextrose and media with high lipid content being a far greater attractant than glucose or the control. It will be fascinating to understand how this unicellular organism invades the shark, burrowing through tissue to reach the brain where it appears to cause severe damage and eventual death with only scurrilous regard for its host.

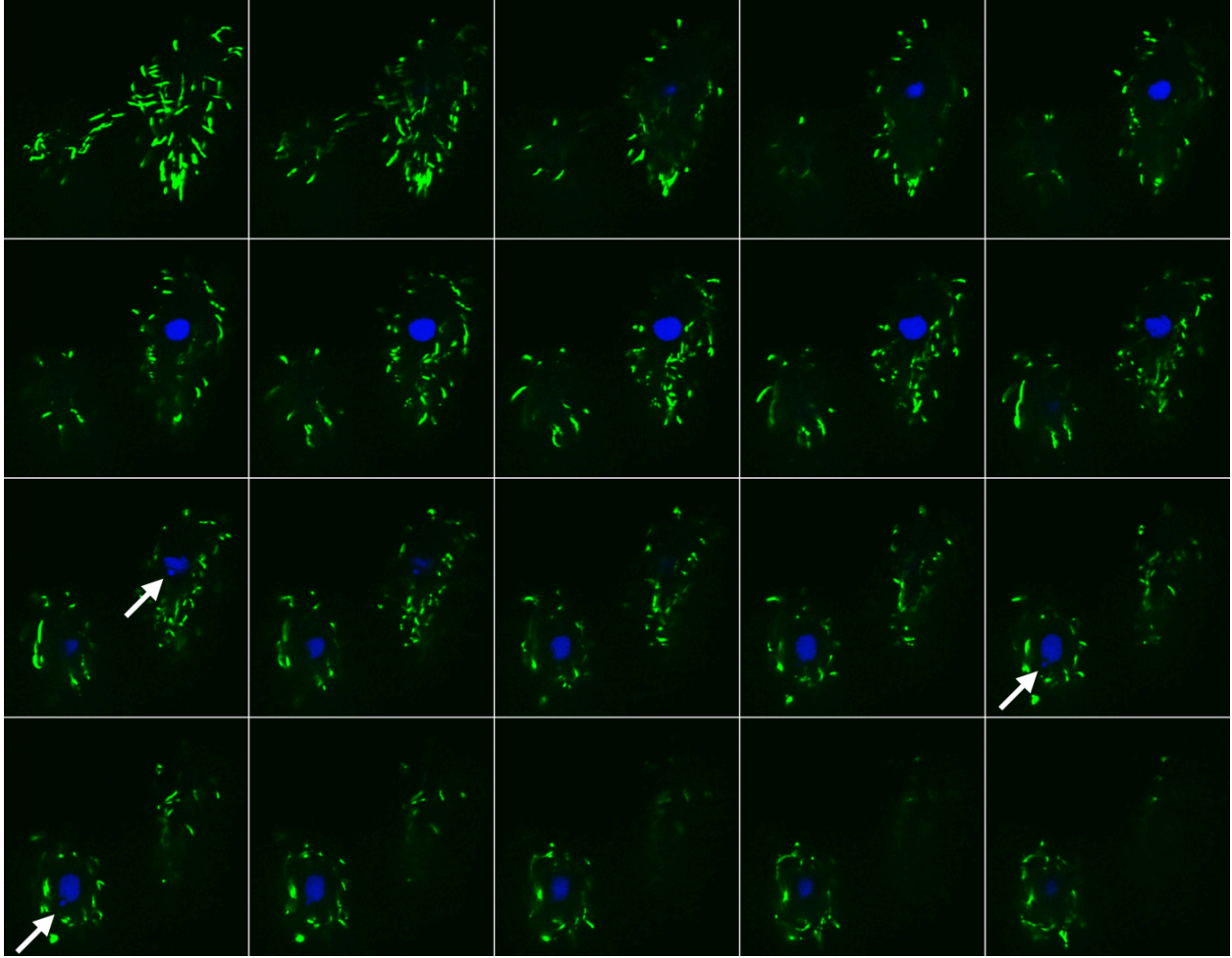


Figure 5.9 Z-stack of cultured *Miamiensis avidus*, showing micronucleus.

Z stack of 2 fixed *Miamiensis avidus* cells (left to right, top to bottom), stained for acetylated tubulin (green) and DAPI (blue). For each organism, the micronucleus is visible in just 1-2 planes (arrows).

References for Chapter 5

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215, 403–410.
- Budiño, B., Lamas, J., Pata, M.P., Arranz, J.A., Sanmartín, M.L., and Leiro, J. (2011). Intraspecific variability in several isolates of *Philasterides dicentrarchi* (syn. *Miamiensis avidus*), a scuticociliate parasite of farmed turbot. *Veterinary Parasitology* 175, 260–272.
- Carlisle, A., and Starr, R. (2009). Habitat use, residency, and seasonal distribution of female leopard sharks *Triakis semifasciata* in Elkhorn Slough, California. *Mar. Ecol. Prog. Ser.* 380, 213–228.
- De Felipe, A.-P., Lamas, J., Sueiro, R.-A., Folgueira, I., and Leiro, J.-M. (2017). New data on flatfish scuticociliatosis reveal that *Miamiensis avidus* and *Philasterides dicentrarchi* are different species. *Parasitology* 144, 1394–1411.
- Dervas, E., Hepojoki, J., Laimbacher, A., Romero-Palomo, F., Jelinek, C., Keller, S., Smura, T., Hepojoki, S., Kipar, A., and Hetzel, U. (2017). Nidovirus-Associated Proliferative Pneumonia in the Green Tree Python (*Morelia viridis*). *J Virol* 91, e00718-17, e00718-17.
- Gao, F., Katz, L.A., and Song, W. (2012). Insights into the phylogenetic and taxonomy of philasterid ciliates (Protozoa, Ciliophora, Scuticociliatia) based on analyses of multiple molecular markers. *Molecular Phylogenetics and Evolution* 64, 308–317.

- Gao, F., Warren, A., Zhang, Q., Gong, J., Miao, M., Sun, P., Xu, D., Huang, J., Yi, Z., and Song, W. (2016). The All-Data-Based Evolutionary Hypothesis of Ciliated Protists with a Revised Classification of the Phylum Ciliophora (Eukaryota, Alveolata). *Sci Rep* 6, 24874.
- Garza, J.B., Bott, N.J., Hammond, M.D., Shepherd, N., and Nowak, B.F. (2017). Molecular characterisation of *Miamiensis avidus* (Ciliophora: Scuticociliata) from farmed Southern bluefin tuna, *Thunnus maccoyii* off Port Lincoln, South Australia. *Aquaculture* 469, 44–49.
- Hight, B.V., and Lowe, C.G. (2007). Elevated body temperatures of adult female leopard sharks, *Triakis semifasciata*, while aggregating in shallow nearshore embayments: Evidence for behavioral thermoregulation? *Journal of Experimental Marine Biology and Ecology* 352, 114–128.
- Hopkins, T.E., and Cech, J.J. (2003). The Influence of Environmental Variables on the Distribution and Abundance of Three Elasmobranchs in Tomales Bay, California. *Environmental Biology of Fishes* 66, 279–291.
- Iglesias, R., Paramá, A., Álvarez, M.F., Leiro, J., Aja, C., and Sanmartín, M.L. (2003). In vitro growth requirements for the fish pathogen *Philasterides dicentrarchi* (Ciliophora, Scuticociliatida). *Veterinary Parasitology* 111, 19–30.
- Jung, S., Kitamura, S., Song, J., Joung, I., and Oh, M. (2005). Complete small subunit rRNA gene sequence of the scuticociliate *Miamiensis avidus* pathogenic to olive flounder *Paralichthys olivaceus*. *Dis. Aquat. Org.* 64, 159–162.

- Jung, S.-J., Im, E.-Y., Strüder-Kypke, M.C., Kitamura, S.-I., and Woo, P.T.K. (2011). Small subunit ribosomal RNA and mitochondrial cytochrome c oxidase subunit 1 gene sequences of 21 strains of the parasitic scuticociliate *Miamiensis avidus* (Ciliophora, Scuticociliatia). *Parasitol Res* 108, 1153–1161.
- Kubinski, S., Affolter, V., Groff, J., and Weber, E. Pathology 11N1368 Final Report (University of California Davis, Veterinary Medical Teaching Hospital).
- Li, W., Lo, C., Su, C., Kuo, H., Lin, S., Chang, H., Pang, V., and Jeng, C. (2017). Locally extensive meningoencephalitis caused by *Miamiensis avidus* (syn. *Philasterides dicentrarchi*) in a zebra shark. *Dis. Aquat. Org.* 126, 167–172.
- Moustafa, E.M.M., Tange, N., Shimada, A., and Morita, T. (2010). Experimental Scuticociliatosis in Japanese Flounder (*Paralichthys olivaceus*) Infected with *Miamiensis avidus*: Pathological Study on the Possible Neural Routes of Invasion and Dissemination of the Scuticociliate inside the Fish Body. *J. Vet. Med. Sci.* 72, 1557–1563.
- Munday, B., O'Donoghue, P., Watts, M., Rough, K., and Hawkesford, T. (1997). Fatal encephalitis due to the scuticociliate *Uronema nigricans* in sea-caged, southern bluefin tuna *Thunnus maccoyii*. *Dis. Aquat. Org.* 30, 17–25.
- Nosal, A.P., Cartamil, D.C., Long, J.W., Lührmann, M., Wegner, N.C., and Graham, J.B. (2013). Demography and movement patterns of leopard sharks (*Triakis semifasciata*) aggregating near the head of a submarine canyon along the open coast of southern California, USA. *Environ Biol Fish* 96, 865–878.

- R. Iglesias, A. ParamÃ¡s, M. F. Alvarez, J. Leiro, J. FernÃ¡ndez, and M. L. SanmartÃ­n (2001). *Philasterides dicentrarchi* (Ciliophora, Scuticociliatida) as the causative agent of scuticociliatosis in farmed turbot *Scophthalmus maximus* in Galicia (NW Spain). *Dis Aquat Org* 46, 47–55.
- Ramos, M., Costa, A., Barandela, T., Saraiva, A., and Rodrigues, P. (2007). Scuticociliate infection and pathology in cultured turbot *Scophthalmus maximus* from the north of Portugal. *Dis. Aquat. Org.* 74, 249–253.
- Russo, R.A. (2015). Observations of predation and loss among leopard sharks and brown smoothhounds in San Francisco Bay, California. *Calif Fish Game* 101, 149–157.
- Russo, R.A., and HERALD, E.S. (1968). The 1967 shark kill in San Francisco Bay. *CALIFORNIA FISH AND GAME, VOL 54, NO 3, P 215-216, 1968.*
- Schaffer, P.A., Lifland, B., Sommeran, S.V., Casper, D.R., and Davis, C.R. (2013). Meningoencephalitis Associated With *Carnobacterium maltaromaticum* –Like Bacteria in Stranded Juvenile Salmon Sharks (*Lamna ditropis*). *Vet Pathol* 50, 412–417.
- Schuster, F.L., and Ramirez-Avila, L. (2008). Current World Status of *Balantidium coli*. *CMR* 21, 626–638.
- Stidworthy, M.F., Garner, M.M., Bradway, D.S., Westfall, B.D., Joseph, B., Repetto, S., Guglielmi, E., Schmidt-Posthaus, H., and Thornton, S.M. (2014). Systemic Scuticociliatosis (*Philasterides dicentrarchi*) in Sharks. *Vet Pathol* 51, 628–632.

- Takagishi, N., Yoshinaga, T., and Ogawa, K. (2009). Effect of hyposalinity on the infection and pathogenicity of *Miamiensis avidus* causing scuticociliatosis in olive flounder *Paralichthys olivaceus*. *Dis. Aquat. Org.* 86, 175–179.
- Whang, I., Kang, H.-S., and Lee, J. (2013). Identification of scuticociliates (*Pseudocohnilembus persalinus*, *P. longisetus*, *Uronema marinum* and *Miamiensis avidus*) based on the *cox1* sequence. *Parasitology International* 62, 7–13.
- Wilson, M.R., Naccache, S.N., Samayoa, E., Biagtan, M., Bashir, H., Yu, G., Salamat, S.M., Somasekar, S., Federman, S., Miller, S., et al. (2014). Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing. *N Engl J Med* 370, 2408–2417.
- Zylberberg, M., Van Hemert, C., Dumbacher, J.P., Handel, C.M., Tihan, T., and DeRisi, J.L. (2016). Novel Picornavirus Associated with Avian Keratin Disorder in Alaskan Birds. *MBio* 7, e00874-16, /mbio/7/4/e00874-16.atom.
- (2018a). Basic local alignment search tool (National Center for Biotechnology Information).
- (2018b). Climate at a glance: Divisional time series (National Oceanographic and Atmospheric Administration National Centers for Environmental Information).

Chapter 6 Genome Sequence of a Divergent Avian

Metapneumovirus from a Monk Parakeet

Authors:

Hanna Retallack^a, Susan Clubb^b, Joseph L. DeRisi^{a,c}

Affiliations:

^a Department of Biochemistry and Biophysics, University of California, San Francisco, CA, USA

^b Rainforest Clinic for Birds and Exotics, Loxahatchee, FL, USA

^c Chan Zuckerberg Biohub, San Francisco, CA, USA

Includes material previously published in:

Retallack H, Clubb S, DeRisi JL. Genome Sequence of a Divergent Avian

Metapneumovirus from a Monk Parakeet (*Myiopsitta monachus*). *Microbiol Resour*

Announc. 2019;8(16):e00284-19. Published 2019 Apr 18. doi:10.1128/MRA.00284-19

Abstract

Here, we report the coding-complete genome sequence of an avian metapneumovirus from a monk parakeet (*Myiopsitta monachus*), identified by metagenomic next-generation sequencing during an investigation into a disease outbreak in a captive parrot breeding facility. Based on divergence from known strains, this sequence represents a new subgroup of avian metapneumovirus.

Metapneumoviruses (genus *Metapneumovirus*, family *Pneumoviridae*) cause disease in birds and mammals, including humans. Avian strains are important pathogens in commercial poultry, causing acute upper respiratory illness often complicated by secondary bacterial infections in chickens and turkeys (J. K. A. Cook 2000; Majó et al. 1997). We observed an unusual cluster of morbidity and mortality among young parrots at a captive breeding facility which could not be explained by routine diagnostics. Difficult-to-control bacterial infections and persistent cryptosporidium infections suggested immunosuppression. This prompted our investigation into underlying infectious etiologies using metagenomic next-generation sequencing.

At necropsy of an affected monk parakeet chick, the lung, liver, and spleen were sampled and stored at -80°C. For RNA extraction, ~50mg of combined tissues was homogenized in 2mL DNA/RNA Shield (Zymo Research) using 2.8mm ceramic beads (Omni) on a TissueLyser II (Qiagen) with 5 cycles of 30Hz for 30sec followed by 1min on ice. Samples were centrifuged at 16,000xg for 10min, and 250µL of homogenized tissue supernatant was added to 750µL Direct-zol (Zymo Research). RNA was extracted using a Direct-zol RNA MiniPrep Plus kit (Zymo Research) with DNase treatment (Qiagen) and quantified by Nanodrop. Sequencing libraries were prepared from 100ng of extracted RNA with 25pg of spike-in control RNA from the External RNA Controls Consortium (ERCC) collection (Thermo Fisher Scientific), using NEBNext Ultra II Directional RNA Lib Prep Kit for Illumina (New England BioLabs). A water sample was processed in parallel. Paired-end 150nt sequencing on an Illumina HiSeq 4000 yielded 35,053,607 raw read-pairs.

A representative host database was built using all genome assemblies and mitochondrial genomes under TaxID 9224 (*Psittacidae*, parrots) in the National Center for Biotechnology Information (NCBI) database as of December 7, 2018. Host subtraction and quality control were performed as described previously (Retallack et al. 2018). The remaining 1,670,686 unique non-host read-pairs (4.8% of raw) were processed using the IDseq pathogen detection pipeline (v3.2, reference nt/nr database December 1, 2018) (Ramesh et al. 2018), which identified metapneumovirus reads in the sample. No other viruses were detected as credible hits by the following criteria: ≥ 10 mapped read-pairs per million non-host read-pairs (rpM) at the nucleotide level, and ≥ 1 rpM at the amino acid level.

These metapneumovirus reads were used as seeds for Paired-End Iterative Contig Extension (PRICE v1.2, settings “-fpp <R1> <R2> 350 99 -mol 30 -target 80 8 2 2 -nc 10 -lenf 500 8”) to assemble the full-length genome (Ruby, Bellare, and DeRisi 2013). Reads were then mapped back to the genome using Bowtie 2 (v2.2.4, “--very-sensitive-local” mode) (Langmead and Salzberg 2012). The final consensus sequence is 13,648nt in length, with 26x mean coverage and 39% GC content (Figure 1). Genome termini were not specifically identified. Consistent with active viral replication, we observed reads from both negative-strand (genomic) and positive-strand (mRNA transcript/anti-genomic) RNA.

The most similar sequences in the NCBI’s nt and nr reference databases by BLAST were metapneumoviruses (Stephen F. Altschul et al. 1990). Phylogenetic analysis of the L gene (RNA-dependent RNA polymerase) at the amino acid level revealed 43-49% identity to representative members of the genus *Orthopneumovirus* and 61-66% identity to

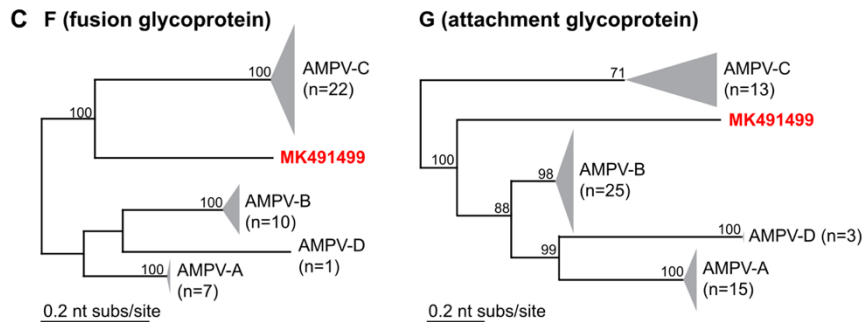
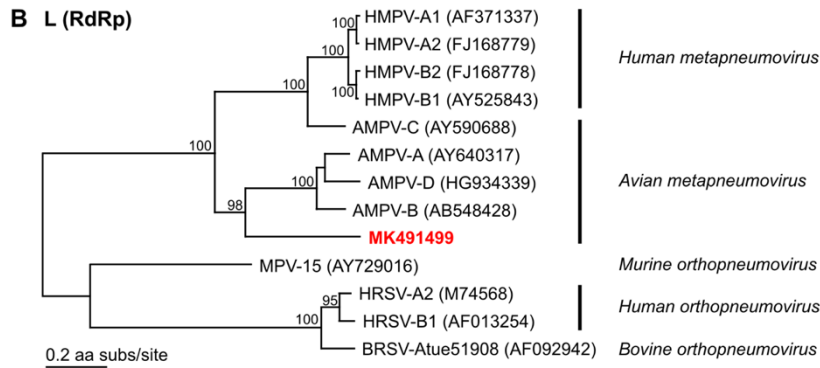
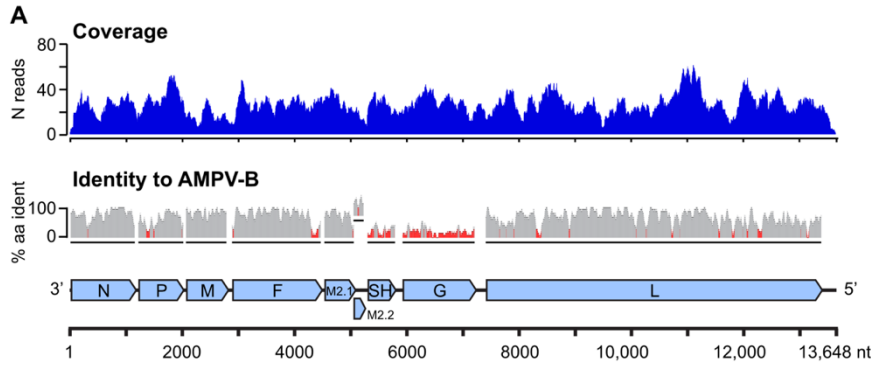


Figure 6.1 Coverage and phylogenetic analysis of sequence representing a new subgroup of avian metapneumovirus.

(A) Top, coverage plot showing number of reads aligning to the consensus sequence (y-axis) along the length of the consensus sequence (x-axis, length in nucleotides, corresponding to the diagram of the viral genome below). Middle, percent identity (y-axis) for a 15-amino acid sliding window across an alignment of the consensus sequence and reference AMPV-B sequence (GenBank accession number AB548428) for each viral protein. Red bars indicate an identity of <30%. Bottom,

representation of likely genomic structure based on open reading frames and homology to other avian metapneumoviruses. (B) Phylogenetic tree of the Pneumoviridae. Amino acid level alignments of L genes (encoding RNA-dependent RNA polymerase [RdRp]) from representative viruses were used to construct a maximum likelihood tree. Multiple-sequence alignment was performed in Geneious (v9.1.8) with default parameters; the phylogenetic tree was built using PhyML v2.2.3 (LG substitution model, 100 bootstraps) (9). The sequence identified in this study is highlighted in red. Values at branch points indicate the fraction of trees with this node, based on a bootstrapping method. Bar, 0.2 amino acid substitutions per site. (C) Maximum likelihood trees (PhyML, default parameters) constructed from nucleotide alignments (Geneious, default parameters) of all available avian metapneumovirus sequences for the fusion glycoprotein (F gene, left) and attachment glycoprotein (G gene, right). Bar, 0.2 nucleotide substitutions per site.

representative members of the genus *Metapneumovirus*, indicating that this sequence represents a new subgroup of metapneumoviruses (Figure 6.1) (Rima et al. 2017). Analysis of the F gene (fusion glycoprotein) and G gene (attachment glycoprotein) further supported this classification.

We have identified the first member of a new subgroup of metapneumoviruses, distinct from avian metapneumoviruses A, B, C, and D. Despite similarities between this outbreak and outbreaks of avian metapneumovirus in commercial poultry, it remains unknown whether the virus identified here directly caused the symptoms observed in this individual and/or flock.

Data availability.

The avian metapneumovirus sequence described here has been deposited at GenBank under the accession number MK491499.

Acknowledgements.

We thank Eric Chow at the UCSF Center for Advanced Technology for assistance with sequencing. This research was supported by the Chan Zuckerberg Biohub (J.L.D.).

Addendum

Following the initial discovery of a new strain of avian metapneumovirus, aMPV-E, I validated a diagnostic PCR (Table 6.1, Figure 6.2), used to screen multiple samples from later outbreaks in an attempt to identify samples for viral isolation. I also used this assay to identify key organs targeted by the virus (Figure 6.3, Figure 6.4).

Table 6.1: Primers for diagnostic testing for aMPV-E (based on MK49199 genome sequence)

Primer Name	Sequence	Tm (C)	Genome L position	Genome R position	Gene	Amplicon length (nt)
oHR3000	AAGGCATTGGGCTCGTCTTC	60	702	721	N	102
oHR3001	CACCCACCTCAGCATAGTC	60	785	804		
oHR3002	TGCTGCACGAGATGGAATCA	60	1,775	1,794	P	141
oHR3003	ACTGCCATTTCCGATCTTGC	60	1,897	1,916		
oHR3010	CATAGCGCGTCTTTTGGTGG	60	2,811	2,830	M	107
oHR3011	AAAAGCTAGGGGCAGGAACC	60	2,723	2,742		
oHR3008	TCCGGTCTGTCTGTTTCGTG	60	6,687	6,706	G	170
oHR3009	CACAACTGCGAAACCGATCC	60	6,536	6,555		
oHR3006	ACACTGCTCTCGGTGATTGG	60	13,114	13,133	L	148
oHR3007	CCGGGCTATTCAGAGGTGAG	60	12,985	13,004		

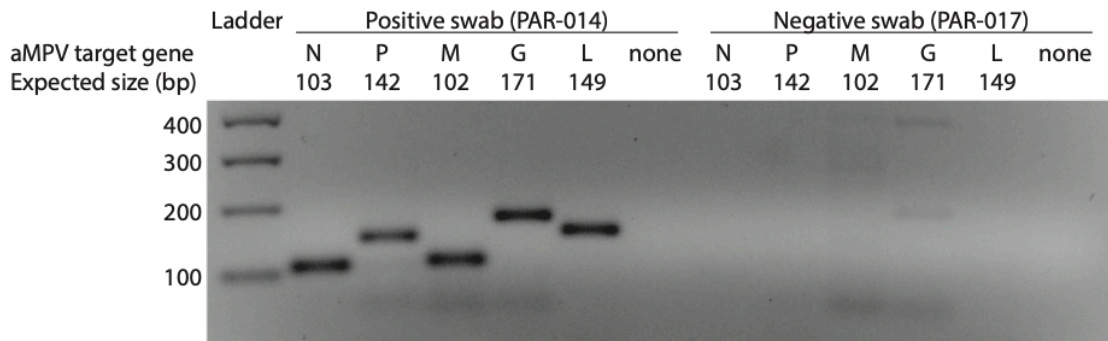
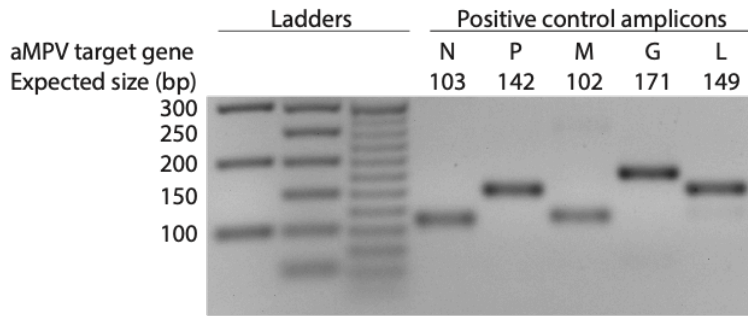
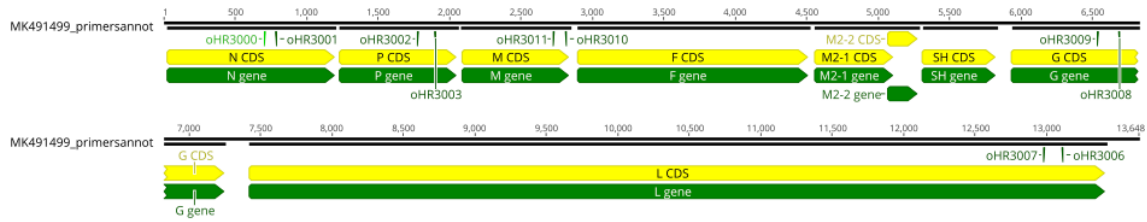


Figure 6.2 Validation of Diagnostic PCR for aMPV-E.

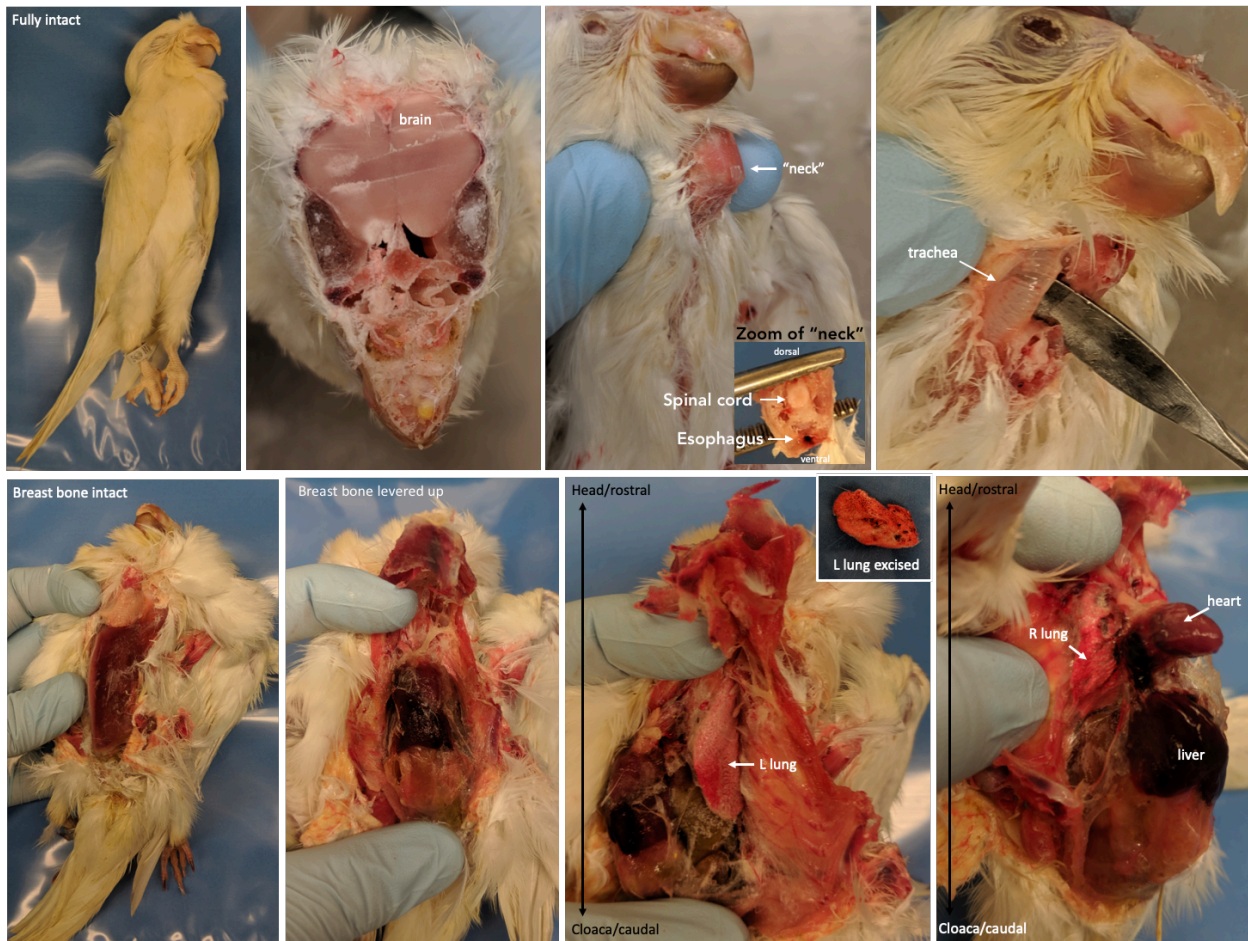


Figure 6.3: Dissection of Quaker parrot (monk parakeet, *Myiopsitta monachus*) that had died during an outbreak of aMPV-E following sudden onset illness.

Organs sampled include brain, "neck" (containing spinal cord and esophagus), trachea, L and R lung, heart, liver, and not shown here: intestines, kidney, and the globule-like bursa of Fabricus located adjacent to the cloaca and involved in immune response.

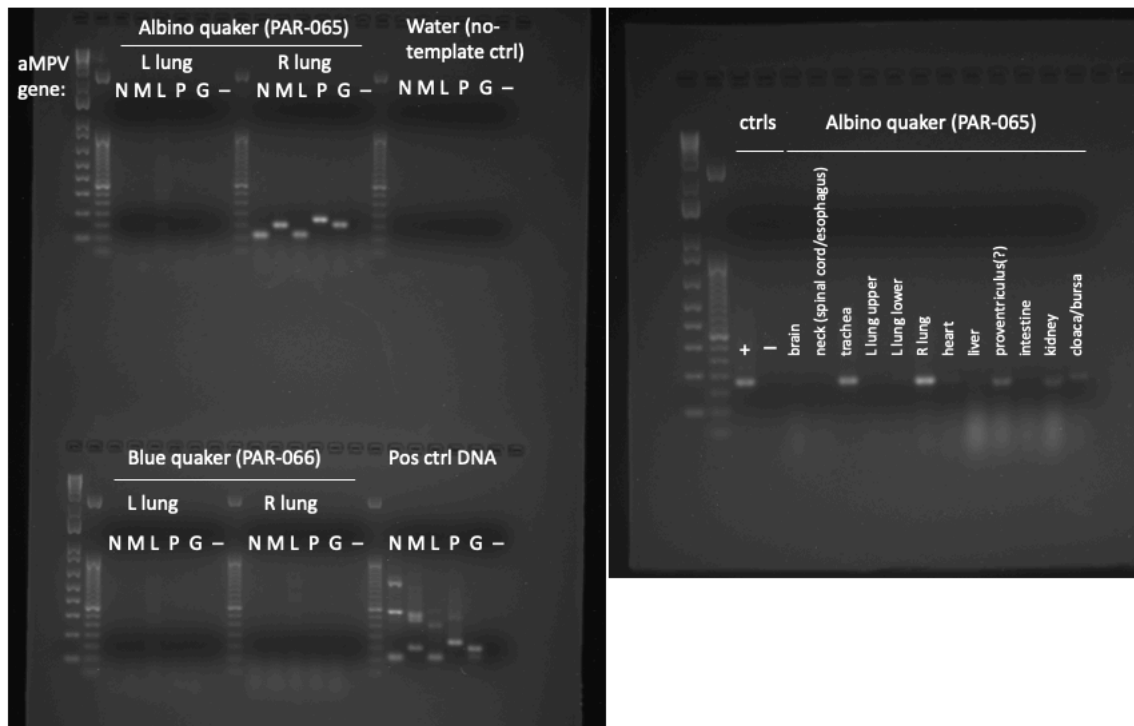


Figure 6.4: PCR identifies aMPV in additional parrots, by organ.

Left: panel of primer pairs to aMPV on each lung of two parrots (PAR-065 and PAR-066).

Right: glycoprotein (G) primer pair across organs from PAR-065, shown during dissection in Figure 6.3.

References for Chapter 6

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215, 403–410.
- Cook, J.K.A. (2000). Avian Pneumovirus Infections of Turkeys and Chickens. *The Veterinary Journal* 160, 118–125.
- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* 59, 307–321.
- Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9, 357–359.
- Majó, N., Gibert, X., Vilafranca, M., O’Loan, C.J., Allan, G.M., Costa, Ll., Pagès, A., and Ramis, A. (1997). Turkey rhinotracheitis virus and *Escherichia coli* experimental infection in chickens: histopathological, immunocytochemical and microbiological study. *Veterinary Microbiology* 57, 29–40.
- Ramesh, A., Nakielny, S., Hsu, J., Kyohere, M., Byaruhanga, O., Bourcy, C. de, Egger, R., Dimitrov, B., Juan, Y.-F., Sheu, J., et al. (2018). Etiology of fever in Ugandan children: identification of microbial pathogens using metagenomic next-generation sequencing and IDseq, a platform for unbiased metagenomic analysis (Genomics).
- Retallack, H., Okihiro, M.S., Britton, E., Sommeran, S.V., and DeRisi, J.L. (2019). METAGENOMIC NEXT-GENERATION SEQUENCING REVEALS MIAMIENSIS

AVIDUS (CILIOPHORA: SCUTICOCILIATIDA) IN THE 2017 EPIZOOTIC OF
LEOPARD SHARKS (TRIAKIS SEMIFASCIATA) IN SAN FRANCISCO BAY,
CALIFORNIA, USA. *Journal of Wildlife Diseases* 55, 375.

Rima, B., Collins, P., Easton, A., Fouchier, R., Kurath, G., Lamb, R.A., Lee, B., Maisner, A.,
Rota, P., Wang, L., et al. (2017). ICTV Virus Taxonomy Profile: Pneumoviridae.
Journal of General Virology 98, 2912–2913.

Ruby, J.G., Bellare, P., and DeRisi, J.L. (2013). PRICE: Software for the Targeted Assembly
of Components of (Meta) Genomic Sequence Data. *G3* 3, 865–880.

Chapter 7 Wildlife investigations with no infectious agent identified by mNGS

Contributions

Includes contributions from Padraig Duignan and Barbie Halaska (Marine Mammal Center, specimen collection and photo documentation, seal), Jill Murray (Oklahoma State University, specimen collection, frogs), David Guzman, Sarah Ozawa, and Tracy Drazenovich (UC Davis, specimen collection and photo documentation, rabbit), Eric Chow and Derek Bogdanoff (UCSF CAT, sequencing).

The investigations on die-offs of animals in Chapter 5 (sharks) and Chapter 6 (parrots) proved fruitful in identifying likely infectious etiologies. In these cases, pathology was severe, the course of illness was definable, and a cluster of affected animals was available for study. Of course, in many situations metagenomic next-generation sequencing (mNGS) does not find a pathogenic culprit. While it can be easy to miss infections if sample collection, handling, or other technical aspects are not optimal, a negative result can also yield valuable information that directs the investigation toward other possible causes of death. Here follow three examples where I found no explanation for the animals' deaths by mNGS.

Toboggan, Ribbon Seal (*Histiophoca fasciata*)

Ribbon seals are normally found on the Arctic pack ice off Alaska and Russia. In December of 2017, staff at the Marine Mammal Center rescued a female juvenile Ribbon seal, Toboggan, that had stranded at Morro Bay, California, with severe alopecia and pneumonia (Fig 7.1) (Duignan, Pdraig 2018). Despite initial recovery, this individual suddenly developed bloody diarrhea and died with hemorrhagic enteritis and colitis and identification of *Clostridium perfringens* enterotoxin A in the colon.

Given the natural range of this species, this was a rare opportunity to investigate a broader phenomenon if this death was related to the unusual mortality event in Alaska in 2011 and 2016, where many seals were observed to have similar forms of alopecia characterized by a defluxion event (when the hair growth cycle abruptly halts) (Burgess, Tristan 2013). No conclusive cause was identified at that time.

Specimens taken from this animal included blood, skin (top of head), and spleen. Next-generation sequencing libraries were prepared from extracted RNA from these samples, yielding 17-32 million read-pairs (Table 7.1)

As no genome existed for this species at the time, to subtract host sequences a database was built from two full genomes, for *Leptonychotes weddellii* (Weddell seal) and *Neomonachus schauinslandi* (Hawaiian monk seal), as well as 13 mitochondrial genomes for related seals within the *Phocidae* (True Seals, taxID 9709). In the computational analysis (methods akin to (Chapter 5 and Chapter 6), 95-97% of the reads were identified as belonging to seal. The remaining non-seal reads were compared to the NCBI database of all nucleotide and



Figure 7.1 Alopecia in female ribbon seal *Toboggan* (*Histriophoca fasciata*), stranded at Morro Bay December 2017.

Photos courtesy of Padraig Duignan (Marine Mammal Center).

Table 7.1 Sequencing metrics for ribbon seal

sample ID	sample type	Total read-pairs	% pass QC	Unique, non-seal reads (includes ERCCs)
SEA1	blood	17658458	87	791609 (4.48 %)
SEA2	spleen	17464057	87	424102 (2.43 %)
SEA3	skin	32285242	82	734354 (2.27 %)
SEA4	NTC	197997	87	113688 (57.42 %)

protein sequences. In the skin sample, a small number of sequences likely derived from *Proteus mirabilis* were observed. In the blood sample, there were 3-4 read-pairs to each of a papillomavirus and an adenovirus. Such low levels are rarely relevant unless truly pathogenic, and these particular viral families have been seen in seals before (Smeele et al. 2018; Chiappetta et al. 2017; Cortés-Hinojosa et al. 2016). Other than these findings, mNGS identified no fungi, eukaryotes, or other viruses or bacteria of note.

Lizzy, Dutch Dwarf Rabbit (Oryctolagus cuniculus)

Papillomatous anorectal masses are relatively common in rabbits, including household pets. While it is believed that such masses may have viral origins, for only a subset has a viral etiology been identified (Shope and Hurst 1933). We received a flash-frozen sample from excision of such a mass in a Dutch Dwarf rabbit in 2017 (Fig 7.2), and processed for mNGS. The data revealed >99.99% rabbit, with no viruses or other pathogens identified.



Figure 7.2 Papillomatous anorectal mass of uncertain etiology in a Dutch Dwarf rabbit (*Oryctolagus cuniculus*).

Photo courtesy of David Guzman (UC Davis).

Blanchard's Cricket Frogs (Acris blanchardi)

We were contacted by a researcher whose colony of wild-caught Blanchard's Cricket Frogs (*Acris blanchardi*) experienced acute and unexplainable mortality. Tadpoles were caught in the spring of 2016 and raised in captivity, developing normally until October, when mortality suddenly approached 30-50% across the colony. Animals became thin in body condition, anorexic, and dehydrated before death. All routine analysis and testing was non-contributory, including standard cultures, and water analysis. Histopathology of a subset of animals revealed mild lymph node distension with edema and seriously atrophic adipose stores.

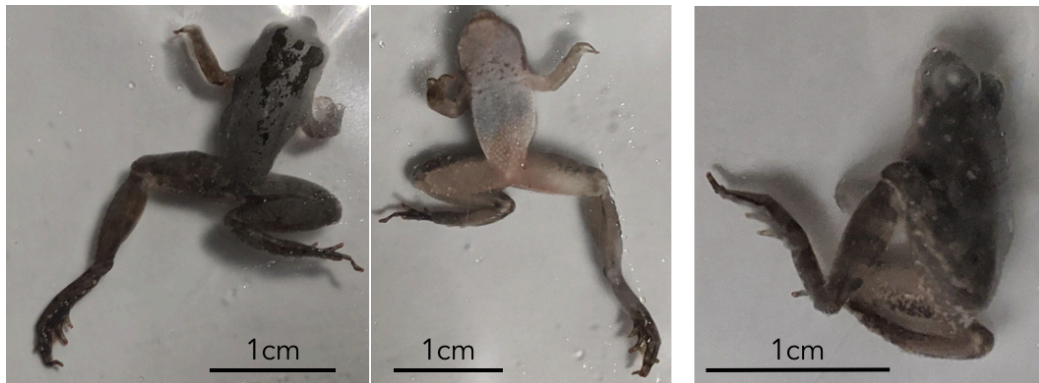


Figure 7.3 Blanchard's Cricket Frogs to be processed for mNGS.

Left and center images are opposing views of the same individual.

We processed samples from 1 asymptomatic animal and 4 diseased animals sacrificed in December 2016, and 1 water control (Figure 7.3). Given the clinical picture and common pathogens that are transmitted on the skin in such animals, we sampled the abdomen, and separately prepared libraries from the digits (1 front and 1 rear foot) and scraped skin from the inner thighs and abdomen. Standard mNGS libraries were prepared akin to Chapter 5 and Chapter 6, yielding metrics in Table 7.2.

Table 7.2 Sequencing metrics for frogs

Sample ID	Sample Type	N raw readpairs	Post-QC % non-host
FRG-D1-abd_S75	Diseased; abdomen	27842334	7.33
FRG-D1-skd_S74	Diseased, skin + digits	24545918	10.52
FRG-D2-abd_S77	Diseased; abdomen	24840430	6.58
FRG-D2-skd_S76	Diseased, skin + digits	30366389	8.21
FRG-D3-abd_S79	Diseased; abdomen	30921287	4.38
FRG-D3-skd_S78	Diseased, skin + digits	28505738	9.83
FRG-D6-abd_S73	Diseased; abdomen	25459198	4.50
FRG-D6-skd_S72	Diseased, skin + digits	27348784	5.17
FRG-H1-abd_S71	Healthy; abdomen	23687950	7.43
FRG-H1-skd_S70	Healthy; skin + digits	1990991	11.82
FRG-H2O_S69	H2O	2045	11.05

As no complete genome existed for this species at the time, we compiled a host subtraction database from related species including the standard laboratory model animal, *Xenopus tropicalis*, as well as *Nanorana parkeri*, and 10 related mitochondrial genomes in order to identify frog-like sequences.

No overwhelming bacterial or fungal pathogens were observed. The bacteria *Aeromonas* was seen in 3 sick animals, and had been cultured by the veterinarian overseeing the colony, but was determined to be low-level environmental and not clinically relevant in this situation. No evidence of common frog pathogens was found, including chytrid, mucor, rhizopus, saprolegnia, ranaviruses, herpesviruses or pathogenic eukaryotes. No viruses were found, apart from Drosophila A virus -- all animals were fed drosophila, and there is no known pathogenicity of DAV in amphibians. In conclusion, we were unable to identify an infectious etiology for this colony's mass mortality.

References for Chapter 7

- Burgess, Tristan (2013). Investigation of a Pinniped Skin Disease Outbreak in the Arctic and Bering Sea Regions. p.
- Chiappetta, C.M., Cibulski, S.P., Lima, F.E.S., Varela, A.P.M., Amorim, D.B., Tavares, M., and Roehe, P.M. (2017). Molecular Detection of Circovirus and Adenovirus in Feces of Fur Seals (*Arctocephalus* spp.). *Ecohealth* 14, 69–77.
- Cortés-Hinojosa, G., Doescher, B., Kinsel, M., Lednicky, J., Loeb, J., Waltzek, T., and Wellehan, J.F.X.J. (2016). COINFECTION OF CALIFORNIA SEA LION ADENOVIRUS 1 AND A NOVEL POLYOMAVIRUS IN A HAWAIIAN MONK SEAL (*NEOMONACHUS SCHAUINSLANDI*). *J Zoo Wildl Med* 47, 427–437.
- Duignan, Pdraig (2018). Alopecia in an Extra-limital Ribbon Seal (*Histiophoca fasciata*) Stranded in Central California: Similarities and Differences to the Alaskan Northern Pinniped Unusual Mortality Event (UME) Alopecia and Dermatitis Syndrome. (Long Beach, CA), p.
- Shope, R.E., and Hurst, E.W. (1933). INFECTIOUS PAPILLOMATOSIS OF RABBITS : WITH A NOTE ON THE HISTOPATHOLOGY. *J Exp Med* 58, 607–624.
- Smeele, Z.E., Burns, J.M., Van Doorsaler, K., Fontenele, R.S., Waits, K., Stainton, D., Shero, M.R., Beltran, R.S., Kirkham, A.L., Bergartt, R., et al. (2018). Diverse papillomaviruses identified in Weddell seals. *J Gen Virol* 99, 549–557.

Chapter 8 Single Mosquito Metatranscriptomics Recovers

Mosquito Species, Blood Meal Sources, and Microbial Cargo,

Including Viral Dark Matter

Authors:

Joshua Batson^{1†}, Gytis Dudas^{3†}, Eric Haas-Stapleton^{2†}, Amy L. Kistler^{1*†}, Lucy M Li^{1†}, Phoenix Logan^{1†}, Kalani Ratnasiri^{1,4†}, Hanna Retallack^{5†}

Manuscript in preparation, publicly available as:

Joshua Batson, Gytis Dudas, Eric Haas-Stapleton, Amy L. Kistler, Lucy M. Li, Phoenix Logan, Kalani Ratnasiri, Hanna Retallack. Single mosquito metatranscriptomics recovers mosquito species, blood meal sources, and microbial cargo, including viral dark matter
bioRxiv 2020.02.10.942854; doi: <https://doi.org/10.1101/2020.02.10.942854>

Abstract

Mosquitoes are a disease vector with a complex ecology involving interactions between transmissible pathogens, endogenous microbiota, and human and animal blood meal sources. Unbiased metatranscriptomic sequencing of individual mosquitoes offers a straightforward and rapid way to characterize these dynamics. Here, we profile 148 diverse wild-caught mosquitoes collected in California, detecting sequences from eukaryotes, prokaryotes, and over 70 known and novel viral species. Because we sequenced singletons, it was possible to compute the prevalence of each microbe and

recognize a high frequency of viral co-infection. By analyzing the pattern of co-occurrence of sequences across samples, we associated 'dark matter' sequences with recognizable viral polymerases, and animal pathogens with specific blood meal sources. We were also able to detect frequent genetic reassortment events in a highly prevalent quaranjavirus undergoing a recent intercontinental sweep. In the context of an emerging disease, where knowledge about vectors, pathogens, and reservoirs is lacking, the approaches described here can provide actionable information for public health surveillance and intervention decisions.

Introduction

Mosquitoes are known to carry more than 20 different eukaryotic, prokaryotic, and viral agents that are pathogenic to humans (WHO, 2017). Infections by these mosquito-borne pathogens account for over half a million human deaths per year, millions of disability-adjusted life years (GBD 2017 Disease and Injury Incidence and Prevalence Collaborators, 2018; GBD 2017 Causes of Death Collaborators, 2018; GBD 2017 DALYs and HALE Collaborators, 2018), and periodic die-offs of economically important domesticated animals (Cohnstaedt, 2018). Moreover, recent studies of global patterns of urbanization, warming, and the possibility of mosquito transport via long-range atmospheric wind patterns point to an increasing probability of a global expansion of mosquito habitat and a potential concomitant rise in mosquito-borne diseases within the next 2-3 decades (Kraemer et al., 2019; Huestis et al., 2019). While mosquito control has played a major role in eliminating transmission of these diseases in many parts of the world, costs and resources associated with basic control measures, combined with emerging pesticide resistance, pose a growing challenge in maintaining these gains (Wilson et al., 2020).

Female mosquitoes subsisting on blood meals from humans and diverse animals in their environment serve as a major source of trans-species introductions of infectious microbes. For well-studied mosquito-borne pathogens such as West Nile virus, an understanding of the transmission dynamics between animal reservoir, mosquito vector, and human hosts has been essential for public health monitoring and intervention

(Hofmeister, 2011). For less well-studied microbes with pathogenic potential, the dynamics are less clear.

We also lack a comprehensive understanding of the composition of the endogenous mosquito microbiota and how it impacts the acquisition, maintenance, and transmission of pathogenic microbes. Evidence supporting the case for a potentially important role of endogenous mosquito microbiota on mosquito-borne infectious diseases stems from decades of research on *Wolbachia*, a highly prevalent bacterial endosymbiont of insects (Werren et al., 2008). *Wolbachia* have been shown to inhibit replication of various mosquito-borne viruses that are pathogenic to humans, including dengue, chikungunya, and Zika viruses, when introduced (or “transinfected”) into susceptible mosquito species in which naturally occurring *Wolbachia* infections are rare or absent (Moreira et al., 2009).

These observations have led to the development of *Wolbachia*-based mosquito control programs for *Aedes aegypti* mosquitoes, which vector yellow fever virus, dengue virus, Zika virus, and chikungunya virus (reviewed in Ritchie et al., 2018; Flores and O’Neill, 2018). Experimental releases of *Aedes aegypti* mosquitoes transinfected with *Wolbachia* have resulted in a significant reduction in the incidence of dengue virus infections in local human populations in several countries (Hoffmann et al., 2011; O’Neill et al., 2019; Anders et al., 2018). Laboratory-based studies have identified additional endogenous mosquito microbes, such as midgut bacteria (Shane et al., 2018) and several insect-specific flaviviruses (Vasilakis and Tesh, 2015), with potential to interfere with mosquito acquisition and competence to transmit pathogenic *Plasmodium* species and

human flaviviruses, respectively. However, definitive evidence for a role of these agents in naturally occurring infections or transmission of known human pathogens has not yet been established (reviewed in Bolling et al., 2015).

Endogenous microbiota of mosquitoes could also provide a source of candidate surveillance biomarkers to follow mosquito movements, sharing of reservoirs, and potential transmission of co-occurring human pathogens carried by mosquitoes. Molecular methods for tracking that rely on mosquito genomic information are hampered by the long generation time of mosquitoes and the relatively large mosquito genome. Incorporating highly prevalent endogenous microbes of mosquitoes could simplify and refine such analyses. For example, RNA viruses have short generation times, extremely compact genomes, high mutation rates, and in some cases, undergo genetic reassortment. They are readily recovered in great numbers from mosquitoes and other insects via metagenomic sequencing (Li et al., 2015). Moreover, modest datasets of RNA virus sequences can reveal information that would be difficult to acquire through direct analysis of host populations. These features have enabled tracking of animal populations in the wild. For example, feline lentiviruses have been used to track native North American felids in the wild (Wheeler et al., 2010; Lee et al., 2014), and rabies virus have been used to track dogs in North Africa (Talbi et al., 2010). Similar approaches may enhance mosquito tracking.

Such exciting possibilities have, in part, motivated a series of recent unbiased metagenomic analyses of pools of mosquitoes collected around the world (Li et al., 2015; Shi et al., 2015, 2016; Fauver et al., 2016; Shi et al., 2017, 2018; Sadeghi et al., 2018, 2017; Xia et al., 2018; Thongsripong et al., 2018; Atoni et al., 2018; Xiao et al., 2018a,b),

(reviewed in Xia et al., 2018; Atoni et al., 2019). However, these studies have primarily focused on analysis of viruses. While these data have had a tremendous impact on our understanding of the breadth of viral diversity present in mosquito populations worldwide, they have not shed light on the potential reservoirs of these viruses or their prevalence within mosquito populations.

To link microbes with one another and with bloodmeal sources, single mosquito analyses are required. A handful of small-scale studies have demonstrated that it is possible to identify divergent viruses and evidence of other microbes in single mosquitoes by metagenomic next-generation sequencing (Chandler et al., 2015; Bigot et al., 2018; Shi et al., 2019). Here, we analysed the metatranscriptomes of 148 individual mosquitoes collected in California, USA, to characterize the composition of their microbiota, identify blood meal sources and associated pathogens, recover complete viral genomes, and infer connectivity between mosquito populations worldwide. This analysis crucially depended on the large number of individuals sequenced, which allowed us to link nucleic acid sequences detected together across many samples even when they had no homology to a reference sequence. Our findings demonstrate how large-scale single mosquito metatranscriptomics can provide insight into the mosquito's complex microbiota and contribute to the development and application of measures to control mosquito-borne pathogen transmission.

Results

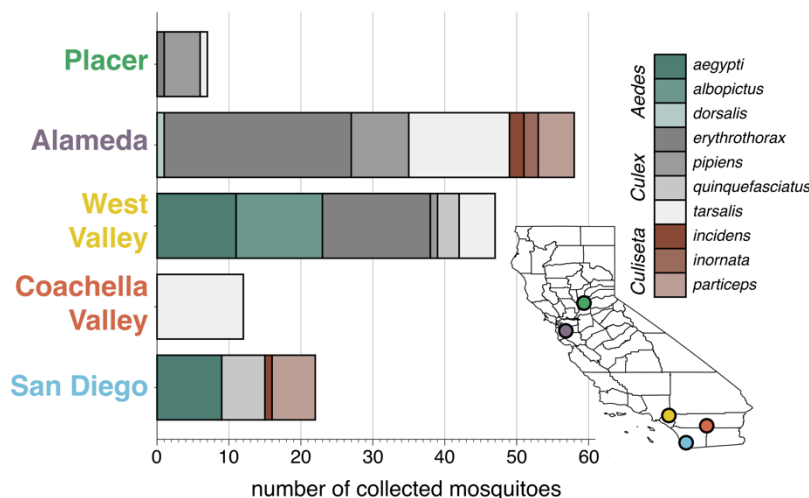
Specimen demographics

Adult mosquito species circulating in California in late fall of 2017 were collected to acquire a diverse and representative set of 148 mosquitoes for analysis. We targeted collections across a variety of habitats within 5 geographically distinct counties in Northern and Southern California (Figure 8.1). Visual mosquito species identification was performed at the time of collection (Materials and Methods), and specimens were skewed to include primarily female mosquitoes to enrich for blood-feeding members of the population responsible for transmission of animal and human diseases.

Total RNA from each mosquito was used as the input template for metatranscriptomic next generation sequencing (mNGS) library preparation to capture

both polyadenylated and non-polyadenylated host, viral, prokaryotic, and non-host eukaryotic RNA. No ribosomal RNA depletion was performed. On average, 52 million reads were obtained per mosquito (range, 11 million - 150 million reads).

Figure 8.1 Number of each genus and species of mosquitoes collected across 5 regions in California Placer, Alameda, West Valley, Coachella Valley, and San Diego. Map inset at lower right shows locations of collection sites. Colors of circles in the map correspond to colors of bar labels in plot at left.



We validated the mosquito

species identification made at the time of collection using hierarchical clustering of the complete metatranscriptomes of each mosquito through a reference-free, kmer-based approach for computing genomic distances (Harris, 2018). The computed species calls based on the metatranscriptomes agreed with the visual calls for most of the specimens (Materials and Methods, Figure 8.8). For the 8 specimens where discrepancies were detected, visual calls were reassigned to their computed calls.

Read distribution among non-host contigs

To investigate the microbiota of each mosquito, host, low quality, and low complexity sequences were filtered from each mosquito (Materials and Methods). The 0.001% - 2.80% reads remaining for each sample were then separately de novo assembled. This yielded a total of 891,000 contigs for further analysis. Contigs were then aligned in parallel to the NCBI nt and nr databases. For all prokaryotic and eukaryotic species, only a fraction of a genome was recovered, thus a lowest common ancestor (LCA) analysis of the BLAST results was used to infer the taxa the contigs represented (Materials and Methods). For viral contigs, recovery of complete or near complete genomes of RNA viruses facilitated viral taxonomic assignment directly from their BLAST results. We further manually curated contigs likely to code for viral RNA-dependent RNA polymerases (RdRps) and contigs that were physically unlinked but showed evidence of co-occurrence with viral RdRps (see Methods). Figure 8.2 shows the taxonomic distribution of reads that were assembled into contigs.

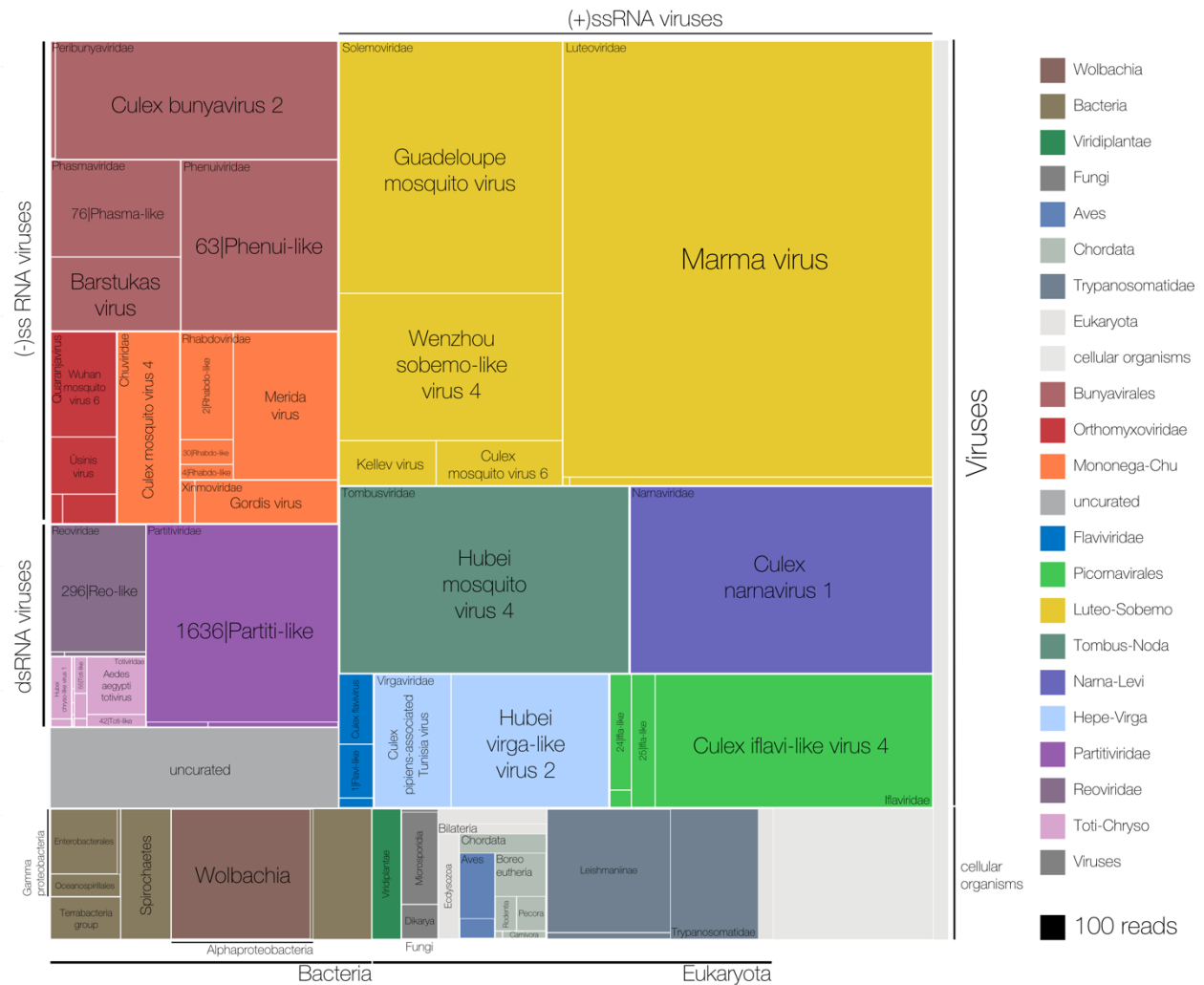


Figure 8.2 Read count distribution across non-host taxonomic groups detected among the set of 148 mosquitoes.

Treemap overview of taxa detected among the assembled contigs are plotted in squares shaded according to taxon identity (legend) and sized according to number of reads supporting each contig (legend). Due to the uncertainty in prokaryotic and eukaryotic taxonomic assignments, cellular organisms are assigned to the lowest plausible taxonomic rank. Each distinct viral lineage is displayed as its own compartment. Reddish hues indicate negative-sense RNA viruses; purple hues indicate double-stranded RNA viruses; blue, green and yellow hues indicate positive-sense RNA viruses. The "uncurated viruses" box indicates contigs that were clearly viral in origin but were too fragmented and/or not associated with a RNA-dependent RNA polymerase.

From the resulting input pool of just under 22.7 million reads, slightly more than 17.1 million (75%) assembled into contigs with more than two reads. A full 4.1 million

reads in these contigs were classified as Hexapoda, likely host sequences too diverged from reference to be recognized in the host filtering stage, and were excluded from further analysis. Metagenomic “dark matter”, i.e. reads in contigs without recognizable homology to previously published sequences, comprise 0.38 million reads. Of the remaining 13.0 million reads, 10.5 million were distinctly of viral origin. Eukaryotic and bacterial contigs comprised 0.85 and 0.68 million reads, respectively, with an additional 0.55 million reads not classifiable as either (either assigned to root or cellular organisms). We identified an additional 384 thousand viral reads from the “dark matter” (i.e. effectively all reads in this category, save for 167 individual reads) using co-occurrence analysis. We thus were able to classify, to at least kingdom level 95.7% of the non-host reads which assembled into contigs.

Viruses comprise the vast majority of the reads in our data, with positive-sense RNA viruses (7.4 million), followed by negative-sense RNA viruses (2.25 million) and double-stranded RNA viruses (0.94 million) with a small portion (0.37 million) being clearly viral in origin but incomplete or not associated with an RdRp and are listed under the “uncurated viruses” category. Reads in this category were largely from RNA viruses of Bunyavirales order and Flaviviridae family, though small numbers of reads belonged to contigs most similar to DNA viruses, such as nucleocytoplasmic large DNA viruses, members of Polydnaviridae, Alphabaculovirus, Nudiviridae, Circovirus-like sequences, phages, and others.

Within the viral taxa, we observed striking heterogeneity in read numbers, both within and across viral genome categories. This likely resulted not only from actual

differences in amounts of virus carried within or on any given mosquito, but also from differences in viral copies in organisms carried on or in the mosquitoes. For example, the recently described Marma virus (Pettersson et al., 2019), a known virus of mosquitoes, comprised nearly a quarter (23.9%) of all viral reads. In contrast, six Botourmiavirus taxa, corresponding to viruses known to infect fungi, were detected at very low levels (<200 reads each). All of the Botourmiaviruses were found in sample CMS002_053a, presumably corresponding to an infection of the ergot fungus (*Claviceps* sp.) that was detectable in the same sample.

Second in abundance by read number were bacterial taxa, with *Wolbachia* comprising most of the reads (*Wolbachia* endosymbiont of *Culex quinquefasciatus* with 0.22 million reads and *Wolbachia pipientis* with 11,000 reads). Various other bacterial taxa were detected at lower abundance (other members of Alphaproteobacteria, Gammaproteobacteria, Terrabacteria group, and Spirochaetes); *Spironema culicis* (73,000 reads) makes up 68% of Spirochaete reads.

Of the eukaryotic non-host sequences, members of Trypanosomatidae comprised 53% of reads (0.45 million). Approximately 56% (0.25 million) of these belonged to members of Leishmaniinae. The second most abundant group of eukaryotes detected in the dataset was the Bilateria (animals) taxa, with 0.20 million reads. This group was made up of the mammals (Boreoeutheria, 73,000 reads), birds (Aves, 51,000 reads), and invertebrates (Ecdysozoa, 36,000 reads). The reads derived from vertebrate taxa almost certainly belong to blood meal hosts, which we investigate in detail below.

The third largest eukaryotic category by read numbers was fungi (79,000 reads), with Microsporidia (56,000 reads) being the largest fungal group. The Microsporidia largely comprised genus-level *Amblyospora* taxa (29,000 reads). A mixture of subkingdom-level Dikarya reads (21,000 reads) were also observed, of which most are assigned to ascomycete yeasts (taxonomic designation Saccharomyceta, 17,000 reads). Plants (Viridiplantae, 62,000 reads) comprise the last category of eukaryotes with notable read prevalence in our data. Many contigs could only be assigned at the highest taxonomic levels: cellular organisms (1.87 million), bacteria (0.68 million), and eukaryotes (0.85 million reads).

Quantifying yield of new genetic information

From the 148 mosquitoes sequenced for this study, a diverse set of 70 unique viral taxa were recovered with identifiable RdRp sequences and complete or near-complete genomes. These viral taxa correspond primarily to RNA viruses and encompass a wide variety of genome types. Twenty-one of these viruses are closely related or identical to previously identified mosquito viruses: 8 correspond to viruses previously detected in California mosquitoes (Sadeghi et al., 2018; Chandler et al., 2015; Tyler et al., 2011), and 13 correspond to viruses identified from mosquito specimens collected outside of California (Göertz et al., 2019; Shi et al., 2019; Parry and Asgari, 2018; Waldron et al., 2018; Bigot et al., 2018; Shi et al., 2016). The remaining 50 viral genomes share less than 85% amino acid sequence identity to any publicly available viral sequences. Despite this divergence, family-level features such as conserved sequences at the 5' and 3' ends of bunyavirus segments allowed us to demonstrate complete genome recovery of a novel

peribunyavirus-like virus from a single mosquito at relatively low abundance (4.9% of non-host reads) (Figure 8.10). To reduce the likelihood of mis-annotating non-infectious endogenous viral elements (EVEs) present within eukaryotic genomes (Ter Horst et al., 2019; Palatini et al., 2017; François et al., 2016; Katzourakis and Gifford, 2010), the set of viruses considered here was restricted to those with complete or nearcomplete genomes. Nevertheless, it is difficult to rule out this possibility with mNGS alone and we identify two viral RdRp-like sequences, RdRp group 88, Tombus-like virus, and RdRp group 246, Reo-like virus, which fall into clades containing both replicating viruses and sequences proposed to be integrated in the genome of *Leptomonas trypanosomatid* and *Ochlerotatus* mosquito, respectively (Grybchuk et al., 2018; Cook et al., 2013).

We quantified the evolutionary novelty of these genome sequences by the total amount of branch length (in expected amino acid substitutions per site) contributed per viral lineage not derived from previously published sequences (Figure 8.3). We define this contribution of evolutionary novelty (CEN) as branch lengths of a substitution phylogeny (inferred from an untrimmed alignment) not visited after tip-to-root traversals from every background (i.e. non-study) sequence (Figure 8.11). Since branch lengths in molecular phylogenies represent discrete and independent periods of evolution, CEN is additive over studies (so long as the tree topologies agree). In contrast, statistics from pairwise comparisons, such as percent identities to the closest previously described sequence, will depend on the order in which new genomes are considered. According to this metric, our data contribute a range of evolutionary novelty across different viral families. While most

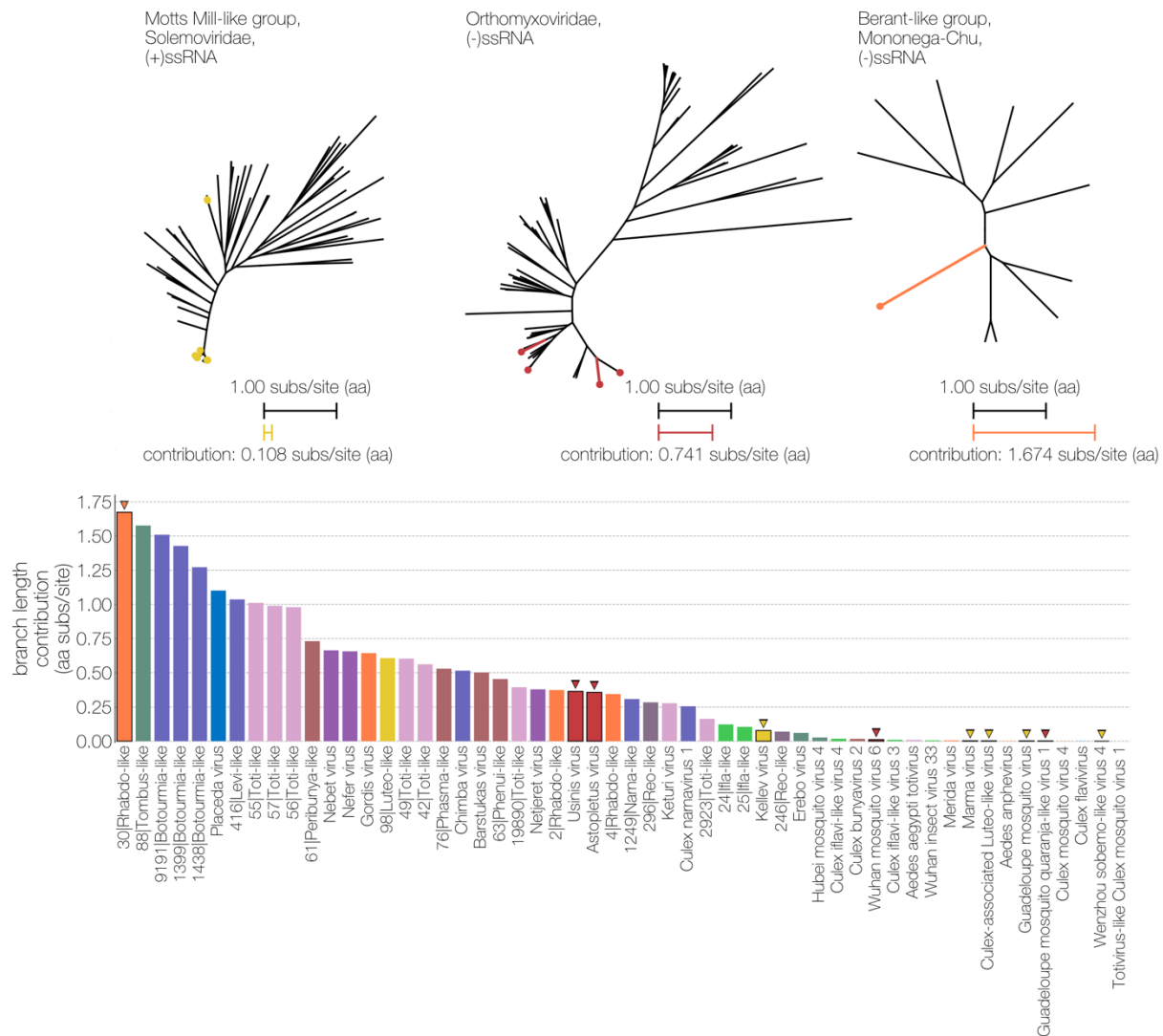


Figure 8.3 Quantifying contribution of new viral sequences.

Top panel, A collection of three unrooted trees representing varying degrees of evolutionary novelty contribution from the current study - four lineages belonging to a subgroup of Solemoviridae related to Motts Mill virus, four lineages of Quarantavirus in Orthomyxoviridae and one lineage distantly related to Berant virus in Mononegavirales. Scale bars under each tree indicate the distance corresponding to 1.0 amino acid substitutions per site (in black) as well as branch length contribution of lineages in the tree from this study (in color). Bottom panel, Bar chart depicting the evolutionary novelty contribution from each lineage described here, colored by viral family/group (same color scheme as Figure 2). Bars outlined in black with a triangle above indicate viral lineages displayed in the trees above.

viral lineages detected here contribute only modest lineages were readily identified, despite mosquitoes being a frequent target of metagenomic studies.

Microbial cargo of single mosquitos

We next delved into the single mosquito distribution of the viral, prokaryotic, and eukaryotic taxa detectable across the specimens included in this study. All confidently called contigs for microbial taxa detected within each mosquito were compiled, and the fraction of non-host reads aligning to each contig was computed to estimate the composition and proportions of microbial agents detectable within each mosquito. Figure 8.4 displays those agents detectable at a level above 1% of non-host reads plotted as bars. Confidently called contigs with less than 1% non-host read support are plotted as symbols above the x-axis coordinate for each relevant mosquito sample.

Viruses detected in single mosquitos

Examining first the 70 unique viral taxa on a single mosquito basis. Among single mosquitoes, co-infections predominated, with 120/136 mosquitoes harboring 2 or more distinct viral taxa (Figure 8.12). This insight is only possible by analyzing mosquitoes individually rather than in bulk. Furthermore, single mosquito analysis also highlighted the variability in the composition and amount of viral reads both within and across mosquito species, and their corresponding collection sites (Figure 8.4, top panel, Figure 8.14). We find that the average abundance of a virus (i.e. the average number of reads across a set of mosquitoes) is not necessarily predictive of the prevalence of that virus (i.e. the number of mosquitoes in which it occurs). For example, *Culex narnavirus 1* and *Culex pipiens*-associated *Tunisia virus* were found in similar abundances in *Culex erythrothorax*

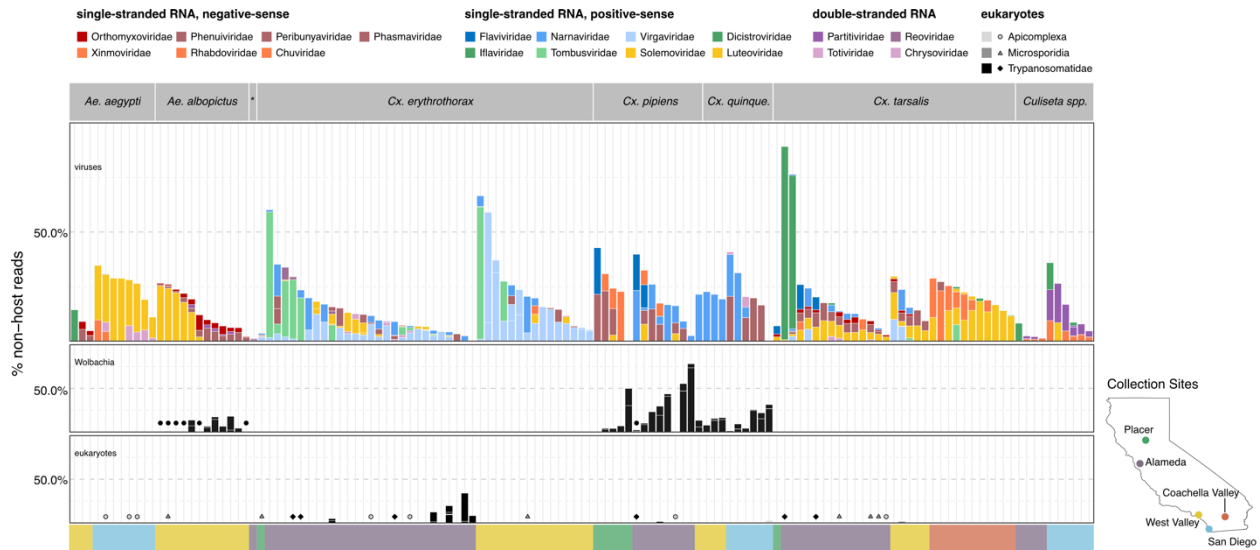


Figure 8.4 Distribution of viruses, Wolbachia, and eukaryotes in single mosquitoes. Individual mosquito proportions of non-host reads supporting assembled contigs corresponding to detectable viruses (top panel), Wolbachia (middle panel), and selected eukaryotic microbes (bottom panel) supported by $\geq 1\%$ of non-host reads (plotted bars). Parallel symbols plotted on Wolbachia and eukaryotic microbes panels indicate microbes detected at $\leq 1\%$ (middle panel, black circles = Wolbachia taxa; bottom panel, light grey circles outlined with black = Apicomplexa taxa; grey triangles = Microsporidia taxa; black diamonds = Trypanosomatidae taxa). Samples were aggregated by mosquito species (top labels), and ordered by collection site location from north to south (see x-axis color bar at the below bottom panel and inset California map at bottom right of plot), and viral abundance (descending order, left to right).

mosquitoes at the same collection site in West Valley, but the latter was 3 times more prevalent (Figure 8.13). The differentiation between abundance and prevalence would not be possible without single mosquito sequencing.

Prokaryotes detected in single mosquitoes

We restricted our single mosquito analysis of prokaryote infections to Wolbachia, which was the most abundant prokaryotic taxon detected in this study (Figure 8.2) and a known endosymbiont of mosquitoes that could impact the microbiota of its mosquito hosts (Werren et al., 2008). We detected Wolbachia in 32 mosquitoes belonging to Culex

quinquefasciatus, *Culex pipiens*, and *Aedes albopictus* species. These observations are consistent with previous observations of wild-caught mosquito species that are naturally infected with *Wolbachia* (Rasgon and Scott, 2004; Kittayapong et al., 2000). Among these 3 species, *Wolbachia* was detected in all or nearly all of the mosquitoes. However, the fraction of non-host reads assigned to *Wolbachia* per mosquito varied from <1% to 74% (Figure 8.4, middle panel, black bar plots and circle symbols). Since all or nearly all the individual mosquitoes among the 3 species were *Wolbachia*-positive, it was not possible to investigate how the presence of different levels of *Wolbachia*, or even the presence or absence of *Wolbachia*, influenced the composition or relative amount of co-occurring viral taxa detectable among these mosquitoes.

Eukaryotic microbes detected in single mosquitoes

For analysis of eukaryotes, we focused on: Trypanosomatidae, the most abundant eukaryotic taxon detected and containing established pathogens of both humans and birds, Microsporidia, a fungal taxon known to infect mosquitoes, Apicomplexa, which encompasses the causative agents of human and avian malaria, Nematoda, which contain filarial species that cause heart worm in canines and filarial diseases in humans.

Twelve mosquitoes were found to harbor Trypanosomatidae taxa. We detected sequences corresponding to monoxenous (e.g. *Crithidia* and *Leptomonas* species), dixenous (*Trypanosoma*, *Leishmania* species), as well as the more recently described *Paratrypanosoma confusum* species. Eight of the Trypanosomatidae-positive mosquitoes corresponded to *Culex erythrothorax* mosquitoes that were all collected from the same

trap site in Alameda County at different times. The four other Trypanosomatidae-positives corresponded to 2 different species of mosquitoes (2 *Culex pipiens* and 2 *Culex tarsalis*).

After the Trypanosomatidae taxa, the next most abundantly detected eukaryotic microbial taxa among the mosquitoes corresponded to the fungal taxa Microsporidia (Figure 8.2). Single mosquito analysis indicated that 6 mosquitoes (4 *Culex tarsalis* and 2 *Culex erythrothorax*) harbor Microsporidia contigs. However, only a single *Culex tarsalis* mosquito collected in West Valley was found to have > 1% non-host reads assigned to this taxon (Figure 8.4, bottom panel, gray bar plot and triangle symbols).

We also investigated the single mosquito distribution of the Apicomplexa contigs and reads we observed within our dataset. This phylum encompasses the *Plasmodium* genus, which includes several pathogenic species that cause avian and human malaria. Single mosquito analysis identified 9 mosquitoes with Apicomplexa contigs. These corresponded to 3 *Aedes aegypti* mosquitoes and 1 *Culex quinquefasciatus* mosquito, both collected in San Diego, and 3 *Culex erythrothorax* mosquitoes, 1 *Culex pipiens* mosquito, and 1 *Culex tarsalis* mosquito, collected in Alameda County. Only the *Culex quinquefasciatus* mosquito harbored Apicomplexa reads at a level above 1% non-host (Figure 8.4, bottom panel, light gray bar plot and circle symbols outlined in black). Interestingly, no viruses were observed in this mosquito.

Finally, we examined taxa falling under Nematoda, a phylum which encompasses a diverse set of more than 50 filarial parasites of humans and animals. Here, we saw evidence of Nematoda carriage in 3 *Culex* mosquitoes: 2 *Culex tarsalis* mosquitoes and 1 *Culex pipiens* mosquito (Figure 8.4, bottom panel, dark gray bar plot and diamond

symbols). Two of these mosquitoes were collected in Alameda County, and showed very low levels Nematoda ($\leq 1\%$ of non- host reads, Figure 8.4, bottom panel, diamond symbols). In the third mosquito, a *Culex tarsalis* collected in Coachella Valley, the Nematoda make up 2% of the non-host reads.

Blood meals and associated microbes

We next investigated the possibility of identifying the blood meal host directly from metatranscriptomic sequencing. We restrict this analysis to the 60 mosquitoes from Alameda County, as they were selected for visible blood engorgement. For 45 of the 60 mosquitoes, there was at least one contig with an LCA assignment to the phylum Vertebrata (range = 1-11 contigs, with 4-12,171 supporting reads). To assign a blood meal host for each of these mosquitoes, we compiled their corresponding Vertebrata contigs and selected the lowest taxonomic group consistent with those contigs. For all samples, the blood meal call fell into one of five broad categories, as shown in Figure 8.5: even-toed ungulates (Pecora), birds (Aves), carnivores (Carnivora), rodents (Rodentia), and

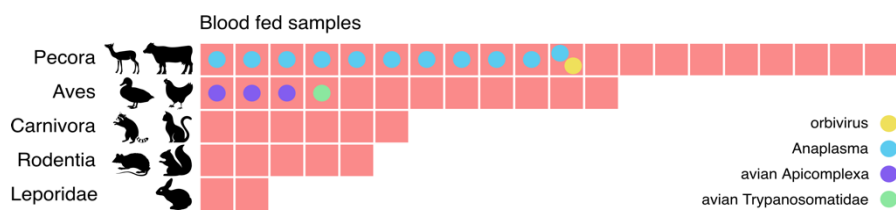


Figure 8.5 Consensus taxonomic calls of vertebrate contigs for 45 of 60 blood fed mosquitoes collected in Alameda County.

The remaining 15 samples had no vertebrate contigs. Red blocks represent individual mosquito samples; colored circles represent co-occurring contigs matching *Orbivirus*, *Anaplasma*, *Avian Apicomplexa* and *Avian Trypanosomatidae* representing possible bloodborne pathogens of the blood meal host.

rabbits (*Leporidae*). For 10 samples, this call was at the species level, including rabbit (*Oryctolagus cuniculus*), mallard duck (*Anas platyrhynchos*), and raccoon (*Procyon lotor*).

These findings are broadly consistent with the habitats where the mosquitoes were collected. For the 25 samples collected in or near the marshlands of Coyote Hills Regional Park, we compare our calls to the wildlife observations in iNaturalist, a citizen science project for mapping and sharing observations of biodiversity. iNaturalist reports observations consistent with all five categories, including various species of squirrel, rabbit, raccoon, muskrat, and mule deer. The mosquitoes with blood meals in Pecora are likely feeding on mule deer, as no other ungulate commonly resides in that marsh (iNaturalist, 2020).

We also investigated whether bloodborne pathogens of the blood meal were detectable. We performed a hypergeometric test for association between each blood meal category and each microbial taxon. The only statistically significant association ($p = 0.0005$, Bonferroni corrected) was between Pecora and *Anaplasma*, an intracellular erythroparasite transmitted by ticks. *Anaplasma* was detected in 11 of the 20 samples with Pecora, indicating a possible burden of anaplasmosis in the local deer population. Additionally, we detected evidence for three other bloodborne pathogens which, because of the small number of observations, could not pass the threshold of statistical significance. These include an orbivirus closely related to those known to infect deer (Cooper et al., 2014; Ahasan et al., 2019a,b), a *Trypanosoma* species previously found in birds (Zídková et al., 2012), and the apicomplexans *Plasmodium* and *Eimeria* from species known to infect birds (Carlson et al., 2018; Harl et al., 2019) (See Methods). The likely hosts of these pathogens were also concordant with the blood meal calls.

Recovery of previously unrecognizable viral genome segments

Although many new viruses can be identified in bulk samples, the majority of these are identified only via their conserved RdRp (Li et al., 2015; Shi et al., 2017, 2019; Pettersson et al., 2019). Recovering complete genomes for segmented viruses from bulk samples is challenging, as genes which are not highly conserved may be unrecognizable by sequence homology. Moreover, the assignment of putative segments to a single genome can be confounded if a mosquito pool contains mosquitoes with multiple infections of related segmented viruses.

By sequencing many single mosquitoes, we can exploit the fact that all segments of a segmented virus will co-occur in the samples where that virus is present and be absent in samples where the virus is absent, enabling identification of previously unidentified viral genome segments. To do so, we first grouped all contigs longer than 500 nucleotides from the study into clusters of highly homologous contigs, then grouped these clusters by cooccurrence across all the samples. This requires only the sequence information from the study, and does not use any external reference. We then scanned each cluster for sequences containing a viral RdRp domain (see Methods). For each RdRp cluster, we consider any other contig cluster whose sample group overlaps the set of samples in the viral RdRp cluster above a threshold of 80% as a putative segment of the corresponding virus (Figure 8.16). We show a cluster-by-sample heatmap for all segments co-occurring with RdRps in Figure 8.15. This resulted in 27 candidate complete genomes for segmented viruses. For the 79 of the 96 putative segments recognizable by homology to published sequences (colored in black), these groupings into genomes were accurate. This supports

the notion that the remaining 17 putative segments (colored in red), which lack homology to any known sequences at either nucleotide or amino acid level, may indeed be part of viral genomes. These putative segments represented 8% of the “dark matter” portion of the reads in the study.

Orthomyxoviruses

Detailed co-occurrence analysis for orthomyxoviruses is shown in

Figure 8.6. These are a family of multi-segmented viruses containing influenza viruses, isaviruses, thogotoviruses, and quaranjaviruses that infect a range of mammalian, avian, and fish species. Many quaranjaviruses infect arthropods, and in this study, we identified four quaranjaviruses, two of which were previously observed in mosquitoes collected outside California (Wuhan Mosquito Virus 6 (WMV6) (Li et al., 2015; Shi et al., 2017) and Guadeloupe mosquito quaranja-like virus 1 (GMQV1) (Shi et al., 2019). While influenza viruses are known to have 7 or 8 segments, the published genomes for quaranjaviruses contained 6 or fewer segments.

For the WMV6 and GMQV1 isolates detected here, we observed all of the previously identified segments (PB1, PB2, PA, and NP segments for both; and two additional segments, gp64 and “hypothetical”, for WMV6). We moreover found evidence for two distinct putative segments, which we name hypothetical 2 and hypothetical 3. We confirmed the existence of the 2 putative segments for WMV6 by assembling homologous segments from reads in the two previously published datasets describing this virus (Li et al., 2015; Shi et al., 2017). For GMQV1, we were able to find reads in the published SRA

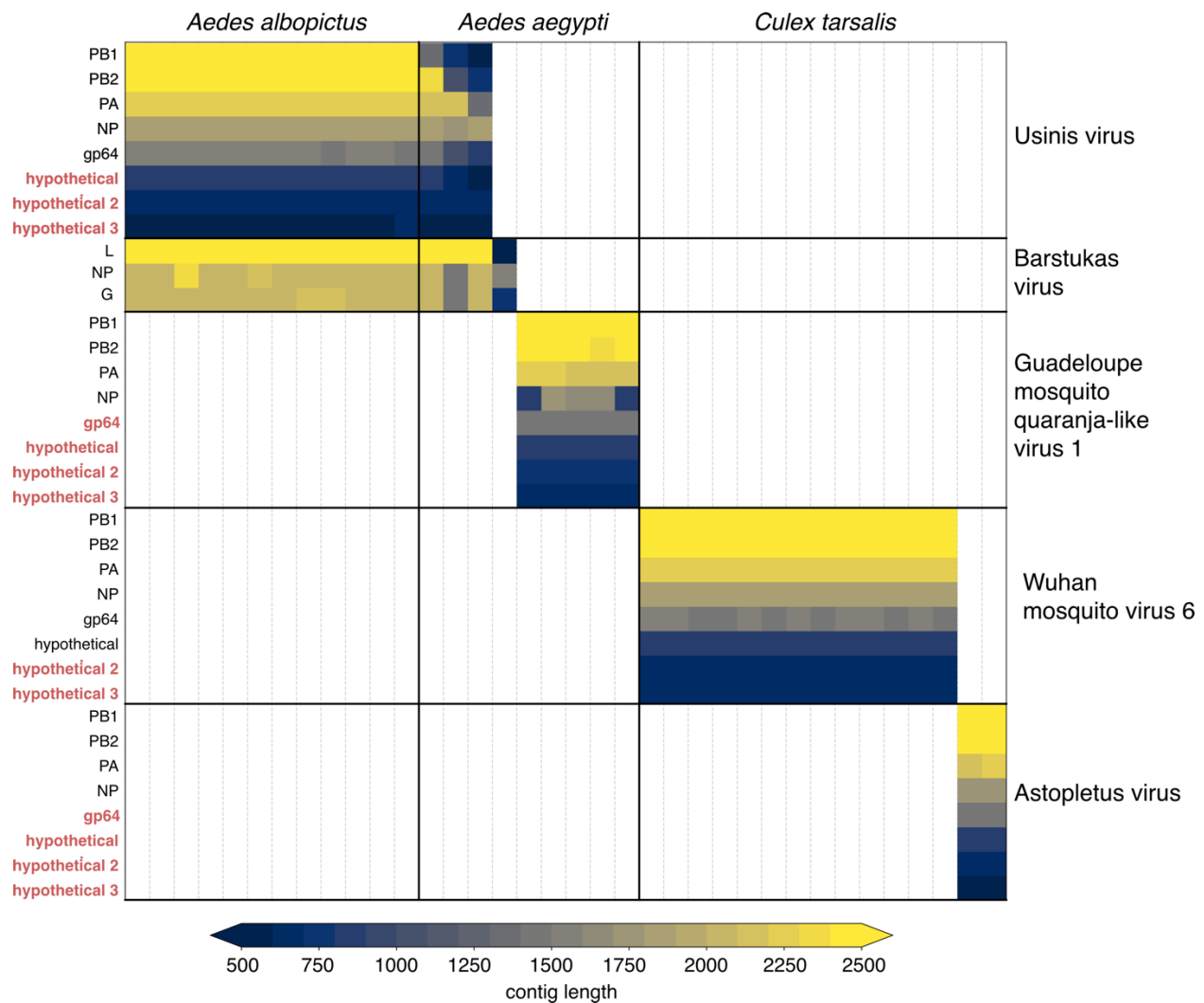


Figure 8.6 Previously unrecognized Orthomyxovirus genome segments identified among unaligned “dark matter” contigs using co-occurrence analysis.

Matrix of contigs derived from four distinct Orthomyxoviruses and one Phasma-like virus that were detected via their distinct co-occurrence pattern across mosquitoes. Rows are clusters of highly similar (>99% identity) contigs and columns are individual mosquito samples. Light grey vertical lines delineate mosquito samples, dark black vertical lines indicate boundaries between mosquito species of each sample. Dark horizontal lines delineate segments comprising viral genomes. Labels on the right indicate viruses, with genomes delineated by horizontal lines. Guadeloupe mosquito quaranja-like virus 1 and Wuhan mosquito virus 6 were previously described and Usinis, Barstukas and Astopletus were named here. At left, plain text indicates putative labels for homologous clusters; black text indicates segments identifiable via homology (BLASTx) and red text indicates contig clusters that co-occur with identifiable segments but themselves have no identifiable homology to anything in GenBank. The Phasma-like Barstukas virus exhibits a

nearly perfect overlap with Usinis virus (except for one sample in which Usinis was not found) but is identifiable as a Bunya-like virus due to having a three-segmented genome with recognizable homology across all segments to other Phasma-like viruses. Cells are colored by contig lengths (see color scale legend), highlighting their consistency which is expected of genuine segments. Deviations in detected contig lengths (e.g., Aedes aegypti samples that harbor shorter Usinis virus genome segments) reflect the presence of partial or fragmented contig assemblies in some of the samples.

entries that are similar at the amino acid level to putative protein products of the two new segments, there was not sufficient coverage to reconstruct whole segments.) Furthermore, phylogenetic trees constructed separately for each of the eight segments have similar topologies (See tanglegram, top panel of Figure 8.7), suggesting that the two new putative segments have evolved in conjunction with the previous six, bringing the total number of segments for each genome to eight.

For the two quaranjaviruses discovered in this study, Usinis virus and Astopletus virus, the co-occurrence analysis also produced 8 segments, 5 and 4 of which, respectively, were recognizable by alignment to NCBI reference sequences. The hypothetical 2 and hypothetical 3 segments we identified from these four quaranjavirus genomes are too diverged from one another to align via BLASTx, but they do share cardinal genomic features such as sequence length, ORF length, and predicted transmembrane domains Figure 8.17. The four viruses discussed here are part of a larger clade of quaranjaviruses, as pictured in Figure 8.18. It is likely that the remaining seven viruses in this clade also have eight segments.

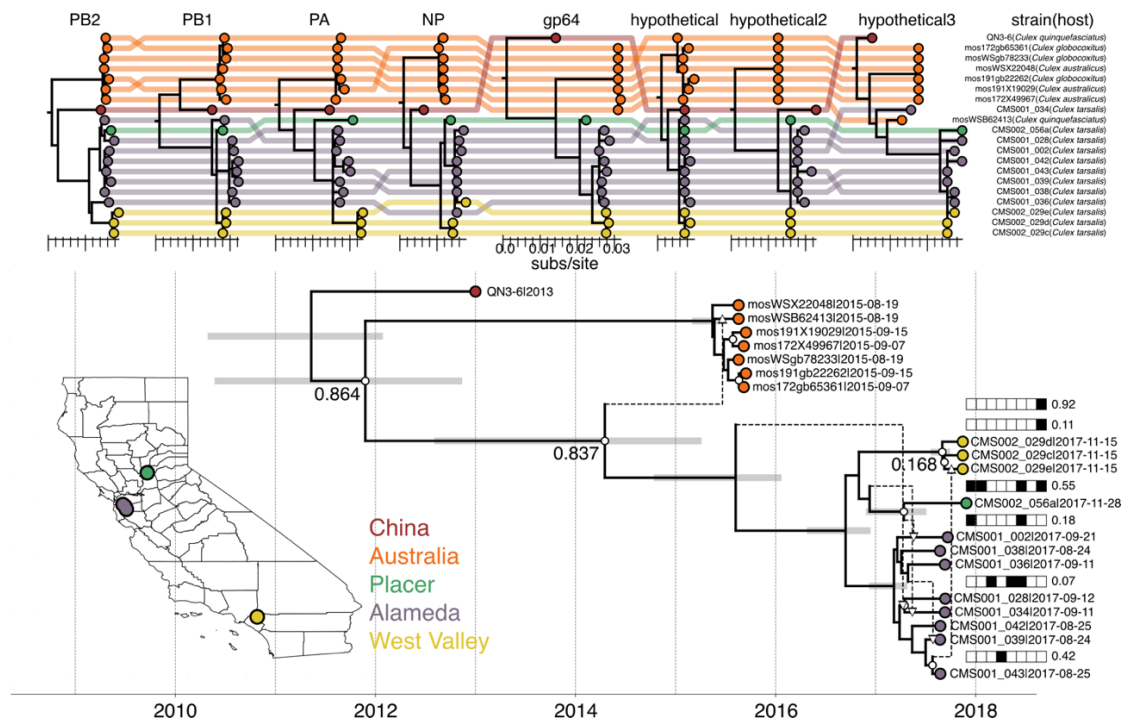


Figure 8.7 Phylogenetic analysis of Wuhan mosquito virus 6.

Top panel, Tanglegram of 20 Wuhan mosquito virus 6 genomes comprised of 8 maximum likelihood (ML) segment trees. The Chinese strain (QN3-6) was originally described from a single PB1 sequence while Australian (orange) viruses were described as having 6 segments. In this study we report the existence of two additional smaller segments (named hypothetical 2 and hypothetical 3) which we have assembled from our samples and the SRA entries of Chinese and Australian samples. Strains recovered in California as part of this study are colored by sampling location (Placer in green, Alameda County in purple, West Valley in yellow). Strain names and hosts are indicated on the far right with colored lines tracing the position of each tip in other segment trees with crossings visually indicating potential reassortments.

Bottom panel, Reassortment time network of the same data as above. Clonal evolution is displayed as solid lines with dashed lines indicating inferred reticulate evolution. Information about which segments are reassorting is represented by eight squares arranged in the same order as the maximum likelihood trees above with black squares indicating traveling segments with posterior probability of reassortment in brackets displayed towards the end of the tree at the same level along the y axis as the origin of the reassorting lineage. Reassortment contributions (i.e. “landing” of a lineage) are indicated by small white arrows and nodes with posterior probability greater than 0.05 are indicated by white circles; posterior probabilities indicated if the node is the ancestor of at least three tips. Grey rectangles mark the 95% highest posterior density (HPD) interval for node heights.

The high rate of viral co-infections (Figure 8.12) detected among the single mosquitoes we analyzed indicated a high likelihood that multiple mosquitoes could harbor more than one multi-segmented virus. A co-occurrence threshold of 0.8 was sufficient to deconvolve those segments into distinct genomes in all cases but one. There were 15 mosquito samples containing both Usinis virus, an orthomyxovirus with eight segments (three of which were unrecognizable by BLASTx), and Barstukas virus, a Phasma-like bunyavirus, with one additional sample where only Barstukas virus was found (

Figure 8.6, top two blocks). In this case, we were able to disentangle the genomes of these two viruses using additional genetic information: Barstukas virus contains all three segments expected for a bunyavirus (L, GP, and NP), all of which had BLASTx hits to other Phasma-like viruses, while the unrecognizable segments of Usinis virus shared features with the other quaranjaviruses in the study (as described above).

Culex narnavirus 1

Beyond detection of missing genome segments for known multi-segmented viruses, the co-occurrence analysis also revealed additional genome segments in “dark matter” contig clusters for viruses with genomes previously considered to be non-segmented. A striking example is an 850 nucleotide contig cluster that co-occurred with the *Culex narnavirus 1* RdRp segment in more than 40 mosquitoes collected from diverse locations across California (Figure 8.15). Like the RdRp segment, the putative new second segment shares the exceptional feature of ambigrammatic open reading frames (ORFs) (DeRisi et al., 2019; Dinan et al., 2019) (Figure 8.17). The phylogenetic tree topology for the set of 42

putative second segments is similar to the tree for the RdRp segments, suggesting co-inheritance (Figure 8.19). Moreover, we were able to recover nearly identical contigs from previously published mosquito datasets, all of which also contained the *Culex narnavirus* 1 RdRp segment. This provides strong evidence that this otherwise unrecognizable sequence is a genuine *Culex narnavirus* 1 segment, which we refer to here as the Robin segment.

Since the Narnaviruses were first described in fungi (Hillman and Cai, 2013) and recent studies have shown other eukaryotes can serve as Narnavirus hosts (Göertz et al., 2019; Charon et al., 2019; Richaud et al., 2019; Dinan et al., 2019), we investigated whether this virus co-occurred with a potential non-mosquito host. There was no significant co-occurrence with a non-mosquito eukaryotic taxon, or between the abundance of *Culex narnavirus* 1 and abundance of fungi (Figure 8.20).

As the putative new Robin segment was pulled from the “dark matter” fraction of contigs, each of the ORFs it encodes corresponds to a hypothetical protein without sequence similarity to any publicly available sequence. To investigate differences in potential functional constraints of the opposing ORFs across both Robin and the RdRp segments, we compared the amino acid alignments among the isolates for each of their ORFs. Among the 42 *Culex narnavirus* 1 RdRp segments we identified and the 3 closely related reference sequences, we detected 139 variable sites in the RdRp compared with 413 in the reverse ORF. In the Robin segment, there were 98 and 126 variable sites in the forward and reverse ORFs, respectively. Taken together, the discovery of a second segment sharing the ambigrammatic structure and the large difference in amino acid

conservation between forward and reverse ORFs, are consistent with a model where the opposition of ORFs is important for the viral lifecycle, rather than simply an efficient encoding of unrelated viral proteins.

Wuhan mosquito virus 6 phylogeography

Thorough identification of complete genomes for segmented viruses also provides a more refined and accurate understanding of both genetic history and reassortment potential. We used the set of 12 complete WMV6 genome sequences recovered from California via the co-occurrence analysis described above, plus the additional complete genomes we assembled from data provided for the previously reported Chinese (Li et al., 2015) and Australian WMV6 isolates (Shi et al., 2017), to investigate the phylogeography of WMV6. Taken together, these provided 4 years of data for our analysis (Drummond et al., 2003). We reconstructed a reassortment network (Figure 8.7, bottom panel) (Müller et al., 2019)), to partition the WMV6 evolution into clonal parts (informed by nearly 12,000 nucleotides) and reticulate parts from reassorting segments. This limited sampling of WMV6 from mosquitoes reveals limited population structure within sampling locations (Australia and California) but clear, presumably oceanic, boundaries between viruses from different studies (Figure 8.7, bottom panel). Viral diversity appears to be distance-dependent where Australian viruses (collected at most 400km apart) have lower diversity than viruses seen in California (collected at most 700km apart). In both cases, however, viral populations are exceedingly homogeneous with times to most recent common ancestor (TMRCA) of ≤ 1 year (with reassortant exceptions). This, combined with an east-to-west-oriented ladder-like phylogeny, suggests that WMV6 is potentially undergoing a

transcontinental sweep restricted to *Culex* mosquitoes where California represents the most recent landing of the virus. Additional sequence data not shown here suggest the origins of the sweep could be as far east as Europe (Pettersson et al., 2019) with Southeast Asian samples falling in the expected position between China and Australia under a hypothesis of a westerly global sweep.

Discussion

We demonstrate how metatranscriptomic sequencing of single mosquitoes, together with reference-free analyses and public databases, provides in a single assay information about circulating mosquito species, prevalence, cooccurrence, and phylodynamics of diverse known and novel viruses, prokaryotes, and eukaryotes, blood meal sources and their potential pathogens. While unbiased sequencing of individual mosquitoes is not practical or appropriate in all contexts, it offers a straightforward and rapid way to characterize a population of mosquitoes, their blood meals, and their microbial cargo. In the context of an emerging disease, where knowledge about vectors, pathogens, and reservoirs is lacking, the techniques here can provide actionable information for public health surveillance and intervention decisions.

Inferring biology from sequence in the context of an incomplete reference

The power of metatranscriptomic sequencing depends on the ability to extract biological information from nucleic acid sequences. For both bulk and single mosquito sequencing studies, the primary link between sequence and biology is provided by public reference databases, and thus the sensitivity of these approaches will depend crucially on

the quality and comprehensiveness of those references. In practice, even the largest reference databases, such as nt/nr from NCBI, represent a small portion of the tree of life, so the best match for a sequence from a sample of environmental or ecological origin is often at a low percent identity. Here, we manage that uncertainty by assigning a sequence to the lowest common ancestor of its best matches in the reference database. However, there is a fundamental limit to the precision of taxonomic identification from an incomplete reference.

An advantage of single mosquito sequencing is that it offers an orthogonal source of information: the ability to recognize nucleic acid sequences detected in many samples even when they have no homology to a reference sequence. This allowed us to associate unrecognizable sequences with viral polymerases, generating hypothetical complete genomes supported, retrospectively, public datasets. The strategy of linking contigs that co-occur across samples is utilized in analysis of human and environmental microbiomes, where it is referred to as "metagenomic binning" (Roumpeka et al., 2017; Breitwieser et al., 2019). In this manner we pulled 8% the reads in the "dark matter" fraction of our dataset into the light. The putative complete genomes were supported, retrospectively, by public datasets, and can be further validated by biological experiments or approaches such as short RNA sequencing that indicate a host antiviral response (Aguiar et al., 2015; Waldron et al., 2018).

Another advantage of single mosquito sequencing is the ability to validate visual species identifications using molecular data. Though only 3 of the 10 mosquito species in this study had complete genome references, it was possible to estimate pairwise SNP

distances between samples in a reference-free way and perform an unsupervised clustering. The clusters were largely concordant with the visual labels, and the outliers were easy to detect and correct (Figure 8.8). In a bulk pool, the microbiota from mislabeled specimens would be blended in with the correctly labeled ones, and difficult or impossible to deconvolute after the fact. This approach generalizes to any collection of metatranscriptomes containing multiple representatives of each species, including the more than 200 mosquito species without full genomes, other arthropods, bats, or any other disease vector.

Distribution of microbes within mosquito populations

Once sequences have been mapped to taxa, it is relatively straightforward to characterize the composition of the microbes within a circulating population of mosquitoes. This information can inform basic research and epidemiologic questions relevant for modeling the dynamics of infectious agents and the efficacy of interventions. A key parameter is the prevalence of a microbe, which cannot be inferred from bulk data. For the 70 viruses in this study, the prevalence ranged from detection in one mosquito (RdRp group 61, peribunya-like virus) to detection in all 36 *Culex tarsalis* samples in the study (Marma virus).

For some questions, the prevalence data supplied by single mosquito sequencing is helpful for experimental design. For example, in our dataset, *Wolbachia* was either absent or endemic in each mosquito species sampled. Thus it is not possible to detect an effect of *Wolbachia* on virome composition or abundance within any species. Single mosquito sequencing could address such questions via more extensive, targeted sampling of

mosquito populations where *Wolbachia* (or any other agent of interest) is expected to have an intermediate prevalence.

Also, the high incidence of viral co-infections has implications for the potential for viral recombination or reassortment within the mosquito population. As described for WMV6 below, such information can provide clues to transmission, virus evolution, and mosquito movements over time.

Blood meal sources and xenosurveillance

The identification of blood meal hosts is important for understanding mosquito ecology and controlling mosquito-borne diseases. Early field observations were supplemented by serology (Washino and Tempelis, 1983), and, more recently, molecular methods based on host DNA. Currently, the most popular method of blood meal identification is targeted PCR enrichment of a highly-conserved "barcode" gene, such as mitochondrial cytochrome oxidase I, followed by sequencing (RATNASINGHAM and HEBERT, 2007; Reeves et al., 2018). To monitor specific relationships between mosquito, blood meal, and pathogen, studies have combined visual identification of mosquitoes, DNA barcode identification of blood meal, and targeted PCR or serology for pathogen identification (Batovska et al., 2018; Tedrow et al., 2019; Boothe et al., 2015; Tomazatos et al., 2019). Here, we extend the spectrum of molecular methods, and show that unbiased mNGS of single mosquitoes can identify blood meal hosts, while simultaneously validating the mosquito species and providing an unbiased look at the pathogens. This allows for both reservoir identification, which seeks to identify the unknown host of a known pathogen, and xenosurveillance, which seeks to identify the unknown pathogens

of specific vertebrate populations (Grubaugh et al., 2015). For example, in this study we found a high prevalence of the tick-borne pathogen *Anaplasma* in mosquitoes that had likely ingested a blood meal from deer. In one deer-fed mosquito, we found Lobuck virus, a member of a clade of orbiviruses implicated in disease of commercially farmed deer in Missouri, Florida, and Pennsylvania (Cooper et al., 2014; Ahasan et al., 2019b,a). Our data suggest that mosquito species are a potential vector for such orbiviruses. For these analyses, it was crucial that single mosquitoes were sequenced—if the mosquitoes had been pooled, it would not have been possible to associate potential vertebrate pathogens with a specific blood meal hosts.

Tracking specific microbes and mosquitos

As the rate of metagenomic sequencing increases, new studies are more likely to find previously characterized viruses. (In fact, while this manuscript was in preparation, a study of viruses in the Caribbean island of Guadeloupe (Shi et al., 2019) described the existence of a virus, Guadeloupe Mosquito Virus 1, which we also detected here in California.) Each virus becomes more interesting the more places it is seen, as the accumulation of observations allows for the construction of more detailed phylogenies. These can reveal the evolution of the virus, its evolutionary rate, reassortment dynamics, and geographic distribution. As the gross structure of the virome is filled out, the rate of uncovering novel genetic diversity (i.e. branch length) will decrease and the fine structure of viral evolution will begin to be elucidated.

The case of Wuhan Mosquito virus 6 illustrates this trend well. In the course of five years, our understanding of a virus has gone from a single segment found in one city (Li et

al., 2015) to eight segments seen across three continents, with evidence for rapid spread, co-infection, and concomitant reassortment. As further studies provide more closely related sequences to the viruses first described in this and other reports, the portrait of mosquito-borne RNA virus evolution and migration will come into focus. In the context of potential habitat changes due to global warming (Kraemer et al., 2019), documented transcontinental transportation of mosquitos by humans (Enserink, 2008), and recent evidence of wind-borne long-distance mosquito migration (Huestis et al., 2019), the fast molecular clock of RNA virus evolution may prove a useful tool for the precise tracking of shifting mosquito populations.

A critical role for public data in public health

This study would have been impossible without rich public datasets containing sequences, species, locations, and sampling dates. These provided the backbone of information allowing us to identify the majority of our sequences. Non-sequence resources, such as the iNaturalist catalog of citizen scientist biodiversity observations, was a valuable complement, providing empirical knowledge of species distributions in the mosquito collection area that resolved the ambiguity we detected in sequence space.

Evidence-based public health interventions could benefit from a combination of unbiased, single mosquito sequencing, targeted analysis of mosquito pools, and field observations of mosquitoes and the animals they bite. As shown here, single mosquito mNGS can map an uncharted landscape that targeted sequencing can cheaply monitor, informing physical inspection and interventions. As mosquitoes and their microbiota

continue to evolve and migrate, posing new risks for human and animal populations, these complementary approaches will empower scientists and public health professionals.

Data and Code Availability

Raw sequencing data is available on the NCBI Short Read Archive at accession PRJNA605178. Code is available on Github at github.com/czbiohub/california-mosquito-study. Derived data (including all contigs) and supplementary data is available on Figshare at dx.doi.org/10.6084/m9.figshare.11832999.

Materials and Methods

Mosquito collection

Adult mosquitoes were collected using encephalitis virus survey (EVS) or gravid traps that were baited with CO₂ or hay-infused water, respectively. The collected mosquitoes were frozen using dry ice or paralyzed using triethyl amine and placed on a -15°C chill table or in a glass dish, respectively, for identification to species using a dissection microscope. Identified female mosquitoes were immediately frozen using dry ice in deep well 96-well plates and stored at -80°C or on dry ice until the nucleic acids were extracted for sequencing.

RNA Preparation

Individual mosquitoes were homogenized in bashing tubes with 200uL DNA/RNA Shield (Zymo Research Corp., Irvine, CA, USA) using a 5mm stainless steel bead and a TissueLyserII (Qiagen, Valencia, CA, USA) (2x1min, rest on ice in between). Homogenates were centrifuged at 10,000xg for 5min at 4°C, supernatants were removed and further

centrifuged at 16,000xg for 2min at 4°C after which the supernatants were completely exhausted in the nucleic acid extraction process. RNA and DNA were extracted from the mosquito supernatants using the ZR-Duet™ DNA/RNA MiniPrep kit (Zymo Research Corp., Irvine, CA, USA) with a scaled down version of the manufacturer's protocol with Dnase treatment of RNA using either the kit's DNase or the Qiagen RNase-Free DNase Set (Qiagen, Valencia, CA, USA). Water controls were performed with each extraction batch. Quantitation and quality assessment of RNA was done by the Invitrogen Qubit 3.0 Fluorometer using the Qubit RNA HS Assay Kit (ThermoFisher Scientific, Carlsbad, CA, USA) and the Agilent 2100 BioAnalyzer with the RNA 6000 Pico Kit (Agilent Technologies, Santa Clara, CA, USA).

Library Prep and Sequencing

Up to 200ng of RNA per mosquito, or 4uL aliquots of water controls extracted in parallel with mosquitoes, were used as input into the library preparation. A 25pg aliquot of External RNA Controls Consortium (ERCC) RNA SpikeIn Mix (Ambion, ThermoFisher Scientific, Carlsbad, CA, USA) was added to each sample. The NEBNext Directional RNA Library Prep Kit (Purified mRNA or rRNA Depleted RNA protocol; New England BioLabs, Beverly, MA, USA) and TruSeq Index PCR Primer barcodes (Illumina, San Diego, CA, USA) were used to prepare and index each individual library. The quality and quantity of resulting individual and pooled mNGS libraries were assessed via electrophoresis with the High Sensitivity NGS Fragment Analysis Kit on a Fragment Analyzer (Advanced Analytical Technologies, Inc.), the High-Sensitivity DNA Kit on the Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA), and via real-time quantitative polymerase chain

reaction (qPCR) with the KAPA Library Quantification Kit (Kapa Biosystems, Wilmington, MA, USA). Final library pools were spiked with a non-indexed PhiX control library (Illumina, San Diego, CA, USA). Pair-end sequencing (2 x 150bp) was performed using a Illumina Novaseq or Illumina NextSeq sequencing systems (Illumina, San Diego, CA, USA). The pipeline used to separate the sequencing output into 150-base-pair pair-end read FASTQ files by library and to load files onto an Amazon Web Service (AWS) S3 bucket is available on GitHub at <https://github.com/czbiohub/utilities>.

Mosquito Species Validation

To validate and correct the visual assignment of mosquito species, we estimated SNP distances between each pair of mosquito transcriptomes by applying SKA (Split Kmer Analysis) to the raw fastq files for each sample. The hierarchical clustering of samples based on the resulting distances was largely consistent with the visual assignments, with each cluster containing a supermajority of a single species. To correct likely errors in the visual assignment, samples were reassigned to the majority species in their cluster, resulting in 8 changes out of 148 samples.

Host and Quality Filtering

Raw sequencing reads were host- and quality-filtered and assembled using the IDseq (v3.2) platform <https://idseq.net>, a cloud-based, open-source bioinformatics platform designed for detection of microbes from metagenomic data.

Host Reference

We compiled a custom mosquito host reference database made up of:

- All available mosquito genome assemblies under NCBI taxid 7157 (Culicidae; n=41 records corresponding to 28 unique mosquito species, including 1 Culex, 2 Aedes, and 25 Anopheles records) from NCBI Genome Assemblies (accession date: 12/7/2018).
- All mosquito mitochondrial genome records under NCBI taxid 7157 available in NCBI Genomes (accession date: 12/7/2018; n=65 records).
- A Drosophila melanogaster genome (GenBank GCF_000001215.4; accession date: 12/7/2018).

Mosquito Genome Assembly and mitochondrial genome accession numbers and descriptions are detailed in Supplemental Data file mosquito_genome_refs.txt.

Read Filtering

To select reads for assembly, we performed a series of filtering steps using the IDSeq platform:

- **Filter Host 1** Remove reads that align to the host reference using the Spliced Transcripts Alignment to a Reference (STAR) algorithm (Dobin et al., 2013).
- **Trim Adapters** Trim Illumina adapters using trimmomatic (Bolger et al., 2014).
- **Quality Filter** Remove low-quality reads using PriceSeqFilter (Ruby et al., 2013).
- **Remove Duplicate Reads** Remove duplicate reads using CD-HIT-DUP (Li and Godzik, 2006; Fu et al., 2012).
- **Low-Complexity Filter** Remove low-complexity reads using the LZW-compression filter (Ziv and Lempel, 1978).
- **Filter Host 2** Remove further reads that align to the host reference using Bowtie2, with flag very-sensitive-local (Langmead and Salzberg, 2012).

The remaining reads are referred to as "putative non-host reads." A detailed description of all parameters is available in the IDseq documentation.

<https://github.com/chanzuckerberg/idseq-dag/wiki/IDseq-Pipeline-Stage-%231:-Host-Filtering-and-QC>

Assembly

The putative non-host reads for each sample were assembled into contigs using SPADES with default settings. The reads used for assembly were mapped back to the

contigs using Bowtie2 (flag very-sensitive), and contigs with more than 2 reads were retained for further analysis.

Taxonomic Assignment

We aligned each contig to the nt and nr databases using BLASTn (discontinuous megablast) and PLAST (a faster implementation of the BLASTx algorithm), respectively (Altschul et al., 1990; Van Nguyen and Lavenier, 2009). (The databases were downloaded from NCBI on Mar 27, 2019.) Default parameters were used, except the E-value cutoff was set to 1e-2. For each contig, the results from the database with a better top hit (as judged by bitscore) are used for further analysis.

For contigs with BLAST hits to more than one species, we report the lowest common ancestor (LCA) of all hits whose number of matching aligned bases alignment length*percent identity is no less than the number of aligned bases for the best BLAST hit minus the number of mismatches in the best hit. (In the case that the same segment of the query is aligned for all hits, this condition guarantees that the excluded hits are further from the best hit than the query is.)

For 172,244 contigs there was a strong BLAST hit was to a species in Hexapoda, the subphylum of arthropods containing mosquitoes. This is likely a consequence of the limited number and quality of genomes used in host filtering, and all contigs with an alignment to Hexapoda of at least 80% of the query length or whose top hit (by e-value) was to Hexapoda were discarded from further analysis.

Contigs with no BLAST hits are referred to as "dark contigs".

For RNA viruses, where complete or near-complete genomes were recovered, a more sensitive analysis was performed.

Viral Polymerase and Segment Assignment

Alignments of viral RNA-dependent RNA polymerases used to detect domains were downloaded from Pfam. These were RdRP_1 (PF00680, Picornavirales-like and Nidovirales-like), RdRP_2 (PF00978, Tymovirales-like and Hepe-Virgali-like), RdRP_3 (PF00998, Tombusviridae-like and Nodaviridae-like), RdRP_4 (PF02123, Toti-, Luteo-, and Sobemoviridae-like), RdRP_5 (PF07925, Reoviridae-like), Birna_RdRp (PF04197, Birnaviridae-like), Flavi_NS5 (PF00972, Flaviviridae-like), Mitovir_RNA_pol (PF05919, Narnaviridae-like), Bunya_RdRp (PF04196, Bunyavirales-like), Arena_RNA_pol (PF06317, Arenaviridae-like), Mononeg_RNA_pol (PF00946, Mononega- and Chuviridae-like), Flu_PB1 (PF00602, Orthomyxoviridae-like). Hidden Markov model (HMM) profiles were generated from these with HMMER (v3.1b2) (Finn et al., 2011) and tested against a set of diverged viruses, including ones thought to represent new families (Obbard et al., 2019). Based on these results only the RdRP_5 HMM was unable to detect diverged Reovirus RdRp, such as Chiqui virus (Contreras-Gutiérrez et al., 2017). An additional alternative Reovirus HMM (HMMbuild command) was generated by using BLASTp hits to Chiqui virus, largely to genera Cypovirus and Oryzavirus, aligned with MAFFT (E-INS-i, BLOSUM30) (Kato et al., 2005).

All contigs of length >500 base pairs were grouped into clusters using a threshold of $\geq 99\%$ identity (CD-HIT-EST (Li and Godzik, 2006; Fu et al., 2012)). Representative contigs from each cluster were scanned for open reading frames (standard genetic code)

coding for proteins at least 200 amino acids long, in all six frames with a Python script using Biopython (Cock et al., 2009). These proteins were scanned using HMM profiles built earlier and potential RdRp-bearing contigs were marked for follow up. We chose to classify our contigs by focusing on RdRp under the assumption that bona fide exogenous viruses should at the very least carry an RdRp (per Li et al. (2015)) and be mostly coding-complete. Contigs that were not associated with an RdRp or coding-complete included Cell fusing agent virus (Flaviviridae, heavily fragmented) and Phasma-like nucleoprotein sequences (potential piRNAs) in a few samples.

Co-Occurrence

For each cluster whose representative contig contained a potential RdRp, we identified as a putative viral segment each other CD-HIT cluster whose set of samples overlapped the set of samples in the RdRp cluster at a threshold of 80%. (That is, a putative segment should be present in at least 80% of the samples that RdRp is present in, and RdRp should be present in at least 80% of the samples that the putative segment is present in).

In cases where a singleton segmented (bunya-, orthomyxo-, reo-, chryso-like, etc) virus was detected in a sample we relied on the presence of BLASTx hits of other segments to related viruses (e.g. diverged orthobunyavirus). We thus linked large numbers of viral or likely viral contigs to RNA-dependent RNA polymerases representing putative genomes for these lineages.

Final Classification

There were 1269 contigs identified as viral either by RdRp detection or co-occurrence, and the resulting specieslevel calls are used for further analysis in lieu of the LCA computed via BLAST alignments. This included 338 “dark contigs” which had no BLAST hits, 748 with LCA in Viruses; the LCAs for the remainder were Bacteria (9), and Eukaryota (4), and Ambiguous (170), a category including (including root, cellular organisms, and synthetic constructs). Reads are assigned the taxonomic group of the contig they map to.

Water Controls and Contamination

There are many potential sources of contaminating nucleic acid, including lab surfaces, human experimenters, and reagent kits. We attempt to quantify and correct for this contamination using 8 water controls. We model contamination as a random process, where the mass of a contaminant taxon t in any sample (water or Mosquito) is a random variable x_t . We convert from units of reads to units of mass using the number of ERCC reads for each sample (as a fixed volume of ERCC spike-in solution was added to each sample well). We estimate the mean of x_t using the water controls. We say that a taxon observed in a sample is a possible contaminant if the estimated mass of that taxon in that sample is less than 100 times the average estimated mass of that taxon in the water samples. Since the probability that a non-negative random variable is greater than 100 times its mean is at most 1% (Markov’s inequality), this gives a false discovery rate of 1%. For each possible contaminant taxon in a sample, all contigs (and reads) assigned to that taxon in that sample were excluded from further analysis. A total of 46,603 reads were

removed as possible contamination using this scheme. (Human and mouse were identified as the most abundant contaminant species.)

For every sample, "classified non-host reads" refer to those reads mapping to contigs that pass the above filtering, Hexapoda exclusion, and decontamination steps. "Non-host reads" refers to the classified non-host reads plus the reads passing host filtering which failed to assemble into contigs or assembled into a contig with only two reads.

Treemap

Treemaps are a way of visualising hierarchical information as nested rectangles whose area represents numerical values. To visualise the distribution of reads amongst taxonomic ranks, we first split the data into two categories: viral and cellular. For cellular taxonomic ranks (Bacteria, Eukaryotes, Archaea and their descendants) we assigned all reads of a contig to the taxonomic compartment the contig was assigned (see above, "Taxonomic Assignment"). For viral taxa we relied on the curated set of viral contigs coding for RdRp and their putative segments, where a putative taxonomic rank (usually family level) had been assigned. All the reads belonging to contigs that comprised putative genomes were assigned to their own compartment in the treemap, under the curated rank. Additional compartments were introduced to either reflect aspects of the outdated and potentially non-monophyletic taxonomy which is nevertheless informative (e.g. positive- or double-strandedness of RNA viruses) or represent previously reported groups without an official taxonomic ID on public databases (e.g. Narna-Levi, Toti-Chryso, Hepe-Virga, etc).

To prototype the cellular part of the treemap, all taxonomic IDs encountered along the path from the assigned taxonomic ID up to root (i.e. the taxonomic ID's lineage) were added to the treemap. Based on concentrations of reads in particular parts of the resulting taxonomic treemap, prior beliefs about the specificity of BLAST hits, and information utility, this was narrowed down to the following taxonomic ranks: cellular organisms, Bacteria, Wolbachia, Gammaproteobacteria, Alphaproteobacteria, Spirochaetes, Enterobacterales, Oceanospirillales, Terrabacteria group, Eukaryota, Opisthokonta, Fungi, Dikarya, Bilateria, Boreoeutheria, Pecora, Carnivora, Rodentia, Leporidae, Aves, Passeriformes, Trypanosomatidae, Leishmaniinae, Trypanosoma, Ecdysozoa, Nematoda, Microsporidia, Apicomplexa, Viridiplantae.

Microbiota distribution in single mosquitos

In Figure 8.4, the denominators are non-host reads. The numerators are numbers of reads from contigs with confident assignments. For viruses, these contigs came from viral curation or co-occurrence. For Wolbachia and eukaryotes, these contigs had LCA assignment within the Wolbachieae tribe (taxid: 952) and Eukaryota superkingdom (taxid: 2759), respectively, and had a BLAST alignment where the percentage of aligned bases was at least 90%. Groups within viruses, Wolbachia, and eukaryotes were excluded for a given sample if the cumulative proportion of nonhost reads was less than 1%. Samples were excluded if the total proportions of non-host reads belonging to viruses, Wolbachia, or eukaryotes were all below 1%.

Branch Length Contributions

Representative contigs with an encoded RdRp from each virus lineage detected were used to perform a BLASTx search under default parameters. The top 100 hits were downloaded, combined based on viral family, and duplicate hits removed. Through an iterative procedure of amino acid alignment using MAFFT (E-INS-i, BLOSUM30) and initially neighbour-joining (Saitou and Nei, 1987) and in later stages maximum likelihood phylogeny reconstructions using PhyML (Guindon and Gascuel, 2003), alignments were reduced and split such that they contained study contigs and some number of BLASTx hits. The goal was to generate alignments with catalytic motif of RdRp (motif C: GDD, GDN, ADN or SDD) perfectly conserved and total number of identical columns in the alignment to be at least 1.0%. This is in contrast to the more commonly used approach where poorly aligning regions of the alignment are irreversibly removed, precluding future re-alignment of sequence data if taxa. By focusing our attention on the closest relatives of our contigs we largely avoided having to deal with acquisition/migration of protein domains that occur over longer periods of evolution. For broader context we relied on previously published large-scale phylogenies to indicate more distant relationships.

Given the trees, the branch length contribution was quantified using "contribution of evolutionary novelty" (CEN), defined to be the total branch length the phylogeny not visited after tip-to-root traversals from every background sequence (as illustrated in Figure 8.11). This is additive in the sense that if one were to consider two studies, study 1 and study 2, added to a fixed set of background sequences, and the phylogeny fit to (background + study 1) were a subset of the phylogeny fit to (background + study 1 +

study 2), then the CEN of study 1 (against the fixed background) plus the CEN of study 2 (against the fixed background + study 1) would be the same as the CEN of the composite study consisting of sequences from both (against the fixed background).

Blood meal Calling

For each of the 60 bloodfed mosquito samples from Alameda County, we selected each contig with LCA in the subphylum Vertebrata, excluding those contained in the order Primates (because of the possibility of contamination with human DNA). For each sample, we identified the lowest rank taxonomic group compatible with the LCAs of the selected contigs. (A taxonomic group is compatible with a set of taxonomic groups if it is an ancestor or descendent of each group in the set.) For 44 of the 45 samples containing vertebrate contigs, this rank is at class or below; for 12 samples, it is at the species level. Each taxonomic assignment falls into one of the following categories: Pecora, Aves, Carnivora, Rodentia, Leporidae. In Figure 8.5, each sample with a blood meal detected is displayed according to which of those categories it belongs to. The remaining sample, CMS001_022_Ra_S6, contained three contigs mapping to members of Pecora and a single contig with LCA Euarchontoglires, a superorder of mammals including primates and rodents; we annotate this sample as containing Pecora.

Notably, 19 samples contain at least one contig with LCA in genus *Odocoileus* and another contig with LCA genus *Bos*. While the lowest rank compatible taxonomic group is the infraorder Pecora, it is likely that a single species endemic in the sampled area is responsible for all of these sequences. Given the observational data in the region

(described in the main text), that species is likely a member of *Odocoileus* whose genome diverges somewhat from the reference.

Phylogenetic analyses

We chose a single Wuhan mosquito virus 6 genome from our study (CMS001_038_Ra_S22) as a reference to assemble by alignment the rest of the genome of strain QN3-6 (from SRA entry SRX833542 as only PB1 was available for this strain) and the two small segments discovered here for Australian segments (from SRA entries SRX2901194, SRX2901185, SRX2901192, SRX2901195, SRX2901187, SRX2901189, and SRX2901190) using magicblast (Boratyn et al., 2018). Due to much higher coverage in Australian samples, magicblast detected potential RNA splice sites for the smallest segment (hypothetical 3) which would extend the relatively short open reading frame to encompass most of the segment. Sequences of each segment were aligned with MAFFT (Auto setting) and trimmed to coding regions. For hypothetical 3 segment we inserted Ns into the sequence near the RNA splice site to bring the rest of the segment sequence into frame.

PhyML (Guindon and Gascuel, 2003) was used to generate maximum likelihood phylogenies under an HKY+ Γ 4 (Hasegawa et al., 1985; Yang, 1994) model. Each tree was rooted via a least-squares regression of tip dates against divergence from root in TreeTime (Sagulenko et al., 2018). Branches with length 0.0 in each tree (arbitrarily resolved polytomies) were collapsed, and trees untangled and visualised using baltic (<https://github.com/evogytis/baltic>).

A reassortment network (Müller et al., 2019) model implemented in BEAST2 (Bouckaert et al., 2019) (v2.6) was used to infer a tree-like network describing the reticulate evolution of Wuhan mosquito virus 6 sequences. Sequences were analysed using a strict molecular clock with tip-calibration, an HKY+ Γ 4 (Hasegawa et al., 1985; Yang, 1994) model of sequence evolution, under a constant population size prior. Four independent MCMC analyses were set up to run 200 million steps, sampling every 20,000 states and combined after the removal of 20 million states as burnin. Sampled networks were summarised by using network-specific software provided with BEAST2. The network was visualised using baltic (<https://github.com/evogytis/baltic>).

Thanks to the presence of closely related viral sequences that were collected a number of years ago (Drummond et al., 2003) we were able to carry out tip-calibrated temporal phylogenetic analyses for two additional viruses - Hubei virga-like virus 2 (Shi et al., 2016) and Culex bunyavirus 2 (another lineage also submitted to GenBank as "Bunyaviridae environmental sample"). For Hubei virga-like virus 2 there were two sequences of RdRp segment (KX883772 and KX883780) collected in 2013 that were $\geq 93\%$ identical at nucleotide level to 35 sample contigs from our study. Hubei virga-like virus 2 sequences were aligned with MAFFT (Kato et al., 2005) and analysed with BEAST v1.10.4 (Suchard et al., 2018) under an HKY+ Γ 4 (Hasegawa et al., 1985; Yang, 1994) substitution model, coalescent constant population size tree prior, no partitioning into codon positions and a strict molecular clock model. The clock model was tip-calibrated (Rambaut, 2000) with an uninformative reference prior (Ferreira and Suchard, 2008) placed on the rate of the molecular clock. MCMC was run for 50 million steps, sampling

every 5000 steps. MCMC convergence was confirmed in Tracer v1.7 (Rambaut et al., 2018) and posterior trees summarised with TreeAnnotator distributed with BEAST.

For *Culex bunyavirus 2/Bunyaviridae* environmental sample there were three sequences of L segment that codes for RdRp (accessions MH188052, KP642114, and KP642115 (Chandler et al., 2015; Sadeghi et al., 2018)) collected in 2016 and 2013. The two sequences collected in 2013 were collected in and around San Francisco (KP642114, and KP642115) are up to 98.7% identical to contigs in our study whereas the 2016 sequence (MH188052) is 99.2% identical. The three previously published sequences were aligned to 32 L segment contigs from our study with MAFFT (Katoh et al., 2005) and analysed with BEAST v1.10.4 (Suchard et al., 2018). Sites were codon partitioned (1+2 and 3) and their evolution modeled with an HKY+ Γ 4 (Hasegawa et al., 1985; Yang, 1994) substitution model with a tip-calibrated (Rambaut, 2000) strict molecular clock model under a coalescent constant population size tree prior. The analysis did not converge when using an uninformative prior but did when being constrained with a uniform prior with bounds (0.000001 and 0.01 substitutions per site per year). MCMC was run for 50 million steps, sampling every 5000 steps, convergence being confirmed with Tracer v1.7 (Rambaut et al., 2018) and posterior trees summarised with TreeAnnotator distributed as part of the BEAST package.

To generate the *Culex narnavirus 1* tanglegrams, 42 sequences of RdRp and 42 co-occurring Robin segment sequences from our samples and three previously published RdRp sequences (MK628543, KP642119, KP642120) as well as their three corresponding Robin segments assembled from SRA entries (SRR8668667, SRR1706006, SRR1705824,

respectively) were aligned with MAFFT (Kato et al., 2005) and trimmed to just the most conserved open reading frame (as opposed to its complement on the reverse strand). Maximum likelihood phylogenies for both RdRp and Robin segments were generated with PhyML (Guindon and Gascuel, 2003) with 100 bootstrap replicates under an HKY+ Γ 4 (Hasegawa et al., 1985; Yang, 1994) substitution model. The resulting phylogenies were mid-point rooted, untangled and visualised using baltic (<https://github.com/evogytis/baltic>).

Acknowledgements

We thank our collaborating partners in the California Mosquito and Vector Control Agency Districts of Alameda, Placer, San Diego, West Valley, and Coachella Valley, who provided all the mosquito specimens and corresponding metadata that made this study possible. We thank Maira Phelps for liaison work with collaborators and in-house specimen management. We thank Rene Sit, Michelle Tan, and Norma Neff of the Chan Zuckerberg Biohub Genomics Platform for supporting all aspects of mNGS sequencing for this study. We thank the IDseq team at the Chan Zuckerberg Initiative for useful discussions and facilitation of analysis over the course of this study. We thank Jack Kamm, Darren J Obbard, and Cristina Tato for useful discussions during the development of this project. We would also like to acknowledge Natalie Whitis and Annie Lo for their contribution to the early phases of the specimen extraction and sequencing library preparation for this project. We thank Cristina Tato, Peter Kim, David Yllanes, and Joe DeRisi for reviewing the manuscript.

Supplemental Figures

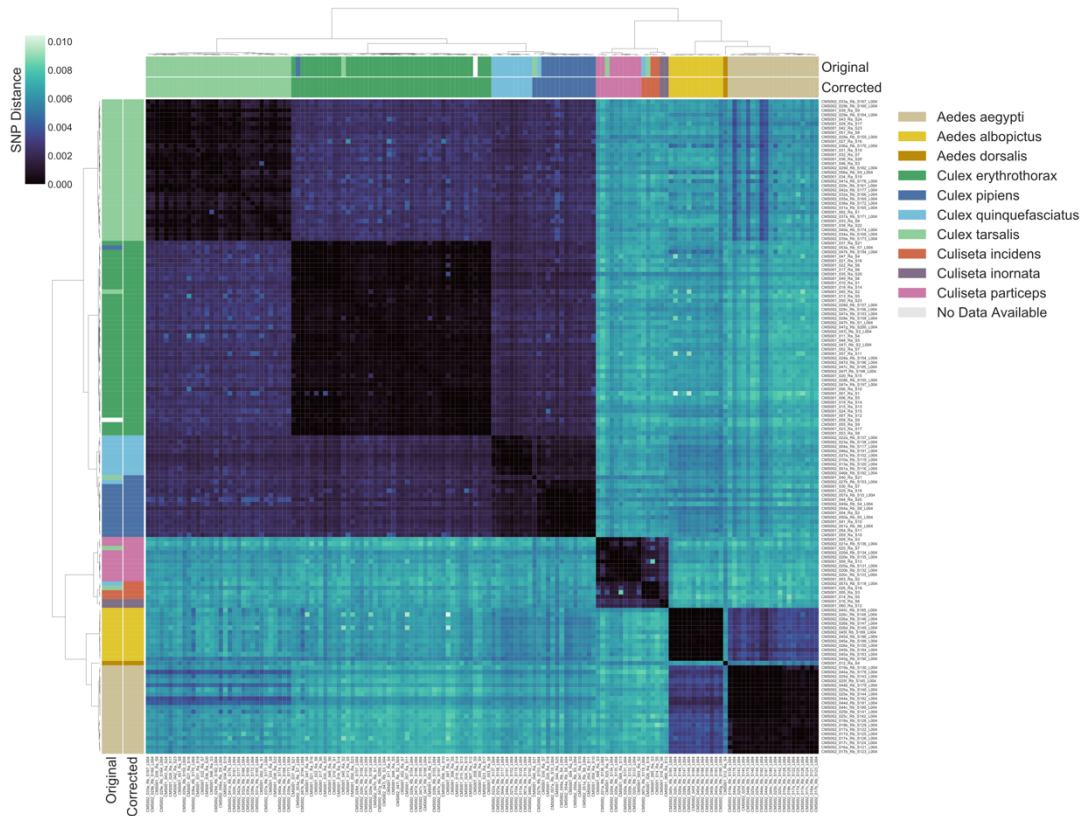


Figure 8.8 Unbiased identification of mosquito species.

Hierarchical clustering of pairwise single-nucleotide polymorphism (SNP) distances between samples estimated using SKA (Harris, 2018), ranging from 0 (no SNPs detected, black) to 0.01 (1% of comparable sites had SNPs, blue). Outside color bar indicates visual species label for each sample (Visual), inside color bar indicates consensus transcriptome species label for each sample's cluster (Computed). Computed mosquito species calls were used when Visual and Computed calls were discordant.

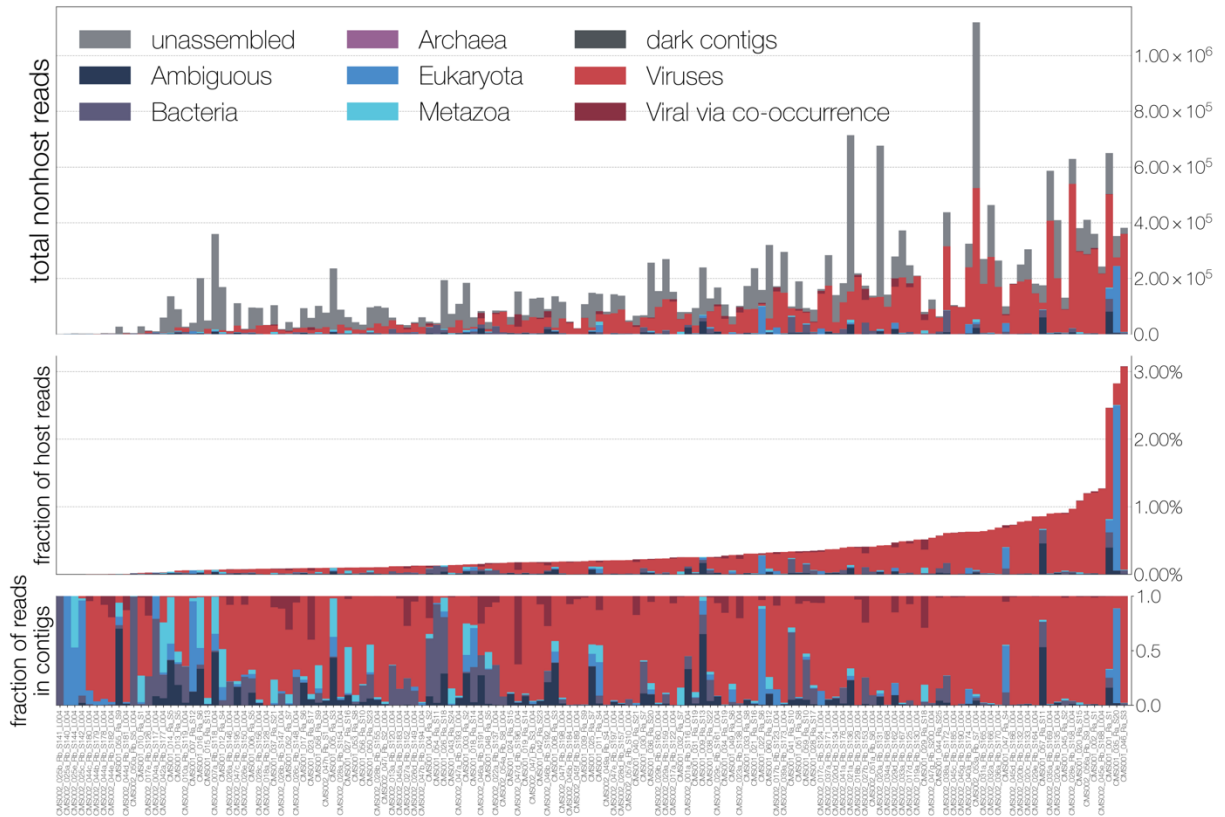


Figure 8.9 Breakdown of reads for each sample into broad categories.

A breakdown of reads for each sample, into unassembled (those which did not assemble into contigs), dark (those which assembled but were not identifiable taxonomically), Ambiguous (those with LCA at root or cellular organisms), Bacteria, Archaea, Eukaryota, Metazoa, Viruses (via alignment to reference), and Viral via co-occurrence analysis (See Figure 6). Top panel is a stacked bar decomposing the total reads per sample. Middle panel shows the breakdown scaled as a fraction of all reads in a sample. Bottom panel shows the fractional breakdown among assembled reads.

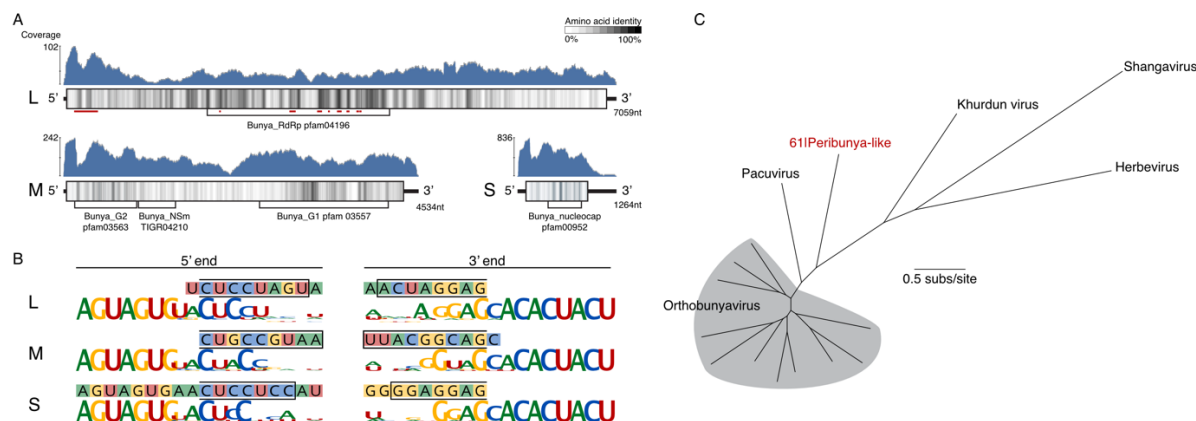


Figure 8.10 Analysis of peribunya-like virus showing completeness.

(A) Read coverage in blue across the length of each genome segment. Diagrams of genome segments display the longest ORF and indicate amino acid identity with a sliding window of 15 amino acids. Regions of homology to known domains are indicated by Pfam labels, and on the L segment, the locations of canonical RdRp motifs are indicated by red underlining. (B) Ends of the assembled contigs are shown (sequence in filled boxes). Regions complementary between 5' and 3' end for each segment are indicated by a black outline. These contig ends were aligned to the conserved end sequences of orthobunya and nearby peribunya viruses (logo diagram shows conservation after alignment of reference sequences). (C) Maximum likelihood phylogenetic tree with representative orthobunyaviruses and nearby peribunyaviruses.

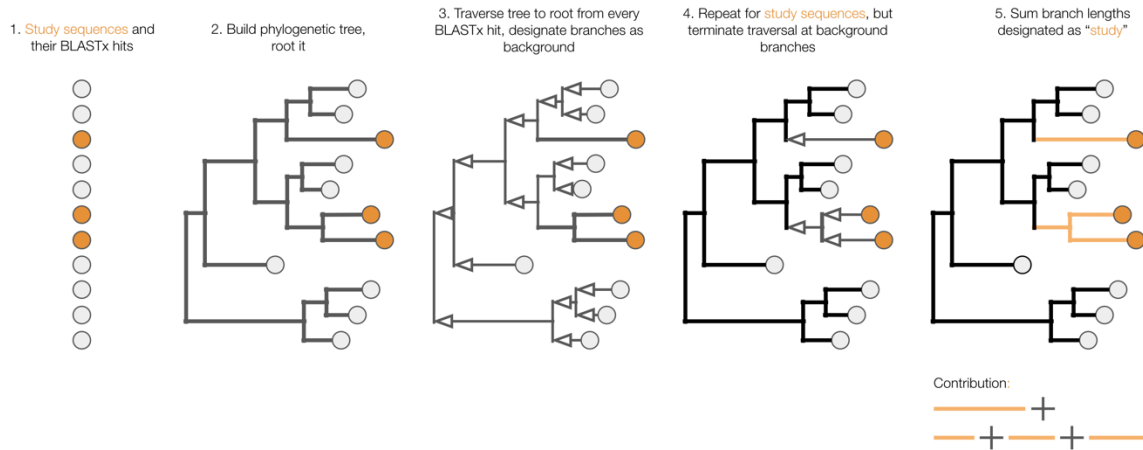


Figure 8.11 Schematic describing method for branch length contribution calculation.

Final calculation of branch length for a given group of viruses quantifies the contribution of evolutionary novelty, in units of substitutions/site.

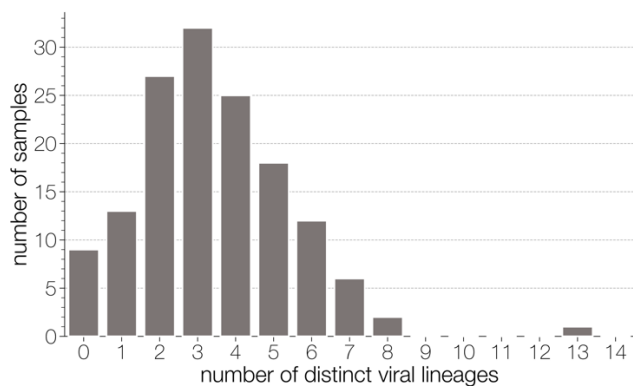


Figure 8.12 Distribution of mosquitoes within the study in which no, one, or multiple viral lineages were detectable.

The sample with 13 distinct viral lineages is sample CMS002_053a which contains six *Botourmia*-like viruses thought to primarily infect fungi and is also the sample where there is evidence for the presence of an ergot fungus.

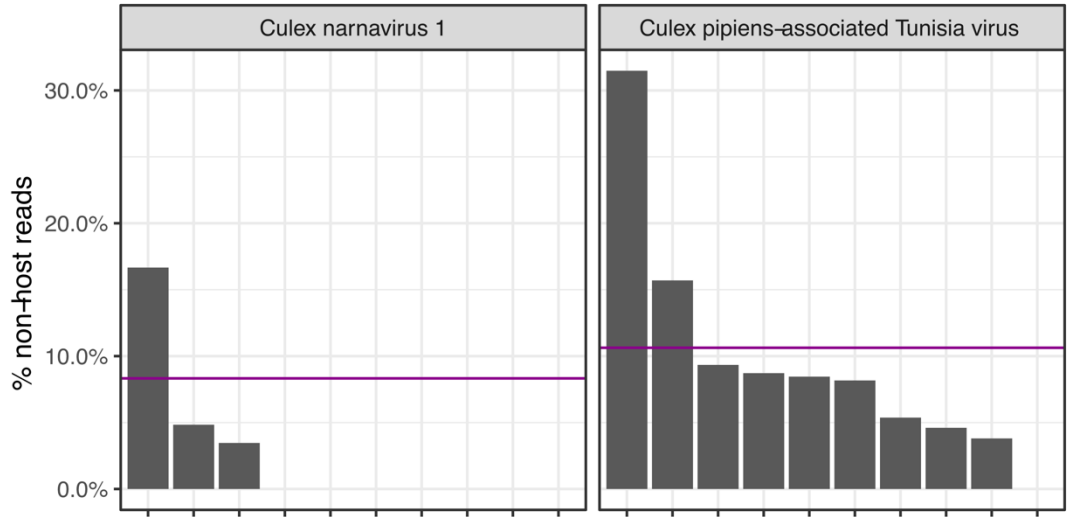


Figure 8.13 An example of viruses with similar bulk abundance but different prevalence.

Viral abundance is often calculated based on bulk mosquito sequencing, which does not provide information about the prevalence or heterogeneity in abundance of a virus across the mosquito population. Both *Culex narnavirus 1* and *Culex pipiens-associated Tunisia virus* were found in *Culex erythrothorax* mosquitoes at the same collection site in West Valley. A total of 10 *C. erythrothorax* mosquitoes were collected at this site, and the bulk abundances as calculated by the mean % non-host reads averaged across the 10 mosquitoes for *Culex narnavirus 1* and *Culex pipiens-associated Tunisia virus* were 8.3% and 10.6%, respectively. The prevalence differed more, at 30% and 90%, respectively.

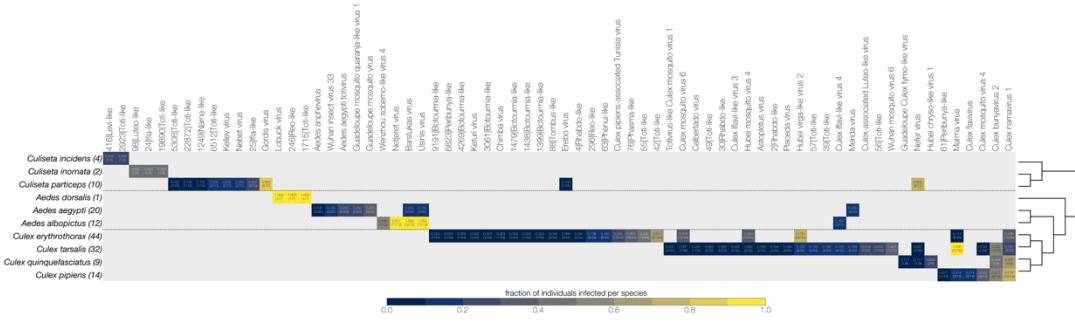


Figure 8.14 Prevalence of each virus by mosquito species.

For each virus, the fraction of individuals infected within each species was calculated, shown on a color scale. Mosquito species arranged according to a phylogeny based on the cytochrome c oxidase subunit I(COI) gene.

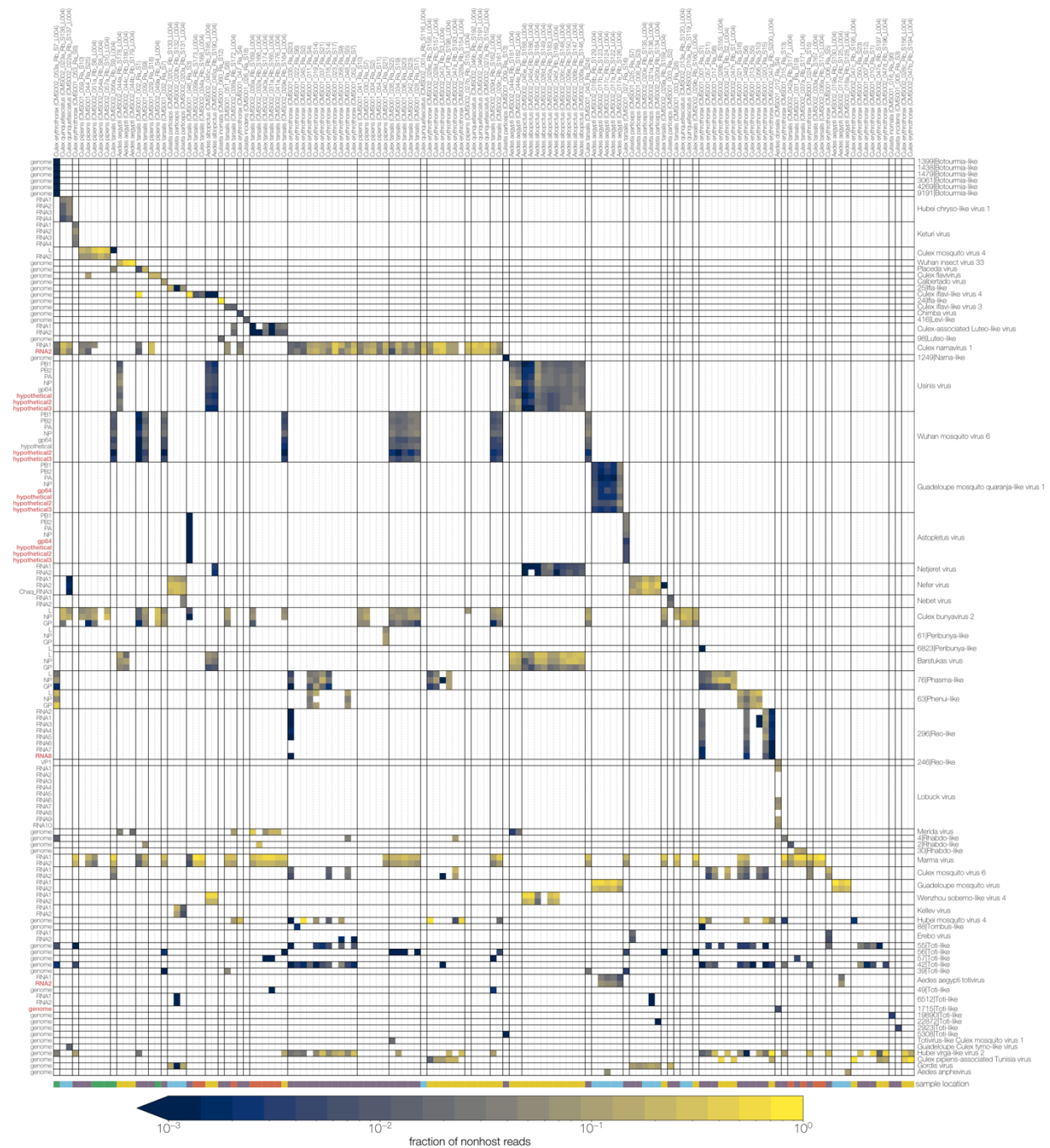
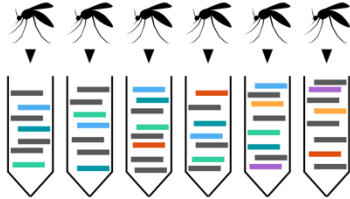


Figure 8.15 Co-occurrence analysis to identify additional viral segments.

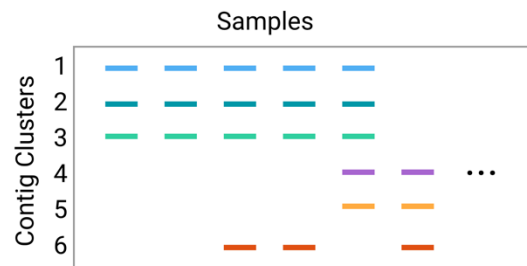
Rows indicate groups of contigs clustered based on nucleotide identity (>99%). Columns are samples with computed mosquito species indicated. Cells are colored by fraction of nonhost reads in sample with color bar legend at the bottom. Clusters of rows likely comprising a single genome are separated from other such putative genomes by black horizontal lines with identified or proposed virus names. Black vertical lines delineate

blocks of neighbouring samples with the same computed mosquito species. Sampling locations are indicated by colored rectangles at the bottom, above the color bar legend.

(1) Identify clusters of homologous contigs across samples.



(2) Group clusters by co-occurrence across samples



(3) Identify polymerases and known segments.

(4) Validate candidate viral genomes

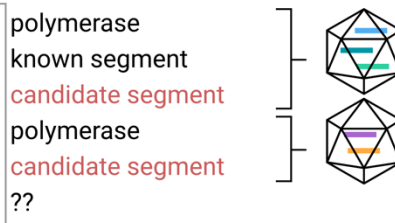


Figure 8.16 Methods for co-occurrence analysis.

Contigs are assembled from each mosquito and grouped into clusters of high sequence identity to find identical and variant sequences of the same viral segment across different mosquitoes. Here, sequences that come from repeated sampling of the same viral segment in different mosquitoes are shown in the same color. Contig clusters are then grouped together based on their co-occurrence across different samples. After identifying known viral segments using Hidden Markov Models (for RdRps) and BLASTx, the remaining contig clusters that co-occur with the known segments are assigned to the same putative viral genome. Finally, bioinformatic validation included consideration of features that supported viral assignment, such as sequence and ORF length, putative protein domains, and consistency with known genomic structure for other viruses in the same family. See Methods for details.

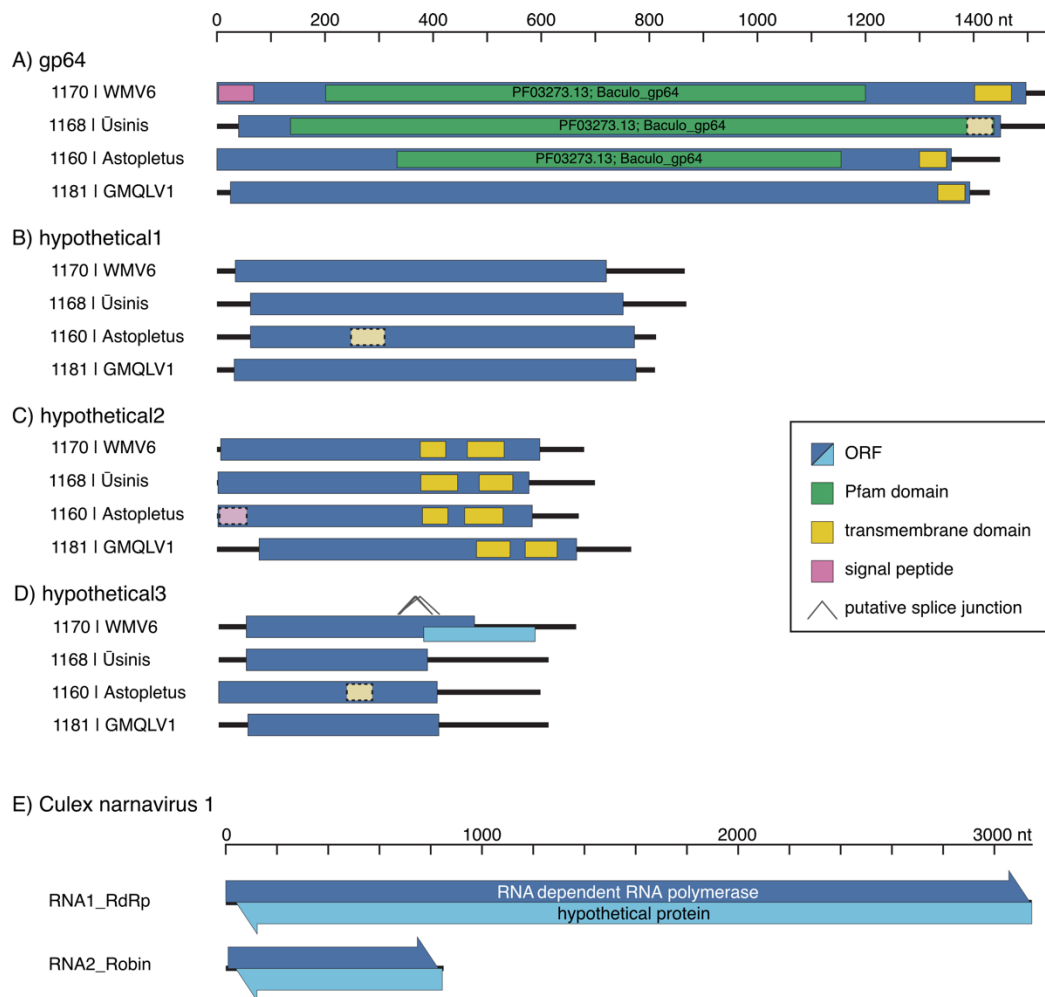


Figure 8.17 Genome diagrams for newly proposed viral segments.

Diagrams of newly discovered genome segments for quaranjaviruses in this study, including Wuhan Mosquito Virus 6 (WMV6), Usinis, Astopletus, and Guadeloupe Mosquito Quaranja-Like Virus 1 (GMQLV1) (panels A-D, respectively), and Culex narnavirus 1 (panel E), are shown (displayed 5' to 3', left to right, not aligned to each other). A-D, The longest ORFs of the new quaranjavirus segments were analyzed for the presence of transmembrane domains and similarity to known domains (see Methods for details). Weaker support for these predictions is indicated by lighter color and dashed border. In the hypothetical 3 segment of WMV6, putative splicing events could result in a jump to a frameshifted ORF, which would substantially increase the coding region of the segment. E, In both segments of the Culex narnavirus 1, long uninterrupted ORFs are found on both RNA strands (arrowheads indicate 5' to 3' direction), and in each case the codon boundaries are aligned.

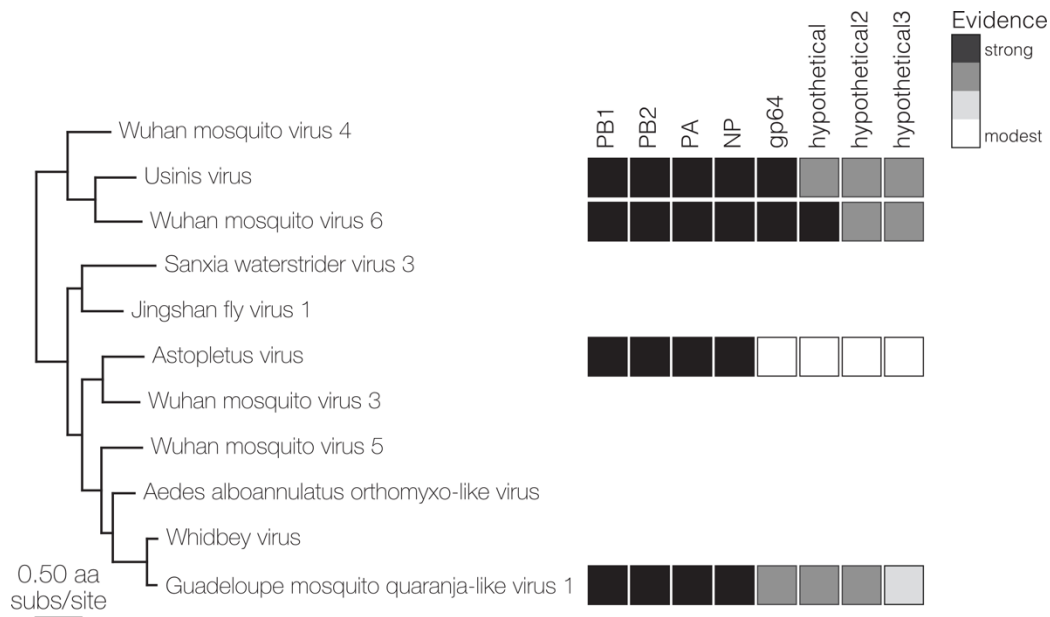


Figure 8.18 RdRp-based maximum likelihood tree spanning the quaranjaviruses in this study for which 8 segments were recovered.

The 8 boxes in the graphic to the right of the tree correspond to each of these 8 genome segments, which are labeled at top right. Levels of evidence for the existence of each segment is highlighted by a grey scale shading of the segment boxes, ranging from strong (black) to modest (white), based on: homology to BLAST hits from NCBI nt/nr database (black); assembly of contigs >500nt from SRA samples containing the RdRp, or co-occurrence in >10 samples in our data (dark grey); reads mapping from SRA and co-occurrence in 5 samples in our data (light grey); or co-occurrence in <5 samples and shared protein domains (white).

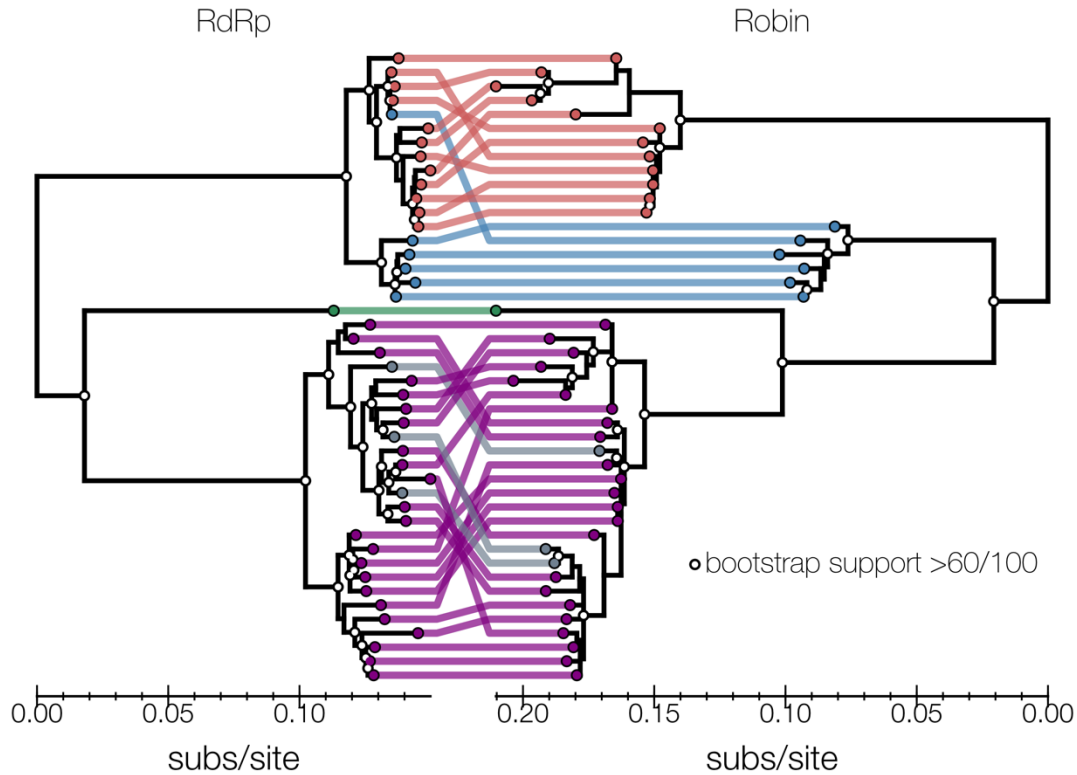


Figure 8.19 Phylogenetic relationship of naravirus RdRp and Robin segments.

Tanglegram shows two mid-point rooted nucleotide phylogenies of *Culex naravirus 1* - RdRp segment on left and Robin segment on right. Tips are colored based on visual presence of clusters on longer branches - clusters in red, blue, green, and purple are sequences reported in this study. Grey tips in the RdRp phylogeny were published by other studies, and their Robin segment counterpart was assembled here from corresponding SRA entries. Reassortment between the two segments is largely restricted to lineages within clusters, but reassortment also took place in either the red or the blue lineage. One possibility is a replacement of a red/blue lineage RdRp with a blue/red lineage or a replacement of red Robin with a diverged and unsampled lineage. White dots on nodes indicate bootstrap support >60 out of 100 replicates. Scale at the bottom of each tree is nucleotide substitutions per site.

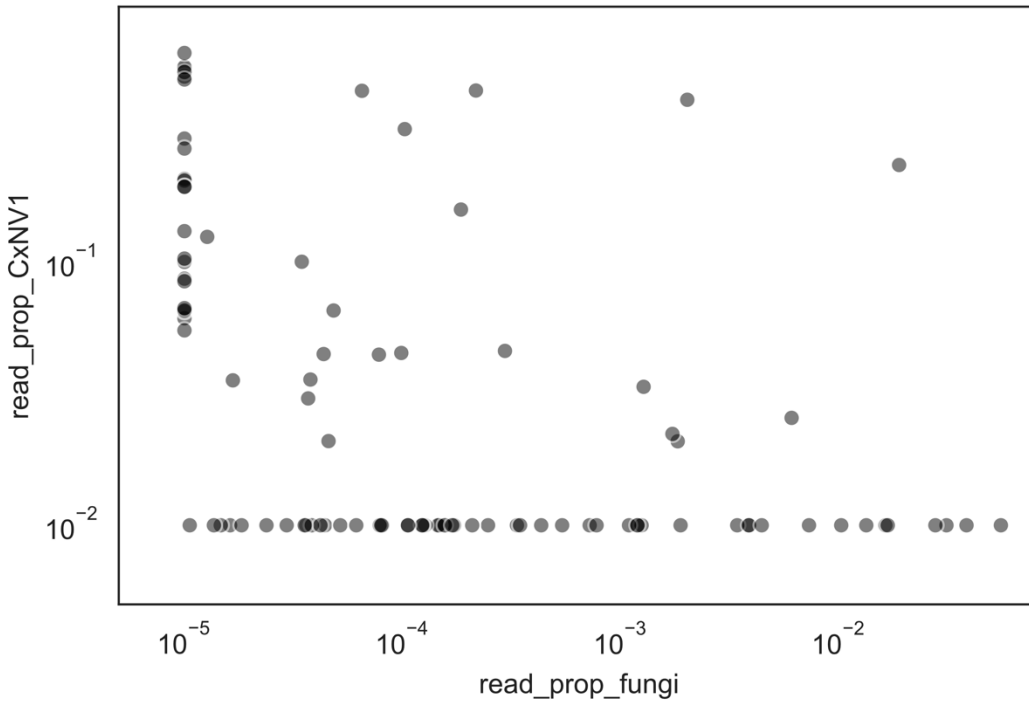


Figure 8.20 Lack of association between abundance of narnavirus and fungi in individual mosquitoes.

Tips are colored based on visual presence of clusters on longer branches -clusters in red, blue, green, and purple are sequences reported in this study. Grey tips in the RdRp phylogeny were published by other studies, and their Robin segment counterpart was assembled here from corresponding SRA entries. Reassortment between the two segments is largely restricted to lineages within clusters, but reassortment also took place in either the red or the blue lineage. One possibility is a replacement of a red/blue lineage RdRp with a blue/red lineage or a replacement of red Robin with a diverged and unsampled lineage. White dots on nodes indicate bootstrap support >60 out of 100 replicates. Scale at the bottom of each tree is nucleotide substitutions per site.

References for Chapter 8

- Aguiar, E.R.G.R., Olmo, R.P., Paro, S., Ferreira, F.V., de Faria, I.J. da S., Todjro, Y.M.H., Lobo, F.P., Kroon, E.G., Meignin, C., Gatherer, D., et al. (2015). Sequence-independent characterization of viruses based on the pattern of viral small RNAs produced by the host. *Nucleic Acids Res* 43, 6191–6206.
- Ahasan, M.S., Campos Krauer, J.M., Subramaniam, K., Lednicky, J.A., Loeb, J.C., Sayler, K.A., Wisely, S.M., and Waltzek, T.B. (2019a). Complete Genome Sequence of Mobuck Virus Isolated from a Florida White-Tailed Deer (*Odocoileus virginianus*). *Microbiol Resour Announc* 8.
- Ahasan, M.S., Subramaniam, K., Campos Krauer, J.M., Sayler, K.A., Loeb, J.C., Goodfriend, O.F., Barber, H.M., Stephenson, C.J., Popov, V.L., Charrel, R.N., et al. (2019b). Three New Orbivirus Species Isolated from Farmed White-Tailed Deer (*Odocoileus virginianus*) in the United States. *Viruses* 12.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215, 403–410.
- Anders, K.L., Indriani, C., Ahmad, R.A., Tantowijoyo, W., Arguni, E., Andari, B., Jewell, N.P., Rances, E., O'Neill, S.L., Simmons, C.P., et al. (2018). The AWED trial (Applying Wolbachia to Eliminate Dengue) to assess the efficacy of Wolbachia-infected mosquito deployments to reduce dengue incidence in Yogyakarta, Indonesia: study protocol for a cluster randomised controlled trial. *Trials* 19, 302.

- Atoni, E., Wang, Y., Karungu, S., Waruhiu, C., Zohaib, A., Obanda, V., Agwanda, B., Mutua, M., Xia, H., and Yuan, Z. (2018). Metagenomic Virome Analysis of Culex Mosquitoes from Kenya and China. *Viruses* 10.
- Atoni, E., Zhao, L., Karungu, S., Obanda, V., Agwanda, B., Xia, H., and Yuan, Z. (2019). The discovery and global distribution of novel mosquito-associated viruses in the last decade (2007-2017). *Rev. Med. Virol.* e2079.
- Batovska, J., Lynch, S.E., Cogan, N.O.I., Brown, K., Darbro, J.M., Kho, E.A., and Blacket, M.J. (2018). Effective mosquito and arbovirus surveillance using metabarcoding. *Molecular Ecology Resources* 18, 32–40.
- Bigot, D., Atyame, C.M., Weill, M., Justy, F., Herniou, E.A., and Gayral, P. (2018). Discovery of Culex pipiens associated tunisia virus: a new ssRNA(+) virus representing a new insect associated virus family. *Virus Evol* 4, vex040.
- Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120.
- Bolling, B.G., Weaver, S.C., Tesh, R.B., and Vasilakis, N. (2015). Insect-Specific Virus Discovery: Significance for the Arbovirus Community. *Viruses* 7, 4911–4928.
- Boothe, E., Medeiros, M.C.I., Kitron, U.D., Brawn, J.D., Ruiz, M.O., Goldberg, T.L., Walker, E.D., and Hamer, G.L. (2015). Identification of Avian and Hemoparasite DNA in Blood-Engorged Abdomens of Culex pipiens (Diptera; Culicidae) from a West Nile Virus Epidemic region in Suburban Chicago, Illinois. *J. Med. Entomol.* 52, 461–468.

Boratyn, G.M., Thierry-Mieg, J., Thierry-Mieg, D., Busby, B., and Madden, T.L. (2018). Magic-BLAST, an accurate DNA and RNA-seq aligner for long and short reads. *BioRxiv* 390013.

Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., Maio, N.D., et al. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology* 15, e1006650.

Breitwieser, F.P., Lu, J., and Salzberg, S.L. (2019). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics* 20, 1125–1136.

Carlson, J.S., Nelms, B., Barker, C.M., Reisen, W.K., Sehgal, R.N.M., and Cornel, A.J. (2018). Avian malaria co-infections confound infectivity and vector competence assays of *Plasmodium homopolare*. *Parasitol. Res.* 117, 2385–2394.

Chandler, J.A., Liu, R.M., and Bennett, S.N. (2015). RNA shotgun metagenomic sequencing of northern California (USA) mosquitoes uncovers viruses, bacteria, and fungi. *Front Microbiol* 6, 185.

Charon, J., Grigg, M.J., Eden, J.-S., Piera, K.A., Rana, H., William, T., Rose, K., Davenport, M.P., Anstey, N.M., and Holmes, E.C. (2019). Novel RNA viruses associated with *Plasmodium vivax* in human malaria and *Leucocytozoon* parasites in avian disease. *PLoS Pathog.* 15, e1008216.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available

Python tools for computational molecular biology and bioinformatics.

Bioinformatics 25, 1422–1423.

- Contreras-Gutiérrez, M.A., Nunes, M.R.T., Guzman, H., Uribe, S., Gallego Gómez, J.C., Suaza Vasco, J.D., Cardoso, J.F., Popov, V.L., Widen, S.G., Wood, T.G., et al. (2017). Sinu virus, a novel and divergent orthomyxovirus related to members of the genus Thogotovirus isolated from mosquitoes in Colombia. *Virology* 501, 166–175.
- Cook, S., Chung, B.Y.-W., Bass, D., Moureau, G., Tang, S., McAlister, E., Culverwell, C.L., Glücksman, E., Wang, H., Brown, T.D.K., et al. (2013). Novel Virus Discovery and Genome Reconstruction from Field RNA Samples Reveals Highly Divergent Viruses in Dipteran Hosts. *PLoS ONE* 8, e80720.
- Cooper, E., Anbalagan, S., Klumper, P., Scherba, G., Simonson, R.R., and Hause, B.M. (2014). Mobuck virus genome sequence and phylogenetic analysis: identification of a novel Orbivirus isolated from a white-tailed deer in Missouri, USA. *J. Gen. Virol.* 95, 110–116.
- DeRisi, J.L., Huber, G., Kistler, A., Retallack, H., Wilkinson, M., and Yllanes, D. (2019). An exploration of ambigrammatic sequences in narnaviruses. *Scientific Reports* 9, 1–9.
- Dinan, A.M., Lukhovitskaya, N.I., Olendraite, I., and Firth, A.E. (2019). A case for a reverse-frame coding sequence in a group of positive-sense RNA viruses. *BioRxiv* 664342.

- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21.
- Drummond, A.J., Pybus, O.G., Rambaut, A., Forsberg, R., and Rodrigo, A.G. (2003). Measurably evolving populations. *Trends in Ecology & Evolution* 18, 481–488.
- Enserink, M. (2008). ENTOMOLOGY: A Mosquito Goes Global. *Science* 320, 864–866.
- Fauver, J.R., Grubaugh, N.D., Krajacich, B.J., Weger-Lucarelli, J., Lakin, S.M., Fakoli, L.S., Bolay, F.K., Diclaro, J.W., Dabiré, K.R., Foy, B.D., et al. (2016). West African *Anopheles gambiae* mosquitoes harbor a taxonomically diverse virome including new insect-specific flaviviruses, mononegaviruses, and totiviruses. *Virology* 498, 288–299.
- Ferreira, M.A.R., and Suchard, M.A. (2008). Bayesian analysis of elapsed times in continuous-time Markov chains. *Canadian Journal of Statistics* 36, 355–368.
- Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 39, W29–W37.
- Flores, H.A., and O’Neill, S.L. (2018). Controlling vector-borne diseases by releasing modified mosquitoes. *Nat. Rev. Microbiol.* 16, 508–518.
- François, S., Filloux, D., Roumagnac, P., Bigot, D., Gayral, P., Martin, D.P., Froissart, R., and Ogliastro, M. (2016). Discovery of parvovirus-related sequences in an unexpected broad range of animals. *Sci Rep* 6, 1–13.
- Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152.

GBD 2017 Causes of Death Collaborators (2018). Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* 392, 1736–1788.

GBD 2017 DALYs and HALE Collaborators (2018). Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* 392, 1859–1922.

GBD 2017 Disease and Injury Incidence and Prevalence Collaborators (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* 392, 1789–1858.

Göertz, G.P., Miesen, P., Overheul, G.J., van Rij, R.P., van Oers, M.M., and Pijlman, G.P. (2019). Mosquito Small RNA Responses to West Nile and Insect-Specific Virus Infections in *Aedes* and *Culex* Mosquito Cells. *Viruses* 11.

Grubaugh, N.D., Sharma, S., Krajacich, B.J., Iii, L.S.F., Bolay, F.K., Li, J.W.D., Johnson, W.E., Ebel, G.D., Foy, B.D., and Brackney, D.E. (2015). Xenosurveillance: A Novel Mosquito-Based Approach for Examining the Human-Pathogen Landscape. *PLOS Neglected Tropical Diseases* 9, e0003628.

Grybchuk, D., Akopyants, N.S., Kostygov, A.Y., Konovalovas, A., Lye, L.-F., Dobson, D.E., Zangger, H., Fasel, N., Butenko, A., Frolov, A.O., et al. (2018). Viral discovery and

- diversity in trypanosomatid protozoa with a focus on relatives of the human parasite *Leishmania*. *Proc Natl Acad Sci USA* 115, E506–E515.
- Guindon, S., and Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst Biol* 52, 696–704.
- Harl, J., Himmel, T., Valkiūnas, G., and Weissenböck, H. (2019). The nuclear 18S ribosomal DNAs of avian haemosporidian parasites. *Malar. J.* 18, 305.
- Harris, S.R. (2018). SKA: Split Kmer Analysis Toolkit for Bacterial Genomic Epidemiology. *BioRxiv* 453142.
- Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22, 160–174.
- Hillman, B.I., and Cai, G. (2013). The family narnaviridae: simplest of RNA viruses. *Adv. Virus Res.* 86, 149–176.
- Hoffmann, A.A., Montgomery, B.L., Popovici, J., Iturbe-Ormaetxe, I., Johnson, P.H., Muzzi, F., Greenfield, M., Durkan, M., Leong, Y.S., Dong, Y., et al. (2011). Successful establishment of *Wolbachia* in *Aedes* populations to suppress dengue transmission. *Nature* 476, 454–457.
- Hofmeister, E.K. (2011). West Nile virus: North American experience. *Integrative Zoology* 6, 279–289.
- Huestis, D.L., Dao, A., Diallo, M., Sanogo, Z.L., Samake, D., Yaro, A.S., Ousman, Y., Linton, Y.-M., Krishna, A., Veru, L., et al. (2019). Windborne long-distance migration of malaria mosquitoes in the Sahel. *Nature* 574, 404–408.
- iNaturalist (2020). iNaturalist Research-grade Observations. Occurrence dataset.

- Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33, 511–518.
- Katzourakis, A., and Gifford, R.J. (2010). Endogenous Viral Elements in Animal Genomes. *PLOS Genetics* 6, e1001191.
- Kittayapong, P., Baisley, K.J., Baimai, V., and O’Neill, S.L. (2000). Distribution and diversity of *Wolbachia* infections in Southeast Asian mosquitoes (Diptera: Culicidae). *J. Med. Entomol.* 37, 340–345.
- Kraemer, M.U.G., Reiner, R.C., Brady, O.J., Messina, J.P., Gilbert, M., Pigott, D.M., Yi, D., Johnson, K., Earl, L., Marczak, L.B., et al. (2019). Past and future spread of the arbovirus vectors *Aedes aegypti* and *Aedes albopictus*. *Nat Microbiol* 4, 854–863.
- Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359.
- Lee, J.S., Bevins, S.N., Serieys, L.E.K., Vickers, W., Logan, K.A., Aldredge, M., Boydston, E.E., Lyren, L.M., McBride, R., Roelke-Parker, M., et al. (2014). Evolution of Puma Lentivirus in Bobcats (*Lynx rufus*) and Mountain Lions (*Puma concolor*) in North America. *Journal of Virology* 88, 7727–7737.
- Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659.
- Li, C.-X., Shi, M., Tian, J.-H., Lin, X.-D., Kang, Y.-J., Chen, L.-J., Qin, X.-C., Xu, J., Holmes, E.C., and Zhang, Y.-Z. (2015). Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *ELife* 4, e05378.

- Moreira, L.A., Iturbe-Ormaetxe, I., Jeffery, J.A., Lu, G., Pyke, A.T., Hedges, L.M., Rocha, B.C., Hall-Mendelin, S., Day, A., Riegler, M., et al. (2009). A *Wolbachia* Symbiont in *Aedes aegypti* Limits Infection with Dengue, Chikungunya, and Plasmodium. *Cell* 139, 1268–1278.
- Müller, N.F., Stolz, U., Dudas, G., Stadler, T., and Vaughan, T.G. (2019). Bayesian inference of reassortment networks reveals fitness benefits of reassortment in human influenza viruses. *BioRxiv* 726042.
- Obbard, D.J., Shi, M., Roberts, K.E., Longdon, B., and Dennis, A.B. (2019). A new lineage of segmented RNA viruses infecting animals. *BioRxiv* 741645.
- O’Neill, S.L., Ryan, P.A., Turley, A.P., Wilson, G., Retzki, K., Iturbe-Ormaetxe, I., Dong, Y., Kenny, N., Paton, C.J., Ritchie, S.A., et al. (2019). Scaled deployment of *Wolbachia* to protect the community from dengue and other *Aedes* transmitted arboviruses. *Gates Open Res* 2.
- Pagès, N., and Cohnstaedt, L.W. (2018). 8. Mosquito-borne diseases in the livestock industry. In *Ecology and Control of Vector-Borne Diseases*, C. Garros, J. Bouyer, W. Takken, and R.C. Smallegange, eds. (The Netherlands: Wageningen Academic Publishers), pp. 195–219.
- Palatini, U., Miesen, P., Carballar-Lejarazu, R., Ometto, L., Rizzo, E., Tu, Z., van Rij, R.P., and Bonizzoni, M. (2017). Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors *Aedes aegypti* and *Aedes albopictus*. *BMC Genomics* 18, 512.

- Parry, R., and Asgari, S. (2018). Aedes Anphevirus: an Insect-Specific Virus Distributed Worldwide in Aedes aegypti Mosquitoes That Has Complex Interplays with Wolbachia and Dengue Virus Infection in Cells. *J. Virol.* 92.
- Pettersson, J.H.-O., Shi, M., Eden, J.-S., Holmes, E.C., and Hesson, J.C. (2019). Meta-Transcriptomic Comparison of the RNA Viromes of the Mosquito Vectors Culex pipiens and Culex torrentium in Northern Europe. *Viruses* 11.
- Rambaut, A. (2000). Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* 16, 395–399.
- Rambaut, A., Drummond, A.J., Xie, D., Baele, G., and Suchard, M.A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol* 67, 901–904.
- Rasgon, J.L., and Scott, T.W. (2004). An initial survey for Wolbachia (Rickettsiales: Rickettsiaceae) infections in selected California mosquitoes (Diptera: Culicidae). *J. Med. Entomol.* 41, 255–257.
- RATNASINGHAM, S., and HEBERT, P.D.N. (2007). bold: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Mol Ecol Notes* 7, 355–364.
- Reeves, L.E., Gillett-Kaufman, J.L., Kawahara, A.Y., and Kaufman, P.E. (2018). Barcoding blood meals: New vertebrate-specific primer sets for assigning taxonomic identities to host DNA from mosquito blood meals. *PLOS Neglected Tropical Diseases* 12, e0006767.
- Richaud, A., Frézal, L., Tahan, S., Jiang, H., Blatter, J.A., Zhao, G., Kaur, T., Wang, D., and Félix, M.-A. (2019). Vertical transmission in Caenorhabditis nematodes of RNA

- molecules encoding a viral RNA-dependent RNA polymerase. *Proc. Natl. Acad. Sci. U.S.A.* 116, 24738–24747.
- Ritchie, S.A., van den Hurk, A.F., Smout, M.J., Staunton, K.M., and Hoffmann, A.A. (2018). Mission Accomplished? We Need a Guide to the ‘Post Release’ World of *Wolbachia* for *Aedes*-borne Disease Control. *Trends in Parasitology* 34, 217–226.
- Roumpeka, D.D., Wallace, R.J., Escalettes, F., Fotheringham, I., and Watson, M. (2017). A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. *Front. Genet.* 8.
- Ruby, J.G., Bellare, P., and DeRisi, J.L. (2013). PRICE: Software for the Targeted Assembly of Components of (Meta) Genomic Sequence Data. *G3* 3, 865–880.
- Sadeghi, M., Popov, V., Guzman, H., Phan, T.G., Vasilakis, N., Tesh, R., and Delwart, E. (2017). Genomes of viral isolates derived from different mosquitos species. *Virus Res.* 242, 49–57.
- Sadeghi, M., Altan, E., Deng, X., Barker, C.M., Fang, Y., Coffey, L.L., and Delwart, E. (2018). Virome of > 12 thousand *Culex* mosquitoes from throughout California. *Virology* 523, 74–88.
- Sagulenko, P., Puller, V., and Neher, R.A. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol* 4.
- Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4, 406–425.
- Shane, J.L., Grogan, C.L., Cwalina, C., and Lampe, D.J. (2018). Blood meal-induced inhibition of vector-borne disease by transgenic microbiota. *Nat Commun* 9, 4127.

- Shi, C., Liu, Y., Hu, X., Xiong, J., Zhang, B., and Yuan, Z. (2015). A metagenomic survey of viral abundance and diversity in mosquitoes from Hubei province. *PLoS ONE* 10, e0129845.
- Shi, C., Beller, L., Deboutte, W., Yinda, K.C., Delang, L., Vega-Rúa, A., Failloux, A.-B., and Matthijnsens, J. (2019). Stable distinct core eukaryotic viromes in different mosquito species from Guadeloupe, using single mosquito viral metagenomics. *Microbiome* 7, 121.
- Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., Qin, X.-C., Li, J., Cao, J.-P., Eden, J.-S., et al. (2016). Redefining the invertebrate RNA virosphere. *Nature*.
- Shi, M., Neville, P., Nicholson, J., Eden, J.-S., Imrie, A., and Holmes, E.C. (2017). High-Resolution Metatranscriptomics Reveals the Ecological Dynamics of Mosquito-Associated RNA Viruses in Western Australia. *J. Virol.* 91.
- Shi, M., Lin, X.-D., Chen, X., Tian, J.-H., Chen, L.-J., Li, K., Wang, W., Eden, J.-S., Shen, J.-J., Liu, L., et al. (2018). The evolutionary history of vertebrate RNA viruses. *Nature* 556, 197.
- Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* 4.
- Talbi, C., Lemey, P., Suchard, M.A., Abdelatif, E., Elharrak, M., Jalal, N., Faouzi, A., Echevarría, J.E., Vazquez Morón, S., Rambaut, A., et al. (2010). Phylodynamics and Human-Mediated Dispersal of a Zoonotic Virus. *PLoS Pathogens* 6, e1001166.

- Tedrow, R.E., Rakotomanga, T., Nepomichene, T., Howes, R.E., Ratovonjato, J., Ratsimbaoa, A.C., Svenson, G.J., and Zimmerman, P.A. (2019). Anopheles mosquito surveillance in Madagascar reveals multiple blood feeding behavior and Plasmodium infection. *PLOS Neglected Tropical Diseases* 13, e0007176.
- Ter Horst, A.M., Nigg, J.C., Dekker, F.M., and Falk, B.W. (2019). Endogenous Viral Elements Are Widespread in Arthropod Genomes and Commonly Give Rise to PIWI-Interacting RNAs. *J. Virol.* 93.
- Thongsripong, P., Chandler, J.A., Green, A.B., Kittayapong, P., Wilcox, B.A., Kapan, D.D., and Bennett, S.N. (2018). Mosquito vector-associated microbiota: Metabarcoding bacteria and eukaryotic symbionts across habitat types in Thailand endemic for dengue and other arthropod-borne diseases. *Ecol Evol* 8, 1352–1368.
- Tomazatos, A., Jansen, S., Pfister, S., Török, E., Maranda, I., Horváth, C., Keresztes, L., Spînu, M., Tannich, E., Jöst, H., et al. (2019). Ecology of West Nile Virus in the Danube Delta, Romania: Phylogeography, Xenosurveillance and Mosquito Host-Feeding Patterns. *Viruses* 11, 1159.
- Tyler, S., Bolling, B.G., Blair, C.D., Brault, A.C., Pabbaraju, K., Armijos, M.V., Clark, D.C., Calisher, C.H., and Drebot, M.A. (2011). Distribution and phylogenetic comparisons of a novel mosquito flavivirus sequence present in *Culex tarsalis* Mosquitoes from western Canada with viruses isolated in California and Colorado. *Am. J. Trop. Med. Hyg.* 85, 162–168.
- Van Nguyen, H., and Lavenier, D. (2009). PLAST: parallel local alignment search tool for database comparison. *BMC Bioinformatics* 10, 329.

- Vasilakis, N., and Tesh, R.B. (2015). Insect-specific viruses and their potential impact on arbovirus transmission. *Curr Opin Virol* 15, 69–74.
- Waldron, F.M., Stone, G.N., and Obbard, D.J. (2018). Metagenomic sequencing suggests a diversity of RNA interference-like responses to viruses across multicellular eukaryotes. *PLOS Genetics* 14, e1007533.
- Washino, R.K., and Tempelis, C.H. (1983). Mosquito Host Bloodmeal Identification: Methodology and Data Analysis. *Annual Review of Entomology* 28, 179–201.
- Werren, J.H., Baldo, L., and Clark, M.E. (2008). Wolbachia: master manipulators of invertebrate biology. *Nat. Rev. Microbiol.* 6, 741–751.
- Wheeler, D.C., Waller, L.A., and Biek, R. (2010). Spatial analysis of feline immunodeficiency virus infection in cougars. *Spatial and Spatio-Temporal Epidemiology* 1, 151–161.
- WHO (2017). Global Vector Control Response 2017-2030.
- Wilson, A.L., Courtenay, O., Kelly-Hope, L.A., Scott, T.W., Takken, W., Torr, S.J., and Lindsay, S.W. (2020). The importance of vector control for the control and elimination of vector-borne diseases. *PLoS Negl Trop Dis* 14, e0007831.
- Xia, H., Wang, Y., Atoni, E., Zhang, B., and Yuan, Z. (2018). Mosquito-Associated Viruses in China. *Virol Sin* 33, 5–20.
- Xiao, P., Han, J., Zhang, Y., Li, C., Guo, X., Wen, S., Tian, M., Li, Y., Wang, M., Liu, H., et al. (2018a). Metagenomic Analysis of Flaviviridae in Mosquito Viromes Isolated From Yunnan Province in China Reveals Genes From Dengue and Zika Viruses. *Front Cell Infect Microbiol* 8, 359.

- Xiao, P., Li, C., Zhang, Y., Han, J., Guo, X., Xie, L., Tian, M., Li, Y., Wang, M., Liu, H., et al. (2018b). Metagenomic Sequencing From Mosquitoes in China Reveals a Variety of Insect and Human Viruses. *Front Cell Infect Microbiol* 8, 364.
- Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J Mol Evol* 39, 306–314.
- Zídková, L., Cepicka, I., Szabová, J., and Svobodová, M. (2012). Biodiversity of avian trypanosomes. *Infect. Genet. Evol.* 12, 102–112.
- Ziv, J., and Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory* 24, 530–536.

Chapter 9 An exploration of ambigrammatic sequences in

narnaviruses

Authors:

Joseph L. DeRisi^{1,2}, Greg Huber¹, Amy Kistler¹, Hanna Retallack², Michael Wilkinson^{1,3} & David Yllanes^{1*}

Affiliations:

¹ Chan Zuckerberg Biohub, 499 Illinois Street, San Francisco, CA, 94158, USA.

² Department of Biochemistry and Biophysics, University of California, San Francisco, California, USA.

³ School of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, England.

*email: david.yllanes@czbiohub.org

Includes material previously published in:

DeRisi JL, Huber G, Kistler A, Retallack H, Wilkinson M, Yllanes D. An exploration of ambigrammatic sequences in narnaviruses. *Sci Rep.* 2019;9(1):17982. Published 2019 Nov 29. doi:10.1038/s41598-019-54181-3

Abstract

Narnaviruses have been described as positive-sense RNA viruses with a remarkably simple genome of ~3 kb, encoding only a highly conserved RNA-dependent RNA polymerase (RdRp). Many narnaviruses, however, are ‘ambigrammatic’ and harbour an additional uninterrupted open reading frame (ORF) covering almost the entire length of the reverse complement strand. No function has been described for this ORF, yet the absence of stops is conserved across diverse narnaviruses, and in every case the codons in

the reverse ORF and the RdRp are aligned. The >3kb ORF overlap on opposite strands, unprecedented among RNA viruses, motivates an exploration of the constraints imposed or alleviated by the codon alignment. Here, we show that only when the codon frames are aligned can all stop codons be eliminated from the reverse strand by synonymous single-nucleotide substitutions in the RdRp gene, suggesting a mechanism for de novo gene creation within a strongly conserved amino acid sequence. It will be fascinating to explore what implications this coding strategy has for other aspects of narnavirus biology. Beyond narnaviruses, our rapidly expanding catalogue of viral diversity may yet reveal additional examples of this broadly-extensible principle for ambigrammatic-sequence development.

Introduction

Narnaviruses (a contraction of naked RNA viruses) are RNA viruses with a seemingly simple genome (Hillman and Cai 2013). The only manifestation of narnaviral infections documented to date is the presence of large concentrations of the viral RNA in the cytoplasm of the host cell, often detected as double-stranded RNA. These infections were first observed in cultured yeast (Kadowaki and Halvorson 1971; Wesolowski and Wickner 1984). Subsequent metagenomic sequencing revealed narnaviruses in other fungi (Osaki et al. 2016), oomycetes (Cai et al. 2012), mosquitoes (S. Cook et al. 2013; Chandler, Liu, and Bennett 2015; Göertz et al. 2019; Shi et al. 2017), other arthropods (Harvey et al. 2019), algae (Waldron, Stone, and Obbard 2018), trypanosomatid (Grybchuk et al. 2018; Akopyants et al. 2016; Lye et al. 2016) and potentially apicomplexans (Charon et al. 2019), although the precise host species is not always clear.

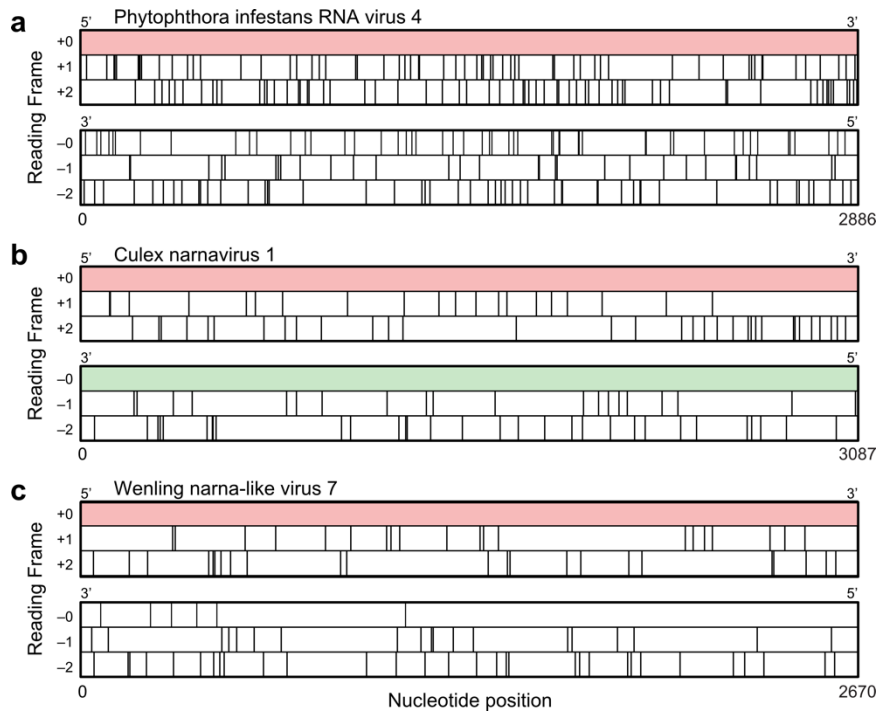


Figure 9.1 Ambigrammatic sequences in narnaviruses. Coding region for the RNA-dependent RNA polymerase (RdRp) of *Phytophthora infestans* RNA virus 4 (a), *Culex narnavirus* 1 (b), and *Wenling narna-like virus* 7 (c) in the reference +0 frame and all five other reading frames (see Fig. 2 for our frame-labelling conventions). Stop codons in each frame are depicted as vertical lines. Large uninterrupted open reading frames (ORFs) are highlighted in colour.

The known examples of narnaviruses are approximately 3 kb in size and code for a single protein, an RNA-dependent RNA polymerase (RdRp), with the exception of two putative bipartite narnaviruses (Grybchuk et al. 2018; Lye et al. 2016; Charon et al. 2019). A remarkable feature of some narnaviruses is the

existence of an additional large open reading frame (ORF), defined as a region devoid of stop codons, which spans nearly the full length of the reverse complement sequence of the virus genome. We refer to these examples, with large reverse open reading frame (rORF) features, as ambigrammatic narnaviruses, where the adjective is derived from ambigram, a set of letters with two, orientation-dependent readings (Hofstadter 1985). Such large uninterrupted rORFs are very unlikely to occur by chance, since for a typical

nucleotide sequence, there are likely to be codons for which the reverse complement read is a stop codon. This is illustrated in Figure 9.1a, which shows a typical viral genome spanned by a single ORF and where all five alternative reading frames contain many stop codons. In contrast, Figure 9.1b shows the genetic sequence of *Culex narnavirus 1*, where one of the reverse reading frames also has an uninterrupted ORF spanning nearly the whole sequence. All known examples of ambigrammatic narnaviruses have the large reverse ORF in a reading frame where the codons are aligned with those in the forward direction (i.e., in “frame -0” in the conventions of Figure 9.2). Figure 9.1c considers an intriguing intermediate case, which will be discussed further below. The existence of these very long rORFs is a surprising observation which demands an exploration. Interestingly, not all narnaviruses have an ambigrammatic genome (indeed, the example of Figure 9.1a is also in the Narnaviridae family). In agreement with previously reported observations (Dinan et al. 2019), overlaying the lengths of detectable rORFs on the narnavirus phylogeny shows that ambigrammatic sequences are present in at least two different

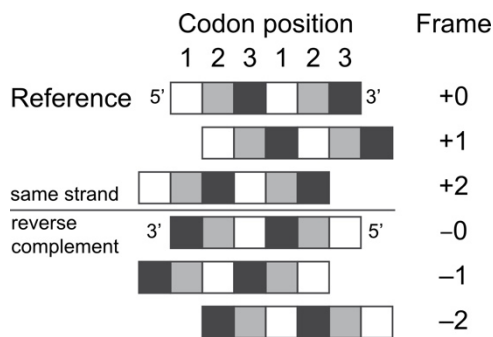


Figure 9.2 Labelling conventions used in this paper for reading frames.

clades. This finding, illustrated in Figure 9.3, indicates that the ambigrammatic feature is polyphyletic and may have been gained and lost multiple times in the evolution of this viral family

In this paper we explore how a large open reading frame might arise in the reverse complement

sequence of the RNA. While there is an extensive literature on the statistics of codons in overlapping genes (discussed in (Smith and Waterman 1980; Shukla 2015; Brandes and

Linial 2016; Lèbre and Gascuel 2017)) and on the evolution of overlapping genes in viruses ((Staden 1984; Krakauer 2002; 2002; Belshaw, Pybus, and Rambaut 2007; Rancurel et al. 2009; Chirico, Vianelli, and Belshaw 2010), the vast majority of these analyses handle cases of overlapping genes being translated in the same direction, with few papers considering antisense overlaps (see, e.g. (Merino et al. 1994)).

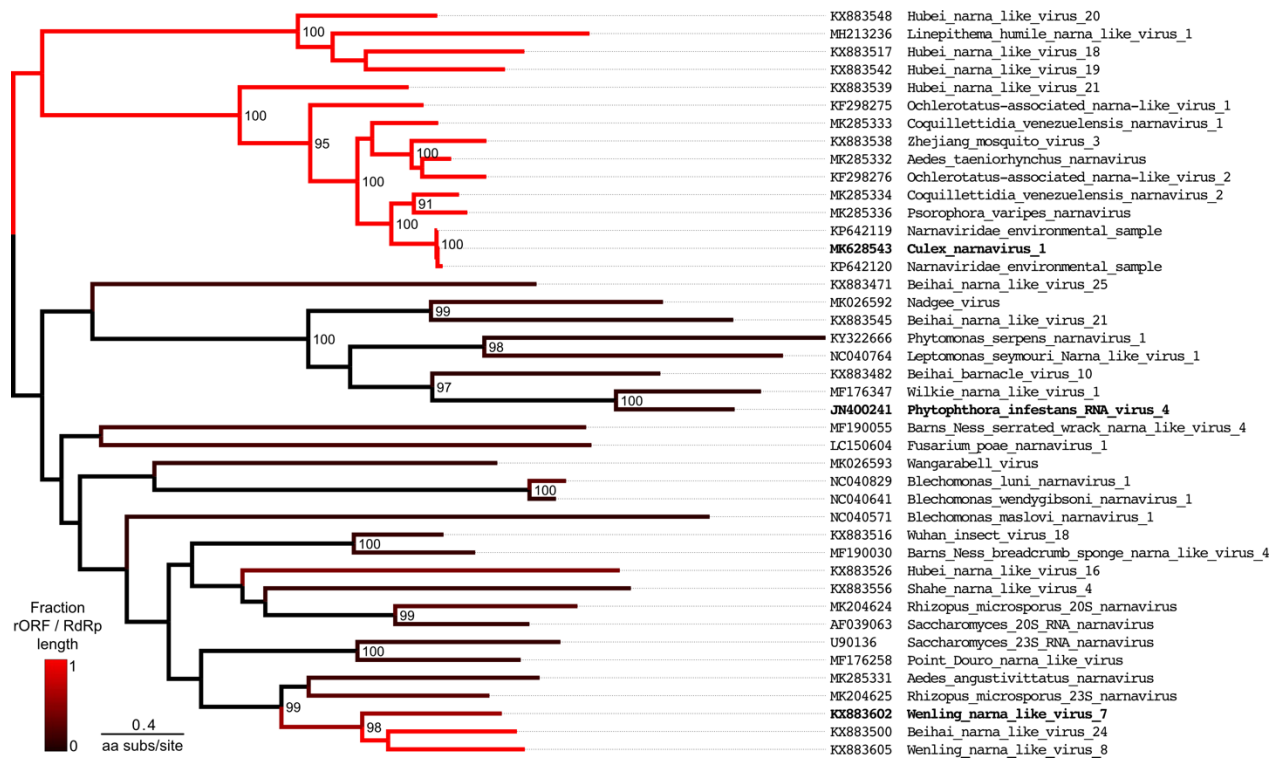


Figure 9.3 Maximum likelihood tree of amino-acid sequences for RNA-dependent RNA polymerase (RdRp) of 42 representative narnaviruses, identified by homology to the narnaviruses observed in culture, *Culex narnavirus 1* and *Phytophthora infestans virus 4*.

Unrooted tree shown with midpoint rooting for display. Branch colouring indicates the fraction of RdRp coding sequence overlapped by the longest open reading frame (defined as a region uninterrupted by stops) in the reverse complement aligned frame (-0 frame) for sequences at tips (see colour bar, bottom left). The sequence names in bold correspond to those shown in Fig. 1. Numbers at nodes indicate bootstrap values (shown when >80). The branch length is given by the amino-acid substitutions per site, as illustrated by the scale bar.

In contrast, narnaviruses are unique in demonstrating long uninterrupted reading frames in the reverse complement sequence of RNA. Moreover, most analyses of overlapping genes are complicated by the fact that, in the general case, the sequences of both genes can evolve. Analysing cases where one of the overlapping genes is more strongly conserved is very difficult in general, and most earlier works treat the two overlapping genes symmetrically (Smith and Waterman 1980). In this work we adopt a different approach and treat one of the genes as having a fixed amino acid sequence. This could represent a gene with a critical function such as viral polymerases which are often strongly conserved. We suggest that even under these very strong constraints it may be possible for a novel large rORF to arise, but only in one of the three possible reverse reading frames. The two alternative forward reading frames, though not directly relevant for experimental observations in narnaviruses, are also discussed for completeness. We discuss how this mechanism may give rise to a narnaviral genome that is ambigrammatic, in the sense that it can code for two proteins, each translated from one of the two complementary strands of RNA. Finally, we note that ambigrammatic coding in other RNA viruses has not been observed to date, and make some speculative remarks about the potential role of the narnavirus rORF.

Results

Synonym sudoku.

Usually, an alternative reading frame for a gene is found to have many stop codons, so that it cannot, therefore, code for a polypeptide or protein of useful length. Let

us assume that a nucleotide sequence already codes for a gene which has an essential function. In this case the sequence of amino acids should not be changed. The nucleotide sequence, on the other hand, can still be altered by replacing codons with synonyms (codons which code for the same amino acid). We can ask whether a sequence of single-nucleotide mutations can remove the stop codons that are expected to occur in alternative reading frames, while replacing codons with others that are synonyms in the original reading frame. For some base sequences this will always be possible, but we consider whether all of the stop codons in an alternative reading frame can be removed by single-nucleotide mutations in an arbitrary polypeptide sequence. There are five possible alternative reading frames (two in the original strand and three in the reverse complement sequence, see Figure 9.2). Like a Sudoku puzzle, there is no alternative but to explore all five possibilities in turn. In our discussion, we frame our argument in terms of an RNA genome, so that the base pairings are A=U and G=C, and the stop codons are UAA, UAG, UGA. For the purposes of understanding ambigrammatic sequences observed in narnaviruses, we focus below on the three reverse complement reading frames; the two alternate forward reading frames are discussed in Methods section.

Frame -0: aligned complementary reading frames.

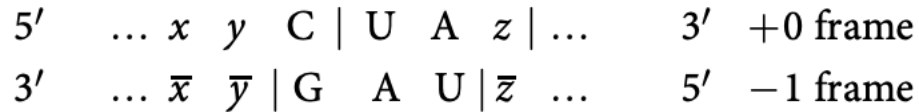
First, consider the case where the reading frames of the forward and reverse complement sequences have their codons aligned (the -0 frame). In this context, stop codons UAA, UAG, UGA become, respectively, UUA, CUA, UCA in the +0 frame, encoding the amino acids Leu, Leu, Ser. Thus, only instances of leucine and serine in the +0 reading frame can result in stop codons in the -0 reading frame. We should now

consider if synonymous substitution of Leu or Ser codons in the +0 frame can remove the stop codons in the -0 frame. The synonyms of Leu are CU*, UUA, and UUG (where * means any nucleotide). The synonyms of Ser are UC*, AGU, and AGC. Hence, the Leu codon UUA can be transformed to UUG by a single substitution. Similarly, the Leu codon CUA can be transformed to CUU, CUG or CUC by single substitutions, while the Ser codon UCA is transformed to UCU, UCG or UCC by single substitutions. Thus, when frames are aligned, 7 types of single-nucleotide synonymous substitutions in the +0 frame are sufficient to remove stops in the reverse direction.

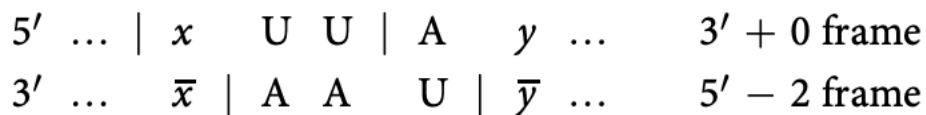
Ambigrammatic narnavirus sequences have been identified in fungi¹⁷, where the mitochondrial genetic code uses only two stop codons. The fact that the narnavirus rORF sequences lack all three possible stop codons suggests that translation of the narnavirus rORF in these hosts is not occurring in the mitochondria, unlike viruses of the related Mitovirus genus (Hillman and Cai 2013).

Frames -1, -2: staggered complementary reading frames.

The cases of staggered reverse complement reading frames are more complex, because a codon in the original +0 reading frame straddles two codons in each of these alternate reading frames. Let us first study the case of the -1 reading frame, where the codons of the reverse reading frame are shifted towards the 3' end (Figure 9.2). Consider the sequence CUA in the forward direction, with a shift of one base between frames, so we have, using | to denote triplet codon boundaries, x, y, ... for unspecified bases and x', y' ... for their pairing complementary bases:



Note that the reversed read UAG is one of the stop codons we wish to avoid, which we could do by changing either UAz or xyC to a synonym. Let us start with UAz, which can only be Tyr (either UAU or UAC; the other two values of y correspond to stops). UAU and UAC are the only codons that translate to Tyr, so it is not possible to find a synonym of UAz that avoids the stop in the reverse sequence. The alternative is to find a synonym of xyC. Here we note that if transforming C to U is the only synonymous change, this will still yield a stop codon in the reverse read frame. This occurs if xyC represents Asn, Asp, Cys, His, Phe, Tyr, and those Ser codons which begin with AG. Exactly the same restrictions arise from considering the U|UA sequence, and no additional cases of non-removable stops arise from considering U|CA. Thus we find that the following combinations of four +0 codon pairs will prevent an ambigrammatic partner rORF in the -1 frame: (Asn,Tyr), (Asp,Tyr), (Cys, Tyr), (His, Tyr), (Phe, Tyr), (Tyr, Tyr), and some cases of (Ser,Tyr). In the case of the -2 frame, the codons are shifted to the 5' end. Let us consider when the stop codon in the reverse-read direction is UAA



Consider the possible synonym changes to the codons in the +0 frame that will remove the stop codon in the -2 frame. The second codon, beginning with A, can be either Arg,

Asn, Ile, Lys, Met, Ser or Tr. Of these, changing the first base can only give a synonym for Arg (AGC and AGT, coding for Ser, yield synonyms by changing two bases). The first codon, ending in UU, can only be Ile, Leu, Phe or Val and it is not possible to obtain a synonym by changing its second base. Finally, it is always possible to obtain a synonym by changing the third base, but, if the first codon is UUU, there is only one synonym, UUC, which also gives a stop in the complementary chain. Considering the case of UC|A yields the same set of excluded combinations, and CU|A does not lead to further examples. We conclude that if the +0 frame contains Phe followed by Asn, Ile, Lys, Met, Ser or Tr; then there is a stop codon in the reverse-read frame which cannot be removed.

The potential of synonymous substitutions for generating a new ORF.

In conclusion, the -0 reading frame is the only one of the three reverse reading frames where stop codons can always be removed by synonymous single-nucleotide mutations in the +0 original (conserved) amino-acid sequence. This is consistent with the observation that the forward and the reverse open reading frames in ambigrammatic narnaviruses have their codons aligned. Thus, synonymous substitutions provide a possible route to the generation of a new ORF (as discussed in Methods section, large ORFs could also be generated in the +2, but not in the +1, reading frame by single-nucleotide synonymous substitutions, although this possibility has not been observed in narnaviruses). We should consider the extent to which this is an effective mechanism for creating new coding potential. In addition to removing stop codons in the -0 reading frame, replacing other codons of the original +0 ORF by synonyms can result in changes to the amino-acid sequence of the rORF, without changing the protein encoded in the +0

frame. It is these synonymous substitutions which allow scope for evolutionary adaptation of the new protein. If the changes in the +0 frame are strictly limited to synonym substitutions, the range of available proteins that can be produced by the new ORFs is quite constrained. We discuss how this can be quantified in the Methods section.

A hypothesis about the evolution of narnaviruses.

First we should consider a null hypothesis, that the rORF (spanning approximately 3 kb) is a chance occurrence. A priori this appears to be extremely unlikely: given that there are three stop codons out of 64 possibilities, we expect that the typical distance between stop codons will be approximately 60 base pairs. If the stop codons are randomly and independently scattered, with mean separation $\langle N \rangle$, the probability of a string of N bases containing no stops is expected to be $P(N) = \exp(-N/\langle N \rangle)$. If $\langle N \rangle = 60$, the probability of finding a sequence of 3 kb lacking a stop codon is approximately $\approx \exp(-50) \approx 2 \times 10^{-22}$. The correct value of $\langle N \rangle$ depends upon the distribution of codons, so as a more refined check we examined the distribution of stop codons in each of the five possible alternative frames which arise from choosing a random permutation of the codons of the RdRp gene. Including both ambigrammatic and non-ambigrammatic narnaviruses to generate a null distribution of lengths of ORFs that may overlap the RdRp, we find that the expected probability of a long sequence without a stop codon is indeed well approximated by an exponential distribution, and that the expected probability of an ORF with the observed length is negligible (Figure 9.4). The scale length $\langle N \rangle$ varies from frame to frame, but is not greatly different from the simple estimate, $\langle N \rangle \approx 60$. Of course, even a highly unlikely feature can arise and become fixed in a population. However, there

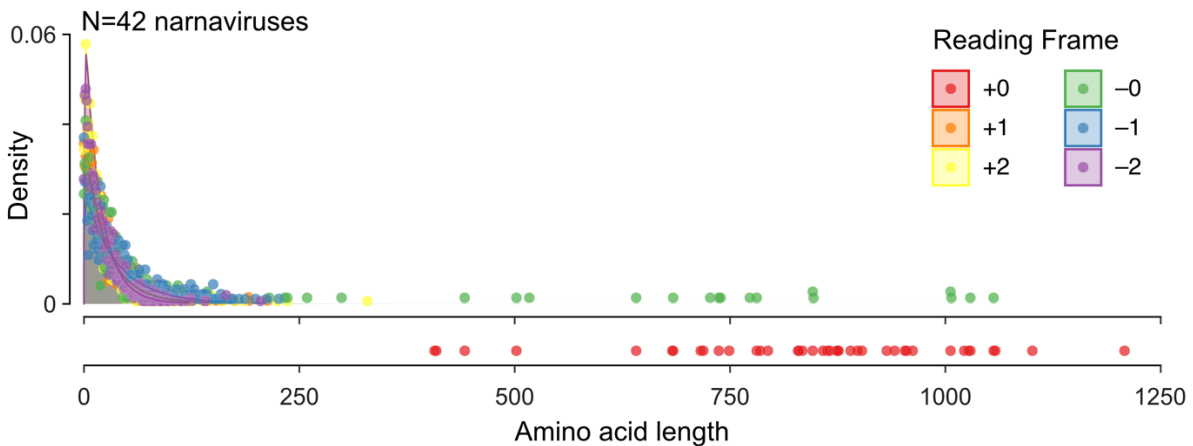


Figure 9.4 Probability distribution for ORF lengths in narnavirus-like sequences.

Shading shows distribution of ORF lengths coloured by reading frame after codon permutation test on RdRp coding sequences of 42 representative narnaviruses as in Fig. 3. In brief, codons are randomly re-ordered and then ORF lengths in the 5 alternate frames are calculated. Points give lengths of actual ORFs in reference sequences, coloured according to reading frame, with the reference RdRp as +0 frame (red, below). Note that some annotated RdRp coding regions in the database may be fragments of the complete coding sequence.

is sufficient variability in the RdRp sequence that stop codons would be expected to arise unless selected against. For instance, the average pairwise identity of the 11 sequences in the clade that includes *Culex* narnavirus 1 in Figure 9.3 is only 51%. These analyses suggest that the chance occurrence and maintenance of a large rORF is highly unlikely, implying that it may offer some evolutionary advantage, broadly speaking. Given this, it is desirable to speculate how the rORF observed in narnaviruses may have arisen. The virus could have originally existed as a sequence which just coded for the RdRp, with stop codons in the complementary reading frame. It could have evolved by gradually removing the stop codons, and at the same time making other synonym changes in the +0 frame, as the coding sequence lengthens. These mutations could result in progressively longer

rORFs, selected by their capacity to increase the fitness of the virus. The Wenling narna-like virus 7 (Figure 9.1c), which has a single stop codon in the middle of a large rORF, is intriguing as a potential transitional form in which the rORF is either being gained or lost. If it is being gained, the mechanism in the previous paragraph indicates that the rORF would increase in length incrementally. If the rORF is being lost (for instance, if transfer to exploit a new host removes its advantage), we expect one or more stops randomly scattered in a large ORF. Because this latter picture is more consistent with the Wenling narna-like virus 7 sequence, we speculate that this strain is in the process of losing the large rORF feature. With the presently available data, however, we cannot conclusively support this intuition. Filling out the phylogeny with more sequences could allow us to clarify the potential direction of change. At this time there is very limited information as to whether the rORF can increase the fitness of the narnavirus. There may be unusual mechanisms in which uninterrupted rORFs alter the translation or protein processing machinery in the cell to the virus's advantage, even if no protein is made or if the amino acid sequence is not important for its function. For instance, RNA forms of several viruses are thought to be subject to nonsense-mediated decay (LeBlanc and Beemon 2004; Balistreri et al. 2014; Fontaine et al. 2018). The narnavirus rORF may be another strategy to increase viral-RNA stability through antagonism of RNA decay pathways or other methods. Alternatively, if we postulate that the rORF does indeed code for a functional protein, we might speculate on what can be learned from its sequence. A search of the Protein Data Bank (PDB) and NCBI's translated sequence database (NCBI nr) revealed no sequences with significant homology to the translated rORF. Explorations of secondary

structure and other protein features were similarly uninformative, but do not rule out the possibility for a functional protein (see Methods). Perhaps a protein translated from the rORF could enable the narnavirus to evade host-cell defenses, allow for movement of the virus between cells, enhance replication by complexing with the genome and RdRp, be required for replication in a particular host species, interact with additional viral elements, or have any of a number of other functions that have not yet been described for this family of viruses.

Discussion

Motivated by observations of ambigrammatic sequences in narnaviruses, we have shown that an existing ORF can give rise to a large uninterrupted ORF in the reverse complement sequence by synonymous substitutions that preserve the amino-acid sequence of a conserved forward ORF and remove stop codons from the rORF. We find that this mechanism for making ambigrammatic genes only works when the forward and reverse read frames are aligned. These findings are consistent with the observed alignment of overlapping +0 RdRp ORFs and -0 rORFs among many naturally occurring narnavirus sequences. Any function for the narnavirus rORF remains an intriguing mystery, as does the machinery and processes that may be involved in translating complementary strands of the same RNA sequence. To our knowledge, there are no biologically validated examples of overlapping genes in the reverse complement orientation among RNA-only viruses (Belshaw, Pybus, and Rambaut 2007; Rancurel et al. 2009; Sabath, Wagner, and Karlin 2012). While some such overlaps have been predicted (Schlub, Buchmann, and

Holmes 2018), they exhibit neither the overlap length nor the conservation across related strains that is seen among narnaviruses. Indeed, the narnavirus overlap is the longest yet observed among RNA viruses (see, e.g. (Brandes and Linial 2016)). As our observations about the structure of the genetic code are extensible beyond this family, it will be interesting to see whether the explosion in metagenomic sequencing data will reveal more ambigrammatic viruses. By virtue of packaging, replication, and transmission requirements, viral genomes display a myriad of diverse innovations that provoke us to consider what is possible at the extremes of sequence evolution. Here, the ambigrammatic genes found in some narnaviruses are one such innovation, and their existence likely points to new biology that may be equally as fascinating.

Methods

Quantifying the evolutionary space for the companion gene.

In a standard gene, a sequence of N codons allows for $N = 20^N$ different combinations of amino acids. In the case where a new gene overlaps an existing gene which is perfectly conserved, we must confine our choices to the set of amino acids which correspond to synonyms in the original gene. If the codon for position j allows n_j different synonyms, the number of possible amino-acid sequences is

$$\mathcal{N} = n_1 n_2 \cdots n_N.$$

This number grows rapidly with the length of the sequence:

$$\mathcal{N} \approx e^{sN}$$

where s describes the amount of freedom that the polypeptide sequence has (in physics, it would be termed an entropy). For unconstrained evolution we have an entropy per codon equal to

$$s_0 = \ln 20.$$

This is a measure of the range of possible proteins that can be constructed in the original gene. In the cases that we analyze, a new coding sequence in an alternate frame is constrained by the requirement that the amino-acid sequence of the original gene is conserved, so that codons must be chosen from the set of synonyms. The number of possible sequences for the new gene is much smaller, but still grows exponentially with the length of the sequence. Because there are 20 amino acids and 64 codons, there are approximately three possible choices for each codon, so that the entropy of the new gene is (approximately) $s_1 \approx \ln 3$. Let us consider this more precisely for the case where the new gene is evolving in the -0 frame (that is, reverse complement sequence, with codons aligned). In this case the entropy of the new sequence is

$$s_1 = \sum P_i \ln n_i$$

where P_i is the fraction of amino acids of type i , and n_i is the number of distinct amino acids which can arise from the reverse complement of each synonym. For example, if the amino acid of the original gene is Ser, there are six possible synonyms (UC*, AGU, AGC). The complementary codons, (UGA, CGA, AGA, GGA, ACU, GCU) represent, respectively Stop, Arg, Arg, Gly, Tr, Ala, so that for $i = \text{Ser}$, the number of distinct codons, excluding

Stop, is $n_{\text{Ser}} = 4$. The corresponding numbers for all of the amino acids are $n_{\text{Phe}} = 2$, $n_{\text{Leu}} = 4$, $n_{\text{Ile}} = 4$, $n_{\text{Val}} = 4$, $n_{\text{Ser}} = 4$, $n_{\text{Pro}} = 3$, $n_{\text{Thr}} = 4$, $n_{\text{Ala}} = 4$, $n_{\text{Tyr}} = 2$, $n_{\text{His}} = 1$, $n_{\text{Gln}} = 1$, $n_{\text{Asn}} = 2$, $n_{\text{Lys}} = 2$, $n_{\text{Asp}} = 2$, $n_{\text{Glu}} = 2$, $n_{\text{Cys}} = 2$, $n_{\text{Trp}} = 1$, $n_{\text{Arg}} = 4$, $n_{\text{Gly}} = 4$. Using the codon frequencies P_i determined from the *Culex narnavirus 1* sequence, we find $s_1 \approx 1.104$, which is remarkably close to the value $s_1 \approx \ln 3 \approx 1.099$ estimated in the previous paragraph. Because s_1 is significantly smaller than s_0 , the set of possible sequences which can be coded by the new gene is quite restricted. In particular, in a sequence of length N the ratio of the number of possible amino-acid sequences between a standard gene (N_0) and the constrained one (N_1) is

$$\frac{\mathcal{N}_0}{\mathcal{N}_1} \approx e^{(s_0 - s_1)N} \approx 6.63^N.$$

For $N \sim 1000$ (as in narnaviruses), this ratio would be $\sim 10^{820}$. The constraint is eased if we allow changes in amino acids of the original protein as well as synonyms. It is expected that the actual evolution of the virus genome will represent a compromise between conserving the function of the original protein and allowing scope for evolution of the new protein.

Exploration of the possible secondary structure in the rORF.

The translated rORFs of ambigrammatic narnaviruses are predicted to have median α -helical and β -strand contents of 22% and 12% respectively (calculated using JPred4 (Drozdetskiy et al. 2015)). This degree of secondary structure is consistent with a structured (or folded) protein, but a significant presence of secondary structure can also be

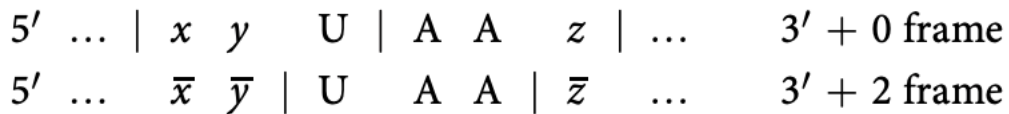
observed in random amino-acid sequences (Tretyachenko et al. 2017). We further note that the isoelectric point (PI) of the RdRp is high (median 10.4, range 7.7–11.6 for sequences >2 kb in Figure 9.3), due to a high frequency of Arg, a basic amino acid (median 9.9%, range 6–13.3%). Notice that this does not necessarily lead to a high concentration of Arg in the rORF (the codons for Arg are CG*, AGA and AGG, none of which leads to an Arg in the –0 frame). The translated rORFs also have high PIs (median 10.2, range 8.3– 11.4), which does not seem to be dictated by the amino-acid composition in the RdRp, and yet is similar to PIs calculated for translations of the –0 frame for non-ambigrammatic narnaviruses (median 10.1, range 6.8–12.7). Basic residues can be involved in binding negatively-charged nucleic acids, but without experimental information we can only speculate on the role for a putative protein translated from the rORF.

Alternative forward reading frames

In the main text we focused on the possibility of ORFs in the reverse strand, because that was the situation relevant for narnaviruses. However, since the arguments presented in this paper are just based on the genetic code and not specific to these viruses, it is worth considering the alternative reading frames in the forward direction as well.

Frame +2.

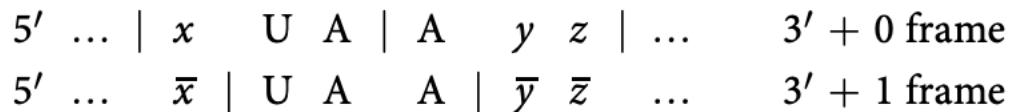
Consider first the case where the new gene is read in the forward direction and left shifted by a single base (frame +2 in the convention of Figure 9.2). For example, if the codons of the +0 sequence are



so that there is a stop codon in the new reading frame. In this case, replacing xyU with xyw , $w \neq U$, removes the stop codon in the new frame. In seven cases (Ala, Arg, Gly, Leu, Pro, Tr, Val) any of the three alternatives for w gives a synonym. For another six cases (Asn, Asp, Cys, His, Phe, Tyr), a synonym substitution is also possible but $w=C$ is the only option. In the case of Ser, UCU allows three changes but AGU allows only AGC. For Ile there are two possible changes and the remaining amino acids can never have U in the final position. Hence, a synonym for the first codon that removes the stop can always be found. The same reasoning holds true if AA in the second codon is replaced by AG or GA (the other two possible combinations that would yield a stop in the +2 frame). So, in the case of a left shift, synonym substitution is possible.

Frame +1.

In the case of a right shift (frame +1), not all stop codons can be removed by a synonym transformation. For example if the codons in the +0 frame are



then we want to remove the stop codon UAA by a single-base substitution which gives a synonym, for any choice of x . If $x=U$, then we have to find a replacement for the second U or the first A that still codes for Leu. The only possible synonym of Leu that involves

changing the second or third letter is UUG, but this is unsatisfactory because UGA is another stop codon. As discussed for the -2 frame, it is only possible to obtain a synonym by changing the A in the second codon if the amino acid is Arg. The same reasoning holds if we consider the UGA stop codon and in the case of the UAG stop codon it is impossible to change the G and obtain a synonym for the second codon in the $+0$ frame. So the Leu codon UUA followed by any codon beginning with either A or G, except for AGA and AGG (Arg), results in a non-removable stop. With the previous exception, the Leu codon UUG followed by a codon beginning with A results in a stop in the new frame that cannot be removed by a single-nucleotide substitution (although it can be removed by two substitutions, for example $UUG \rightarrow CUG \rightarrow CUC$). We conclude that it is possible for new ORFs, transcribed in the forward direction, to overlap a perfectly conserved gene, using single-nucleotide mutations to eliminate stops. If we exclude cases requiring more than one substitution, this can only happen in the $+2$ frame.

Data availability

All the sequences analyzed in this paper are available in online public repositories

Acknowledgements

The authors thank Yun S. Song and John Pak for illuminating discussions on yeast genomes and on rORF protein coding potential, respectively, and Timothy Schlub and Edward Holmes for generously sharing the code used to produce Figure 1. This work and JDR, GH, AK and DY were supported by the Chan Zuckerberg Biohub. HR acknowledges

support from the UCSF Medical Scientist Training Program. MW thanks the Chan Zuckerberg Biohub for its hospitality.

Author contributions

M.W., D.Y. and G.H. conceived of the proposal and performed reading frame analysis. H.R. analysed sequencing data that stimulated this investigation, and generated the figures. J.D.R. and A.K. provided critical input on viral biology. H.R., M.W. and D.Y. wrote the manuscript with input from all authors.

References for Chapter 9

- Akopyants, N.S., Lye, L.-F., Dobson, D.E., Lukeš, J., and Beverley, S.M. (2016). A Narnavirus in the Trypanosomatid Protist Plant Pathogen *Phytomonas serpens*. *Genome Announc.* 4, e00711-16, /ga/4/4/e00711-16.atom.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215, 403–410.
- Balistreri, G., Horvath, P., Schweingruber, C., Zünd, D., McInerney, G., Merits, A., Mühlemann, O., Azzalin, C., and Helenius, A. (2014). The Host Nonsense-Mediated mRNA Decay Pathway Restricts Mammalian RNA Virus Replication. *Cell Host & Microbe* 16, 403–411.
- Belshaw, R., Pybus, O.G., and Rambaut, A. (2007). The evolution of genome compression and genomic novelty in RNA viruses. *Genome Research* 17, 1496–1504.
- Brandes, N., and Linial, M. (2016). Gene overlapping and size constraints in the viral world. *Biol Direct* 11, 26.
- Cai, G., Myers, K., Fry, W.E., and Hillman, B.I. (2012). A member of the virus family Narnaviridae from the plant pathogenic oomycete *Phytophthora infestans*. *Arch Virol* 157, 165–169.
- Chandler, J.A., Liu, R.M., and Bennett, S.N. (2015). RNA shotgun metagenomic sequencing of northern California (USA) mosquitoes uncovers viruses, bacteria, and fungi. *Front. Microbiol.* 06.

- Charon, J., Grigg, M.J., Eden, J.-S., Piera, K.A., William, T., Rose, K., Davenport, M.P., Anstey, N.M., and Holmes, E.C. (2019). Matryoshka RNA virus 1: a novel RNA virus associated with Plasmodium parasites in human malaria (Evolutionary Biology).
- Chirico, N., Vianelli, A., and Belshaw, R. (2010). Why genes overlap in viruses. *Proc. R. Soc. B* 277, 3809–3817.
- Cook, S., Chung, B.Y.-W., Bass, D., Moureau, G., Tang, S., McAlister, E., Culverwell, C.L., Glücksman, E., Wang, H., Brown, T.D.K., et al. (2013). Novel Virus Discovery and Genome Reconstruction from Field RNA Samples Reveals Highly Divergent Viruses in Dipteran Hosts. *PLoS ONE* 8, e80720.
- Dinan, A.M., Lukhovitskaya, N.I., Olendraite, I., and Firth, A.E. (2019). A case for a reverse-frame coding sequence in a group of positive-sense RNA viruses (Genomics).
- Drozdetskiy, A., Cole, C., Procter, J., and Barton, G.J. (2015). JPred4: a protein secondary structure prediction server. *Nucleic Acids Res* 43, W389–W394.
- Fontaine, K.A., Leon, K.E., Khalid, M.M., Tomar, S., Jimenez-Morales, D., Dunlap, M., Kaye, J.A., Shah, P.S., Finkbeiner, S., Krogan, N.J., et al. (2018). The Cellular NMD Pathway Restricts Zika Virus Infection and Is Targeted by the Viral Capsid Protein. *MBio* 9, e02126-18, /mbio/9/6/mBio.02126-18.atom.
- Göertz, G., Miesen, P., Overheul, G., van Rij, R., van Oers, M., and Pijlman, G. (2019). Mosquito Small RNA Responses to West Nile and Insect-Specific Virus Infections in *Aedes* and *Culex* Mosquito Cells. *Viruses* 11, 271.

- Grybchuk, D., Akopyants, N.S., Kostygov, A.Y., Konovalovas, A., Lye, L.-F., Dobson, D.E., Zangger, H., Fasel, N., Butenko, A., Frolov, A.O., et al. (2018). Viral discovery and diversity in trypanosomatid protozoa with a focus on relatives of the human parasite *Leishmania*. *Proc Natl Acad Sci USA* 115, E506–E515.
- Harvey, E., Rose, K., Eden, J.-S., Lawrence, A., Doggett, S.L., and Holmes, E.C. (2019). Identification of diverse arthropod associated viruses in native Australian fleas. *Virology* 535, 189–199.
- Hillman, B.I., and Cai, G. (2013). The Family *Narnaviridae*. In *Advances in Virus Research*, (Elsevier), pp. 149–176.
- Hofstadter, D.R. (1985). *Metamagical themas: questing for the essence of mind and pattern* (New York: Basic Books).
- Kadowaki, K., and Halvorson, H.O. (1971). Appearance of a new species of ribonucleic acid during sporulation in *Saccharomyces cerevisiae*. *J. Bacteriol.* 105, 826–830.
- Krakauer, D.C. (2002). Evolutionary Principles of Genomic Compression. *Comments on Theoretical Biology* 7, 215–236.
- Krakauer, D.C., and Plotkin, J.B. (2002). Redundancy, antiredundancy, and the robustness of genomes. *Proceedings of the National Academy of Sciences* 99, 1405–1409.
- LeBlanc, J.J., and Beemon, K.L. (2004). Unspliced Rous Sarcoma Virus Genomic RNAs Are Translated and Subjected to Nonsense-Mediated mRNA Decay before Packaging. *JVI* 78, 5139–5146.
- Lèbre, S., and Gascuel, O. (2017). The combinatorics of overlapping genes. *Journal of Theoretical Biology* 415, 90–101.

- Lye, L.-F., Akopyants, N.S., Dobson, D.E., and Beverley, S.M. (2016). A Narnavirus -Like Element from the Trypanosomatid Protozoan Parasite *Leptomonas seymouri*. *Genome Announc.* 4, e00713-16, /ga/4/4/e00713-16.atom.
- Merino, E., Balbás, P., Puente, J.L., and Bolívar, F. (1994). Antisense overlapping open reading frames in genes from bacteria to humans. *Nucl Acids Res* 22, 1903–1908.
- Osaki, H., Sasaki, A., Nomiyama, K., and Tomioka, K. (2016). Multiple virus infection in a single strain of *Fusarium poae* shown by deep sequencing. *Virus Genes* 52, 835–847.
- Rancurel, C., Khosravi, M., Dunker, A.K., Romero, P.R., and Karlin, D. (2009). Overlapping Genes Produce Proteins with Unusual Sequence Properties and Offer Insight into De Novo Protein Creation. *JVI* 83, 10719–10736.
- Sabath, N., Wagner, A., and Karlin, D. (2012). Evolution of Viral Proteins Originated De Novo by Overprinting. *Molecular Biology and Evolution* 29, 3767–3780.
- Schlub, T.E., Buchmann, J.P., and Holmes, E.C. (2018). A Simple Method to Detect Candidate Overlapping Genes in Viruses Using Single Genome Sequences. *Molecular Biology and Evolution* 35, 2572–2581.
- Shi, M., Neville, P., Nicholson, J., Eden, J.-S., Imrie, A., and Holmes, E.C. (2017). High-Resolution Metatranscriptomics Reveals the Ecological Dynamics of Mosquito-Associated RNA Viruses in Western Australia. *J Virol* 91, e00680-17, e00680-17.
- Shukla, A. (2015). Analysis of overlapping reading frames in viral genomes. Ph.D. University of Lübeck.

- Smith, T.F., and Waterman, M.S. (1980). Protein constraints induced by multiframe encoding. *Mathematical Biosciences* 49, 17–26.
- Staden, R. (1984). Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. *Nucl Acids Res* 12, 551–567.
- Tretyachenko, V., Vymětal, J., Bednářová, L., Kopecký, V., Hofbauerová, K., Jindrová, H., Hubálek, M., Souček, R., Konvalinka, J., Vondrášek, J., et al. (2017). Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Sci Rep* 7, 15449.
- Waldron, F.M., Stone, G.N., and Obbard, D.J. (2018). Metagenomic sequencing suggests a diversity of RNA interference-like responses to viruses across multicellular eukaryotes. *PLoS Genet* 14, e1007533.
- Wesolowski, M., and Wickner, R.B. (1984). Two new double-stranded RNA molecules showing non-mendelian inheritance and heat inducibility in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 4, 181–187.

Chapter 10 Functional relevance of the narnavirus' unique ambigrammatic genome

Contributions

Includes contributions from Katerina Popova (assistance with design and performance of ribosome profiling assays), Sara Sunshine (library preparation for ribosome profiling), Aaron Brault (CT and Hsu cell lines), Lori Kohlstaedt (mass spectrometry), Amy Kistler (guidance on experimental design), Joe DeRisi (guidance on experimental design and data interpretation).

Narnaviruses are simple RNA viruses with a surprising genomic feature: the open reading frame (ORF) encoding the viral RNA-dependent RNA polymerase (RdRp) is overlapped for nearly the entire length by an ORF on the reverse complementary strand of RNA. This negative strand RNA is typically thought of as simply a template for replication of the genome and production of translatable RdRp-encoding RNAs. However, previous analyses have shown that this feature is highly unlikely to have arisen or been maintained by chance (Dinan et al. 2020; DeRisi et al. 2019). In recent work we extended these analyses for the *Culex narnavirus 1* member of this family, showing that mosquitoes are the likely host and identifying a second segment (“Robin”) that was always found as a side-kick to the RdRp in wild-caught mosquitoes in the study (Chapter 8, Figure 8.15, Figure 8.19, Figure 8.20). The Robin segment displays the same overlapping ORF (“ambigrammatic”) feature. In addition, we noted that only with the observed frame alignment (codon boundaries aligned), could single nucleotide mutations preserve forward amino acid sequence of the RdRp while eliminating stops in the reverse ORF (Chapter 9).

These computational analyses derived from metagenomic sequencing data leave a major unanswered question: what is the functional role for an ambigrammatic genome? Here, I describe *Culex narnavirus 1* in a cultured cell line, test the hypothesis that Robin depends on RdRp, and explore how ribosomes may interact with the viral RNAs in functional experiments.

In the CT cell line derived from the *Culex tarsalis* mosquito, there are sequences of 4 persistent viruses: *Culex narnavirus 1* (CxNV1), Calbertado virus (CALV, *Flaviviridae*), Phasi Charoen-like virus (PCLV, *Phenuiviridae*), and Flock house virus (FHV, *Nodaviridae*).

Since insect-specific viruses are frequently integrated into the DNA genome as endogenized viral elements (EVEs), I first completed the genome for CxNV1 from RNAseq and showed that it is primarily found in the RNA fraction, although small regions amplified weakly from DNA (Figure 10.1). Further experiments will be required to determine whether these represent integration into the genome (fragmented or complete insertions), or possibly a reverse transcriptase-mediated episomal DNA copy of viral RNA. The latter is known to occur in *Aedes* and *Culex* mosquito cells such that DNA copies of even well-known RNA viruses like West Nile Virus can be found after infection (Tassetto et al. 2019; Rückert et al. 2019). In addition, mass spectrometry on the CT cell lysate identified 2 peptides from the RdRp and 4 peptides from the hypothetical protein encoded by the reverse ORF that overlaps the RdRp, suggesting that both ORFs are translated (Figure 10.1E).

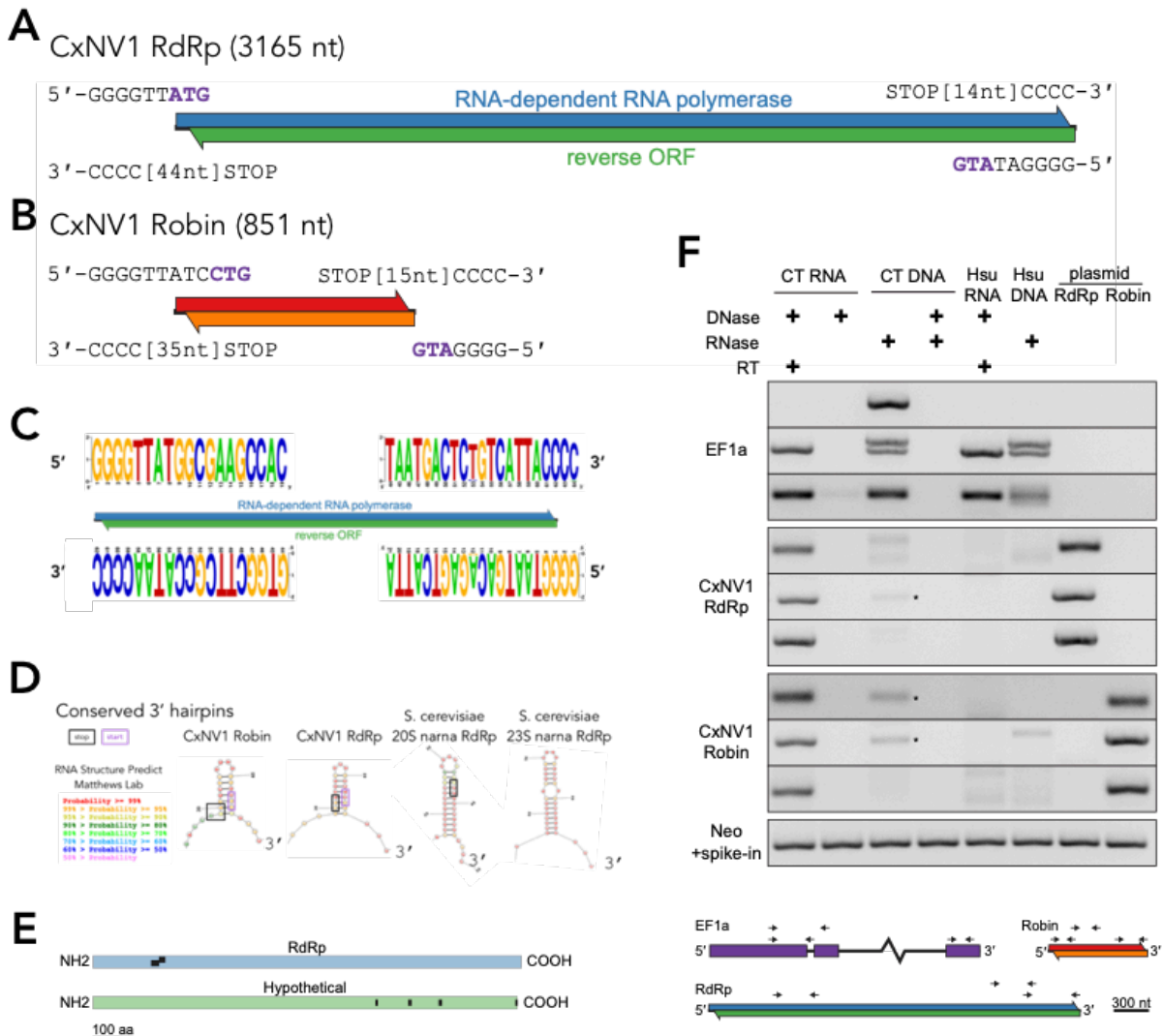


Figure 10.1 Description of Culex naravirus 1 (CxNV1) in Culex tarsalis (CT) cell line.

Genome diagram of consensus sequence by NGS for RNA-dependent RNA polymerase (A), and for Robin segment (B), showing overlapping open reading frames (ORFs) on opposite strands. Ends of RNA segments were verified using an adapter-ligation method (see methods), with consensus shown in (C). D) Using RNA structure predict, the 3' terminal 21 nucleotides form a strong hairpin structure containing the forward ORF stop site and reverse ORF start site. The hairpin structure is conserved across far distant naraviruses such as those of *S. cerevisiae*. E) Alignment of peptides (in black) found by LC-MS/MS of CT cell lysates to the amino acid sequence for the predicted RdRp and Hypothetical proteins encoded by the CxNV1 RdRp RNA segment. F) RT-PCR on RNA and DNA from CT and Hsu cell lines, or on plasmids containing full-length clones of CxNV1 RdRp or Robin, treated with nucleases as indicated, with or without reverse transcriptase (RT) in the reaction. Primer sets for *C. tarsalis* EF1a, CxNV1 RdRp, and CxNV1 Robin are

ordered from top to bottom in according to 5' to 3' position of primers shown in diagram below. Intronic primer specific to *C. tarsalis* in EF1a (top) indicates DNA recovery (absent from *C. quinquefasciatus* Hsu cells). Asterisks indicate faint bands at expected size in CT DNA. For reactions in the lowest row, a plasmid containing the Neo gene was spiked in to the reaction and amplified using Neo-specific primers, to verify that PCR amplification was not inhibited by prior DNase treatment of CT/Hsu input.

Next, I tested the hypothesis that the Robin segment depends on the CxNV1 RdRp for replication/transcription. For these experiments I attempted to launch the virus in a narnavirus-free cell line derived from *Culex quinquefasciatus* mosquitoes: Hsu cells. Full-length viral RNAs were expressed from plasmids with 3' ribozymes, but without selection, driving transient expression of the RNAs and eventual loss of the DNA plasmids. Utilizing a mutant RdRp that has a catalytic site mutation (GDD>RHY), the data suggest that both RdRp and Robin segments depend on RdRp for replication, and can persist in these cells long-term (Figure 10.2A,B). Although I attempted to compare several mutants which eliminated the reverse ORF selectively, the comparable non-synonymous mutants displayed decreased fitness (Figure 10.2C). These mutations identify regions of the RdRp segment that are either important at the RNA level, or in hypothetical protein of the translated reverse ORF. Nevertheless, a small difference was observed in that the RdRp with stops in the reverse ORF was already absent by 2 wks while the non-synonymous mutant control was diminished at 2wks then absent at 6wks. Stops in the reverse ORF of the Robin did not affect the Robin segment. Intriguingly, a Robin segment with stop mutations early in the forward ORF was completely absent at 2 wks, suggesting that either that region of RNA or the primary Robin protein is important for replication (Figure 10.2C).

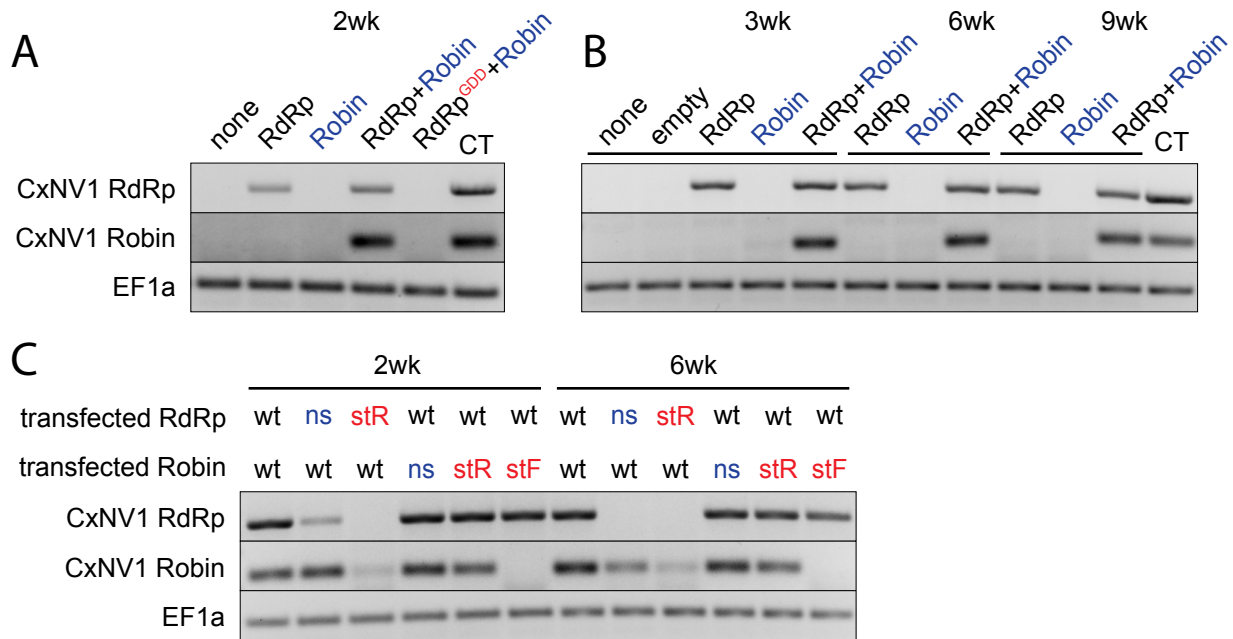


Figure 10.2 Persistence of CxNV1 RdRp and Robin RNA depends on active RdRp.

A) RT-PCR targeting CxNV1 RdRp, Robin, or EF1a RNA at 2 wks post-transfection of Hsu cells with indicated plasmid and sort/counter-sort (see methods), or CT cells as positive control. Plasmids drive expression of full-length viral segments, either wildtype RdRp, active-site mutant RdRp (GDD), and/or wildtype Robin.

B) RT-PCR as in (a), for cells collected at 3, 6, and 9 wks post-transfection.

C) RT-PCR as in (a), at 2 and 6 wks post-transfection of Hsu cells with indicated plasmids: wildtype (wt); mutations introducing stop codons in the reverse ORF while remaining synonymous in the forward ORF (stR); mutations introducing non-synonymous changes in the reverse ORF while remaining synonymous in the forward ORF (nsR), all at the same nucleotide positions as stR; mutations introducing stop codons beginning at 14 codons from the predicted start site of the forward ORF while remaining synonymous in the reverse ORF (stF).

Finally, I set out to determine whether ribosomes were translating both ORFs of each segment, and made an unexpected observation. Using ribosome profiling, I observed a unique pattern of footprints specific to the CxNV1 RNAs. In particular, footprints had a longer length, were clustered in “plateaus” spaced 30-40nt apart, and did not show enrichment of positive strand RNA when compared to the ratio of positive:negative strand RNA in the total RNA libraries (Figure 10.3). There are potential technical explanations for these data, including the possibility of other RNA-binding proteins that are not ribosomes, or RNA structure creating artifacts during library preparation. Nevertheless, I propose a model where ribosomes process haltingly along each strand of each CxNV1 RNA segment, translating protein but also pausing at predictable sites. These sites may be defined by RNA structure, interaction with other proteins, and/or by a block that prevents elongation at the 3' terminus or multiple locations along the RNA, such that colliding ribosomes back up behind the stalled ribosomes in a traffic jam, leaving the observed pattern of footprints. For instance, perhaps the conserved 3' hairpin, which contains both the forward ORF stop and the reverse ORF start codon, is involved in regulating ribosome initiation, translocation, or dissociation. Such a model needs to be tested, by assessing the relative speed of ribosomes on CxNV1 vs other RNAs in the cells, and how mutations to the RNA affect footprints.

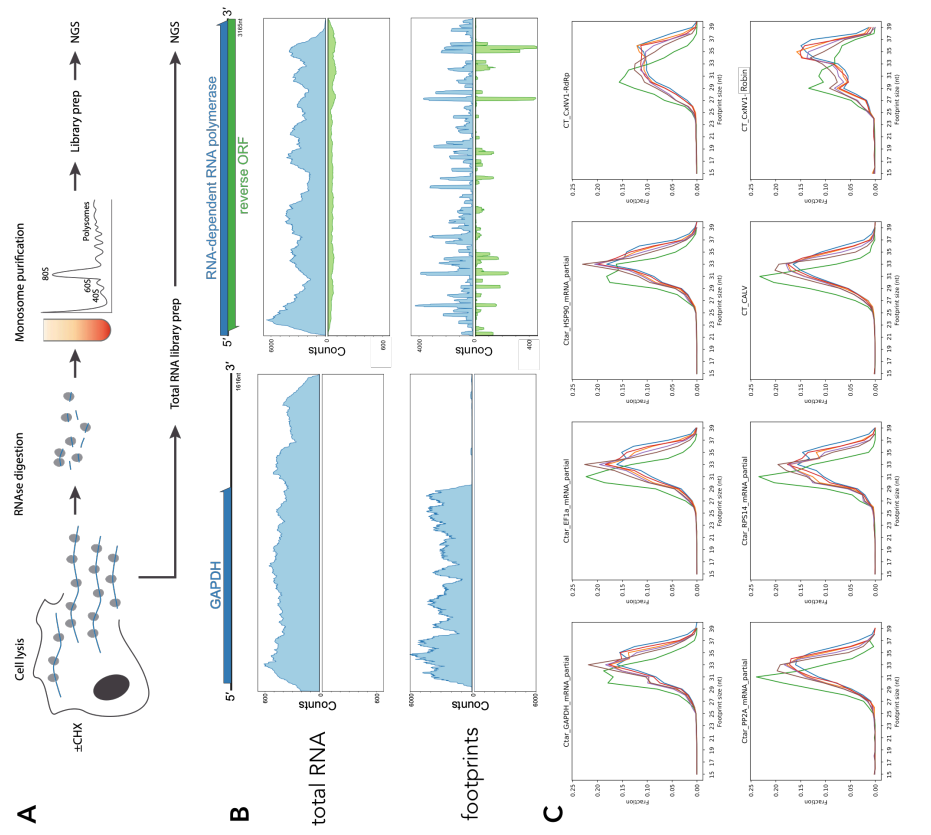
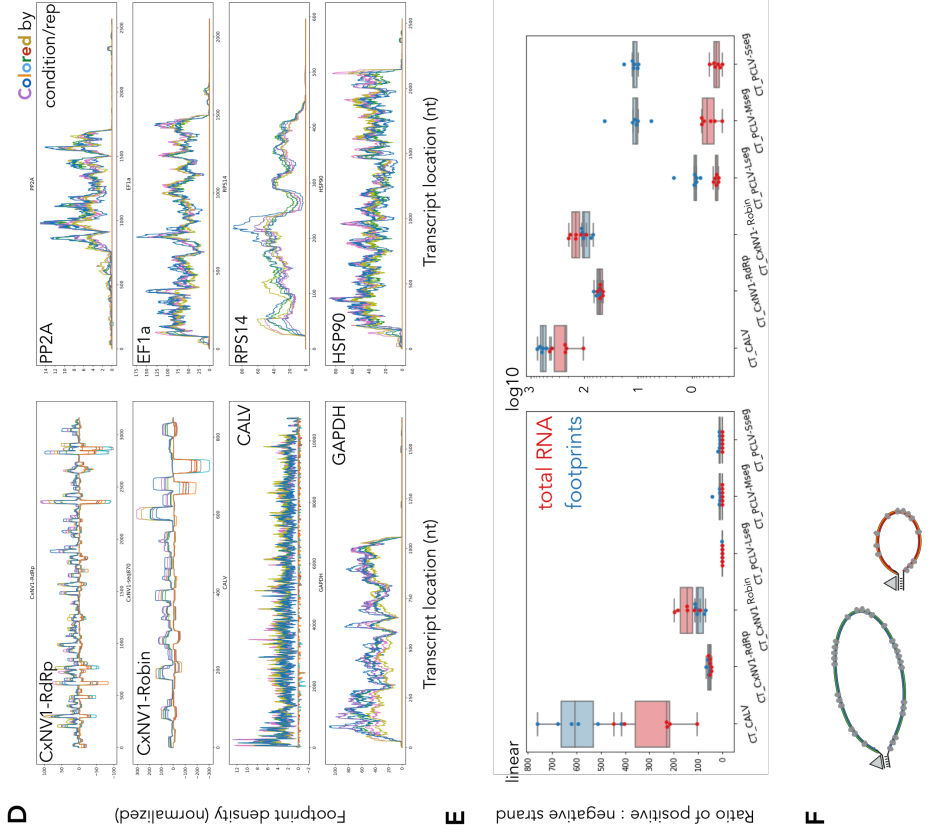


Figure 10.3 Ribosome profiling of CT cells shows a unique pattern on transcripts of CxNV1 RNAs.

A) CT cells were treated with cycloheximide then lysed, underwent RNase digestion with micrococcal nuclease, then centrifuged through a sucrose gradient to purify monosomes by isolating 80S-containing fractions, followed by library preparation. In parallel a library was made of total RNA.

B) Left: total and footprint RNA show expected distribution in the complete transcript or ORF, respectively, of positive (blue) strand of the housekeeping gene GAPDH. Right: total and footprint RNA found across entire length of CxNV1 RdRp segment on both strands (negative strand shown at 10X y-axis scale in green). Footprint coverage shows pile-ups in a pattern resembling plateaus spaced ~30-40nt apart.

C) Across multiple conditions/replicates, size of footprints is comparable for housekeeping genes including GAPDH, EF1a, HSP90, PP2A, RPS14, and a flavivirus persistently found in this cell line, Calbertado virus (CALV), peaking at 33nt. In contrast, the footprint size is larger and possibly bimodal for CxNV1 RdRp and Robin (right top and bottom).

D) Plateau pattern is similar for both segments of CxNV1 (RdRp and Robin), and possibly correlated in location between positive and negative strands (note 10X scale for negative strand). Pile-up location is consistent across conditions/replicates for CxNV1 and housekeeping genes.

E) For other persistent viruses in this cell line, including the positive-sense flavivirus Calbertado virus (CALV), and negative-sense bunyavirus Phasi Charoen-like virus (PLCV), the ratio of positive-to-negative strand is as expected in total RNA vs. in footprints, with a strong enrichment of positive strand RNA in footprints. In contrast, for CxNV1, the footprint-protected RNA is not enriched for positive-strand RNA beyond the total RNA pool.

F) Putative model with the viral polymerase (RdRp, triangle) copying both segments of CxNV1 with the aid of complementary ends to form a panhandle structure. Robin protein may be required for replication of the Robin segment by the RdRp. In this model, ribosomes process haltingly along both positive and negative strand RNA of each segment, translating some protein but also pausing perhaps due to RNA structure, RNA-binding proteins, characteristics of the elongating viral protein, or due to an extending line of collided ribosomes from the 3' hairpin back along the length of the RNA. Instead of immediately being cleared from the RNA, these ribosomes may protect the viral RNA from detection or destruction by the host cell, or may regulate the production of viral proteins to sustain the persistent state of the virus intracellularly.

Let us return to the question of how the ambigrammatic genome might increase this virus' fitness, given the remarkable conservation of this feature. One plausible hypothesis is that the ribosomes coating the viral RNA protect the RNA from detection and/or destruction by the host cell's immune responses. Or, the unique interaction of ribosomes with both strands may be related to the organization of positive, negative, and double-stranded RNA in the cell – while the ratio is consistently 50:1 and 150:1 for positive:negative strand on RdRp and Robin respectively, it is unclear how much remains dsRNA which could be more easily detected as foreign. Alternatively, this feature may regulate the production of proteins, producing levels that are optimized for persistence.

While it is possible that this ambigrammaticity is simply a highly efficient encoding of two proteins in a minimalist genome, it seems far more likely that this narnavirus exhibits a unique feature with interesting regulatory consequences. In the style of Isidor Isaac Rabi's exclamation upon the unexpected discovery of the muon: Who ordered that ORF?

References for Chapter 10

DeRisi, J.L., Huber, G., Kistler, A., Retallack, H., Wilkinson, M., and Yllanes, D. (2019).

An exploration of ambigrammatic sequences in narnaviruses. *Sci Rep* 9, 17982.

Dinan, A.M., Lukhovitskaya, N.I., Olendraite, I., and Firth, A.E. (2020). A case for a

negative-strand coding sequence in a group of positive-sense RNA viruses. *Virus*

Evolution 6.

Rückert, C., Prasad, A.N., Garcia-Luna, S.M., Robison, A., Grubaugh, N.D., Weger-

Lucarelli, J., and Ebel, G.D. (2019). Small RNA responses of *Culex* mosquitoes and

cell lines during acute and persistent virus infection. *Insect Biochem Mol Biol* 109,

13–23.

Tassetto, M., Kunitomi, M., Whitfield, Z.J., Dolan, P.T., Sánchez-Vargas, I., Garcia-Knight,

M., Ribiero, I., Chen, T., Olson, K.E., and Andino, R. (2019). Control of RNA

viruses in mosquito cells through the acquisition of vDNA and endogenous viral

elements. *Elife* 8.

Publishing Agreement

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

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

DocuSigned by:

Hanna Retallack

F0B35E09CE1C48B...

Author Signature

5/27/2020

Date