The Dissertation Committee for Stephanea Sotcheff Certifies that this is the approved version of the following dissertation:

The development and use of cutting edge next generation sequencing methodologies to study RNA viruses

Committee:

Andrew Routh, PhD - Supervisor

Kay Choi, PnD - Chair

Pei-Yong Shi, PhD

Yogesh Wairkar, PhD

Anthony Mustoe, PhD (BCM)

The development and use of cutting edge next generation sequencing methodologies to study RNA viruses

by

Stephanea Sotcheff, BS

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas Medical Branch

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas Medical Branch

April 2022

Acknowledgements

I would like to begin by thanking my mentor, Dr. Andrew Routh, for his support, patience, and encouragement. Andrew always found time to meet with me to discuss new projects and collaborations, my career goals, and opportunities to build skills outside of the laboratory. These conversations were invaluable to me and I am grateful to have completed graduate school working with him as my mentor.

I would also like to thank the rest of my committee members for their attention, support, and guidance regarding completion of my projects and career goals. I appreciated each for the following: Dr. Choi for her keen attention to detail, Dr. Shi for his constant support in my work, Dr. Wairkar for providing perspective, as a faculty member who is not a virologist, and Dr. Mustoe for serving as the external member of this committee and his knowledge in next generation sequencing based technologies.

Additionally, I am grateful for the advice and comradery of my colleagues in the Routh Lab (Dr. Yiyang Zhou, Dr. Rose Langsjoen, Dr. Daniele Swetnam, Dr. Elizabeth Jaworski, Dr. Doreen Lugano, Dr. Nathan Elrod, and Victoria Morris). Noteably, Dr. Langsjoen taught me how to do plaque assays and Dr. Jaworski optimized our library preparation protocols – contributing greatly to my dissertation work.

I would also like to thank the many people in the department who help students navigate the paperwork and course requirements. During my time here that has been JoAlice Whitehurst, Dr. Eric Wagner, Dr. Muge Martinez, and Carmen Duplan. They have made a process that could be very time consuming much more manageable and for that myself and surely many others are grateful.

Last, but certainly not least, I would like to thank my friends and family for their constant support. It truly does take a village, and without mine I could not have accomplished this feat.

The development and use of cutting edge next generation sequencing methodologies to study RNA viruses

Publication No._____

Stephanea Lea Sotcheff, PhD

The University of Texas Medical Branch, 2022

Supervisor: Andrew Routh

With new next generation sequencing technologies and methodologies being published often, our ability to sequence viruses and host transcriptomes (in response to viral infection) has expanded – driving virology forward. Here we show the development and use of novel methodologies to study RNA viruses from their impact on the host, to variations in viral genomes, to how context of an infection may alter disease outcomes. We have published methods such as Tiled Click-Seq (TCS) and poly-A Click-Seq (PAC-Seq) with corresponding pipelines (Virus Recombination Mapper and Differential Poly-A Cluster [*DPAC*]) to study variation in viral genomes and transcriptomic changes respectively. We have also used single nuclei RNA sequencing (snRNA-Seq) for a more granular look at transcriptomic changes using the package Seurat in R.

As the 2015-2016 Zika virus (ZIKV) outbreak in S. America was associated with development of microcephaly in infants born to expectant mothers infected early in pregnancy we wanted to study the transfer of ZIKV from mother to fetus. As this involves placental infection, we extracted total cellular RNA from ZIKV infected (or mock-infected) human placental (JEG3) cells and used it to construct PAC-Seq libraries. Subsequent *DPAC* analysis provided data on differential gene expression, alternative poly-adenylation (APA), and use of alternative terminal exons. We found that up-regulated poly-A sites (PASs) lacked the sequences for canonical poly-adenylation (AAUAAA ~20 nts upstream of the PAS or a GU region just downstream) that were found in down-regulated PASs. Here we present a potential mechanism for the large-scale APA occurring in response to ZIKV infection in JEG3 cells.

Microcephaly, and other CNS issues, can be symptoms of ZIKV infection we wanted to look at the brain as well. As opioid overdose deaths have increased in recently in the U.S. (coinciding with the COVID-19 pandemic) we investigated the potential impact of opioid use on severity of neurological disease from RNA virus infection, looking at expression of genes involved in SARS-CoV-2 or ZIKV infection in the mesolimbic pathway, using both snRNA-Seq and PAC-Seq. Our results suggest that opioid use may exacerbate symptoms of these infections by up-regulating inflammation and down-regulating anti-viral pathways in this brain region.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

ZIKV	Zika virus
DENV	dengue virus
NGS	Next-generation sequencing
PAC-seq	Poly(A)-ClickSeq
DPAC	*Differential Poly-A Cluster* analysis
Flair	*Full length alternative isoform of RNA* analysis
VTA	ventral tegmental area (brain region)
NAc	nucleus accumbens (brain region)
PAC-seq	Poly(A)-ClickSeq
ViReMa	Virus Recombination Mapper
CoVaMa	Co-Variation Mapper
ONT	Oxford Nanopore Technologies
COVID-19	Coronavirus disease 2019
OUD	Opioid use disorder

THE DEVELOPMENT AND USE OF CUTTING EDGE NGS METHODOLOGIES TO STUDY RNA VIRUSES

Chapter 1 Introduction

In this chapter some sections are directly lifted from my published works. The proper citation for these works is provided below. This includes two review articles, one of which I am the first author. On this review and research article I did the majority of writing and had help with revisions from Dr. Andrew Routh. For the research article I also extracted RNA, prepared RNA libraries, did all of the computational analysis, and prepared all of the figures. In the second review article I am second author, I contributed by writing a section and generating figures. These citations are also included in the references section.

Sotcheff, S.; Routh, A. Understanding Flavivirus Capsid Protein Functions: The Tip of the Iceberg. Pathogens. 2020 Jan, 9, 42. https://doi.org/10.3390/pathogens9010042

Sotcheff, S.; Elrod, N.; Chen, J.; Cao, J.; Kuymuycu-Martinez, M.; Shi, P-Y.; Routh, A. Zika virus infection alters gene expression and poly-adenylation patterns in placental cells. (*in Preparation*)

Zhou, Y.; **Sotcheff, S**.; Routh, A. Next Generation Sequencing: a new approach to understanding viral RNA-protein interactions. *(submitted to JBC 11 Nov 2021, in Press)*

Flaviviruses are arthropod-borne (arbo-) viruses that plague both tropic and sub-tropic regions. These viruses belong to *Flaviviridae* and genus *Flavivirus*. There are slightly over 70

species of flaviviruses that have been discovered so far (*1*). Of these, roughly half are mosquito-borne, including the heavily studied: yellow fever (YFV), West Nile (WNV), dengue (DENV), Japanese encephalitis (JEV), and Zika (ZIKV) viruses (*2*). The other half are tick-borne including tick-borne encephalitis virus (TBEV). In this dissertation we focus on ZIKV and DENV.

GLOBAL HEALTH RISK OF FLAVIVIRUSES

Flaviviruses pose an increasingly serious global health risk due to geographic expansion of mosquito vectors (*3*-*5*). Only a small subset of infections results in symptoms, these range from mild fever to hemorrhagic fever or encephalitis to potentially death. There are only a few Food and Drug Administration (FDA) approved vaccines for humans currently available for a few mosquito-borne flaviviruses (namely YFV, JEV, and DENV). However, the DENV vaccine has shown limited efficacy against all DENV serotypes (*6, 7*) and resulted in injury to children in the Philippines, causing safety concerns (*8*). Even with developments in the design of safe and efficacious vaccines, there are unfortunately no antiviral treatments clinically available for infected individuals. As the features of the life cycle appear to be conserved across flaviviruses, there has been much work done to identify pan-flaviviral antiviral therapeutics and to engineer drugs to halt pathogenesis at various stages (*1*). Thus, it is important to work towards a greater understanding of the molecular mechanisms throughout the viral life cycle.

Zika virus (ZIKV)

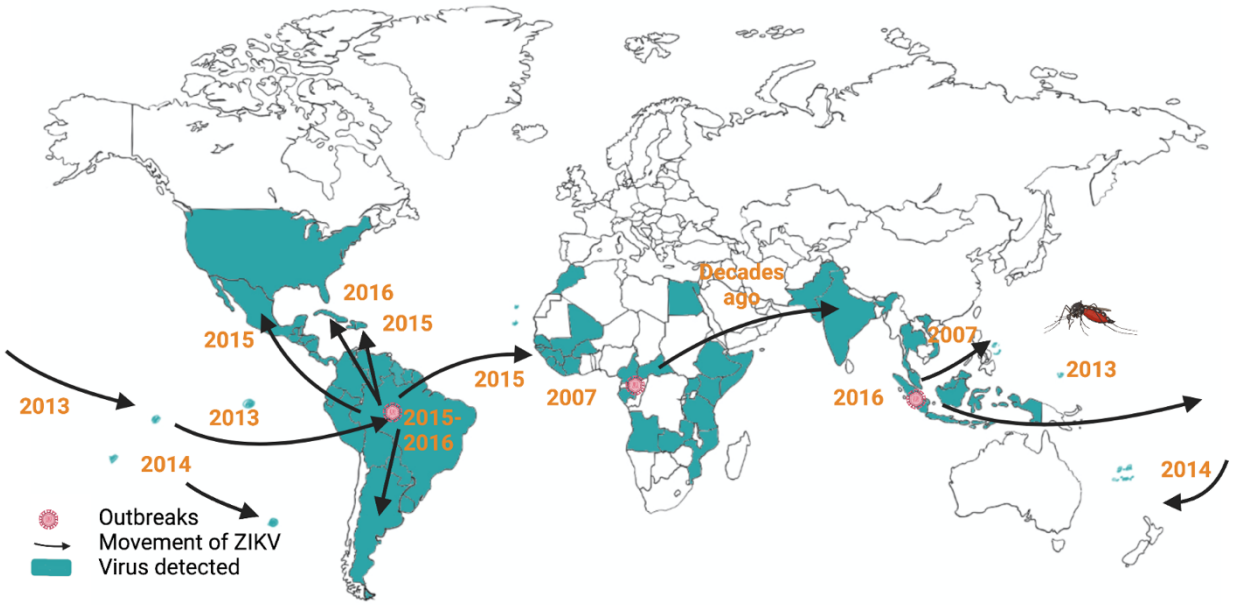
Zika virus (ZIKV) was originally isolated from rhesus monkeys in the Zika forest of Uganda in 1940s. Cases were only reported in Africa until about twenty years ago when ZIKV

began spreading east, first to southeast Asia, then to the Americas in the early 2010s (**Figure 1.1**) (*9*). In 2015–2016 there was an outbreak of ZIKV in South America, which coincided with the Olympic games held in Brazil. Prior to this, symptoms from ZIKV infection were quite mild. However, this outbreak (responsible for 80K+ infections) was associated with microcephaly in infants born to women infected via mosquito bite early in pregnancy and the development of Guillain-Barre syndrome in adults (*2, 10*). Since then, it has also been made apparent that ZIKV persistently infects the testes of infected males and can be spread sexually for months after initial infection (*9*). ZIKV is the first flavivirus with a known arthropod vector to be able to spread sexually (*11*). Although the total number of cases is not exceedingly large for this particular outbreak, it is important to consider the novel means of transmission when considering the global public health risk of ZIKV. At present there are no antiviral treatments available to treat and no FDA approved vaccines to protect against ZIKV infection. There are however, a number of vaccine candidates (*12*).

Dengue virus (DENV)

In 1943 Ren Kimura and Susumu Hotta first isolated dengue virus (DENV) from patients who contracted the virus during an epidemic in Nagasaki, Japan that year (*13*). A year later two additional scientists, Albert Sabin and Walter Schlesinger, independently isolated DENV(*14*). Both groups had isolated what is now known as dengue virus 1 (DEN-1), however there are four serotypes (including DEN-2, DEN-3, and DEN-4). These serotypes share about 65% sequence identity in their genomes, but interact differently with antibodies found in human blood serum (*15*). Like ZIKV, it took decades for DENV to be found outside its region of origin. In the 1970s both DEN-1 and DEN-2 could be found in Africa and Central America, but all four serotypes

Figure 1.1: Distribution of ZIKV infections and outbreaks over the last few decades, originating in Africa and spreading east towards SE Asia and the Americas. Note that geographic expansion of arthropod vector *Aedes* mosquitos plays an additional role. Adapted (with permission) from Weaver *mBio* 2017(*9*). Image created by S. Sotcheff using BioRender under license.

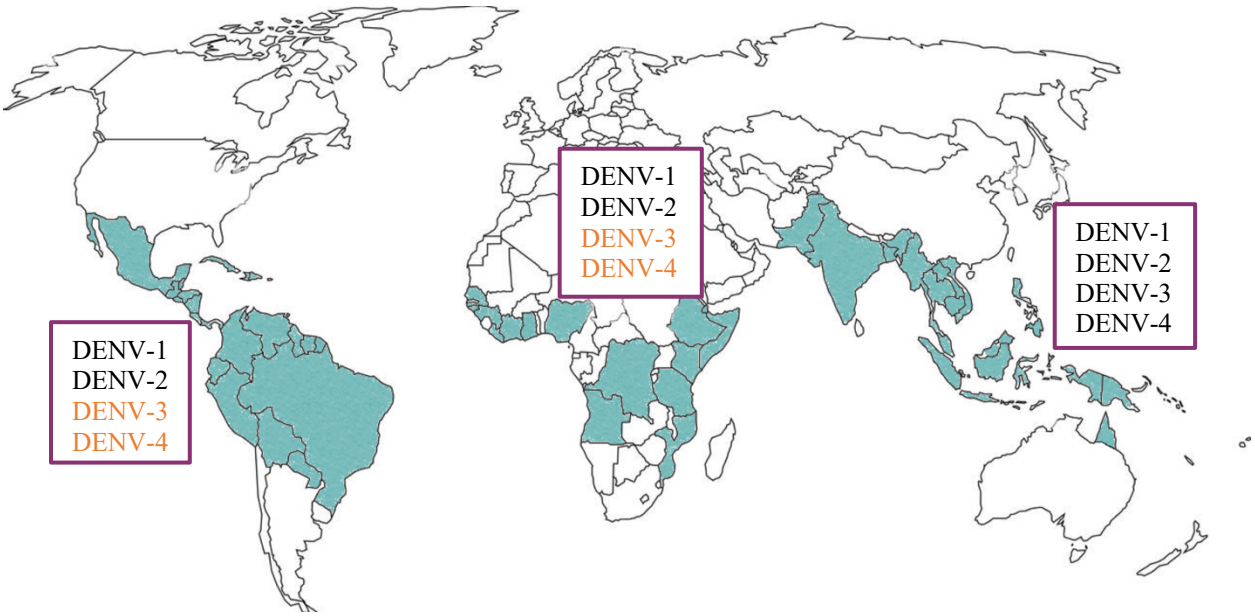


could be found in southeast Asia. By the early 2000s, all four serotypes could be found world-wide in tropic and sub-tropic regions for the most part, with the exception of the Arabian peninsula (**Figure 1.2**) (*16*). There are a documented 400 million cases of the combined DENV serotypes world-wide each year. Roughly 25% of cases are symptomatic with symptoms ranging from febrile illness to hemorrhagic fever and encephalitis. Each year about 40,000 people die from severe DENV infection (*17*). Noted above, there is currently a vaccine for DENV, but its efficacy leaves much to be desired. Dengvaxia® has only shown moderate efficacy (low compared to other commercially available vaccines at 44% across all four serotypes) against serotype DEN-2 (efficacy 64.5%) (*7, 18*). Additionally, the vaccine is only approved for individuals between the ages of 9 and 45 who have already contracted DENV and live in an endemic region. Further, there is additional controversy surrounding this "tetra-valent" vaccine because of studies involving Pilipino children that resulted in vaccine injury (*6, 7, 18*). There has also been reports of this vaccine causing antibody-dependent enhancement (ADE), meaning that individuals who receive the vaccine may experience worse symptoms upon infection compared to those that are not vaccinated (*19*). This is because the antibodies produced by the vaccine bind to virus and make cell entry easier than if the virus were present on its own (*20*). With a lack-luster vaccine, no antiviral treatments available for DENV, and hundreds of millions of cases world-wide each year DENV poses an extremely large global public health risk.

TROPISM OF ZIKV AND DENV

Symptoms of flavivirus infection can range from mild (asymptomatic or febrile illness) to extremely severe including hemorrhagic fever, shock syndrome, encephalitis or hepatitis failure (in the case of Yellow Fever virus or hepatitis C virus) [reviewed in (*21*)] . DENV and ZIKV

Figure 1.2: Global distribution of dengue virus (DENV) has increased over-time. Reported DENV serotype prevalence globally in the 1970s listed in black. Additionally reported DENV serotype prevalence globally in the early 2000s (orange). Image created by S. Sotcheff using BioRender under license.



can both cause either visceral or neurotropic disease. DENV can infect myeloid cells resulting in thrombocytopenia and hypotension (*22-26*). ZIKV also infects myeloid cells as well as epithelium and peripheral tissues such as the eye and reproductive tract. (*27*) In fact, ZIKV can persistently infect the testes (*28*) causing impaired fertility and oligospermia (*29*) and when a high enough viral load is reached ZIKV, can be sexually transmitted. Infection of neuroprogenitor cells by ZIKV or blood-brain barrier cells in severe DENV cases results in damage caused by neuronal death or immune-mediated mechanisms (*30, 31*). Ultimately ZIKV infection of the brain can result in microcephaly as seen in the outbreak in S. America in 2015-2016 (*27, 32, 33*). This outbreak was also associated with increased incidence of development of Guillain-Barré in adults as the CNS was damaged by the immune response to ZIKV infection in these patients (*2, 27*). DENV infections of the nervous system can result in patients experiencing seizures, dizziness, loss of balance, and confusion. In extremely severe cases, individuals may even experience paralysis. Additionally, infection of and virus-mediated damage to the placenta indicates ZIKV is teratogenic (*33-36*). This contributes to congenital Zika syndrome (CZS), although some other viruses have been shown to infect and damage the placenta in pregnant mice (WNV, POWV) (*21*).

FLAVIVIRUSES AND OPIOID USE

Opioids are potent pain relievers typically prescribed to patients after surgeries or other serious procedures in North America, Europe, and Oceania (*37*)**.** However, these drugs must be prescribed sparingly, as they also cause the release of dopamine resulting in a euphoric high. Incidentally, the United States has the highest rates of patients developing opioid use disorder (OUD) (*38-40*), where an individual may overdose from opioid use, survive, and continue using

to the point of another overdose. Although this number was steadily decreasing prior to the SARS-CoV-2 pandemic (*41*), we have recently seen a rise in opioid overdose deaths in the U.S. and globally since COVID-19 made the news at the end of 2019. Individuals with OUD may be treated with an opioid receptor antagonists (*42, 43*). These drugs, like naltrexone or naloxone, bind to opioid receptors, preventing binding of opioids and thus mitigating the release of dopamine and the taker to experience feelings of euphoria.

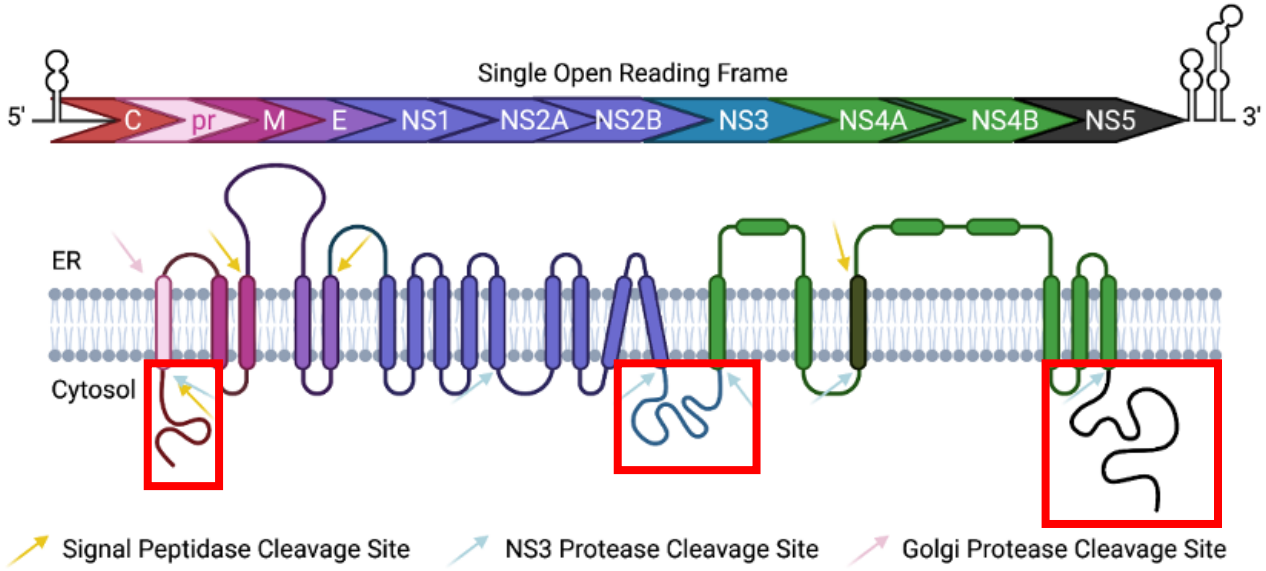
Opioid use has been shown to stimulate the immune response in the brain. In particular, opioid use causes down-regulation of antiviral genes and increases production of IL-6, resulting in local inflammation (*44*). Interestingly, studies have shown that opioid receptor antagonists aid in protection of neuronal tissues from viral infection. For example, SDM25N (a δ-opioid receptor antagonist) appears to inhibit flavivirus NS4 in mammalian (BHK and Hela) but not mosquito (C6/36) cells (*45*). Additionally, naloxone has been shown to prevent induction of IL-1β and increase expression of glutamate transporter GLT-1, preventing paralysis of the hind limbs of mice as a result of infection with murine neuro-adapted Sindbis virus (*46*). This would suggest that opioid usage might exacerbate an individual's response to flavivirus infection but treatment of OUD with an opioid receptor antagonist may prevent neural infection. To date, ZIKV infections that result in congenital Zika syndrome or microcephaly have only occurred in regions with relatively low opioid use. Some of this may be explained by geographic location, that the mosquitos carrying ZIKV are typically not found in Europe of Oceania. However, these mosquitos can be found in Mexico and southern regions of the United States. Therefore, it is interesting to consider if the use of opioid receptor antagonists as treatment of OUD in the U.S. may explain why we have not seen such high incidence of severe neurological disease associated with ZIKV or other flaviviruses in Texas, Southern California, Florida, etc. Alternatively, is it

possible that extreme opioid use, to the point of OUD, offers some amount of protection in the brain from virus? Future studies may be able to tease out the impact of OUD on disease outcomes from viral infections.

GENOME STRUCTURE AND REPLICATION CYCLE

Flaviviruses are enveloped positive sense single-stranded RNA (+ssRNA) viruses that package their ~11 kb genome into individual virus particles that are approximately 50 nm in diameter (*2*). The virions enter the cell via receptor-mediated endocytosis and fuse with the endosomal membrane, releasing viral nucleocapsid into the cytoplasm (*1*). Uncoating is complete when the genome is released from the capsid proteins. The viral genome contains a single open reading frame that must be translated at the endoplasmic reticulum (ER) membrane as a viral polyprotein (*8, 10, 47, 48*) which is cleaved to generate the viral proteins including the RNA-dependent RNA polymerase (RdRP) required for genome replication (**Figure 1.3**). Thereafter, the positive sense genome can be either used to generate a negative-sense template or for translation. Both during and after translation into the ER membrane, the polyprotein is processed to produce 10 viral proteins: three structural (capsid: C, pre-membrane: prM, and envelope: E), as well as seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5) (**Figure 1.3**). This yields three soluble viral proteins in the cytoplasm: C, NS3 (protease/helicase), and NS5 (RdRP with methyl-transferase activity) as indicated by red boxes in **Figure 1.3**. Following synthesis of nascent positive-sense genome in the cytoplasm, the RNA is encapsidated into nucleocapsid particles which begin budding into the ER lumen. These particles traverse through the secretory pathway, undergoing furin-mediated cleavage of prM to

Figure 1.3: Genome structure of flaviviruses consists of a single open reading frame where co- and post-translational cleavage results in 10 viral proteins, 3 structural near the N-terminus and 7 non-structural. Red boxes designate viral proteins that are able to interact with nucleic acids and are cleaved from the ER membrane, able to localize in other cellular compartments. Image created by S. Sotcheff using BioRender under license.

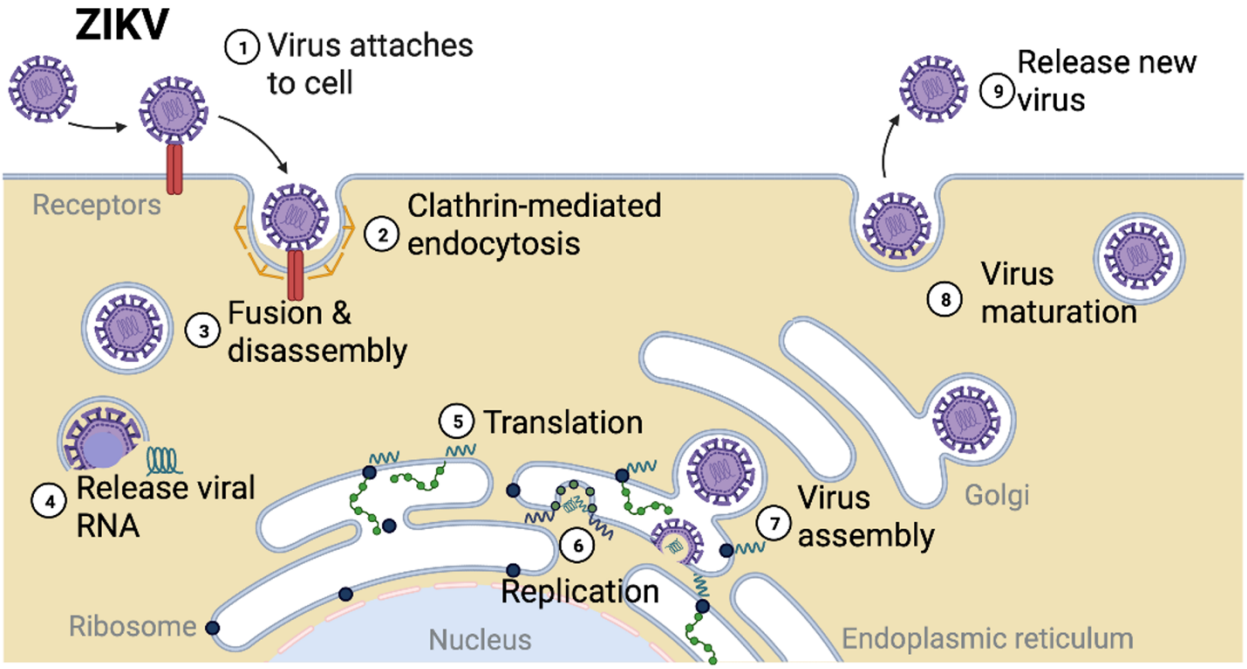


produce mature virus particles that are expelled from the cell via exocytosis (**Figure 1.4)** (*10, 47*).

As translation, replication, and packaging are all distinct processes of the viral life cycle, they are separated spatiotemporally into compartments generated by the rearrangement of the ER membrane (*48*). The rearrangement results in invaginations into the ER, which resemble vesicles with a pore connecting the interior to the cytoplasm. This creates a replication-favorable environment for the virus. These compartments have been viewed using 3D-electron tomography (ET), transmission and scanning electron microscopy (TEM and SEM, respectively) in both mammalian and insect cells. Similar compartments have been seen for alpha- and nodaviruses (*48*). These compartments have been referred to as "replication factories" and are roughly 60 to 90 nm in diameter, depending on cell type (*49*). Within flaviviral replication factories there are three viral proteins known to interact with the primarily double-stranded viral RNA: NS3, NS5, and C. The NS3 helicase separates nascent (+) strand from template (−) strand starting at the 3′ end (*50*) and NS5 binds the 5′ UTR of the (+) sense viral genome and translocates to the 3′ end upon cyclization of the RNA to begin genome replication (*51, 52*). Note that translation and replication (processes that both occur on the + strand but in opposite directions) cannot be occurring on a single strand at the same time. Studies have shown that microRNA 122 functions as a switch for Hepatitis C virus between translation and genome replication by changing the affinity of Poly-C binding protein 2 (PCBP2) for the + strand and re-configuring the structure of the RNA producing an internal ribosomal entry site (IRES) (*53, 54*).

Although capsid protein's primary function is to interact with and protect the genome within virus particles, all of the structural proteins are necessary for the formation of virus

Figure 1.4: Replication cycle of flaviviruses. Virus enters cell through clathrin-mediated endocytosis after interacting with receptor on cell surface. Fusion with the endosomal membrane releases the positive sense genome into the cytoplasm. Translation occurs at the ER membrane and particles begin to bud into the ER. They traverse golgi via the secretory pathway and furin cleavage matures particles prior to their release. Image created by S. Sotcheff using BioRender under license.



particles. Of these proteins, (pr)M and E are integral membrane proteins, with E protruding on the particle surface. This protein is the primary antigen associated with recognition by neutralizing antibodies (*55, 56*). However, specific mutations in NS2A appear to hinder packaging (*57, 58*). It has recently been shown that DENV and ZIKV NS2A recruits the viral genome by binding specifically to the highly-structured 3′ UTR and the C-prM-E complex and protease to site of virion assembly coordinating capsid loading and subsequent virion assembly (*59, 60*).

NEXT GENERATION SEQUENCING (NGS) APPLICATIONS FOR VIROLOGY

There are a number of questions about viruses and their pathogenesis that next generation sequencing (NGS) can help answer. Since the birth of NGS, researchers have been developing exciting new methodologies to investigate evolution of viral genomes, the impact a viral infection has on the host transcriptome, and how protein-RNA interactions make the replication cycle and pathogenesis possible. In this section we will provide a brief overview of NGS methods to study viruses in these contexts.

Sequencing Viral Genomes

Advances in NGS technologies have allowed for the reliable detection of virus and identification of variants or recombination events in viral genomes from both live attenuated vaccines and patient samples. Here we will attempt to summarize the use of NGS to sequence RNA virus genomes. Typically, these methods involve some selection for viral RNA, as viral RNA would be a small portion of total cellular RNA. This can require complete purification of virus using centrifugation, but can be as simple as using RNA from supernatant of infected cells. For

example, Flock House virus (FHV, alphanodavirus) has been purified by ultracentrifugation and sequenced using ClickSeq, a variation of RNAseq that uses azido-nucleotides to stochastically terminate the reverse transcription reaction (*61*). These were sequenced on an Illumina platform and analyzed using various pipelines. Traditional RNAseq has been used to determine the diversity of a live attenuated ZIKV vaccine candidate with paired-end reads on an Illumina platform (*62*). Recently, Tiled-ClickSeq (*63*) and ARTIC (*64, 65*) have made it possible to synthesize cDNA libraries of SARS-CoV-2 from patient samples using sequence specific primers in the reverse transcription step. These are all methods that utilize an Illumina platform which is great for getting a high number of reads and therefore great coverage of the whole genome. However, Illumina reads are short (75-150 nt in length) and therefore much overlap is needed to generate a consensus sequence for a particular sample. Another option may be to pull down viral genomes using anti-sense oligonucleotides with a tag or label. In any case, selection for the viral genome is important to increase genome coverage and minimize the total number of reads necessary.

There are other platforms that allow for sequencing of full-length virus, notably Oxford Nanopore Technologies (ONT). ONT produces a number of sequencers, but the most commonly used is the MinION. This is a hand-held device that can be used in a lab or in the field with the appropriate computer. This is especially important for the sequencing of viruses in regions of the world where storage of samples at the appropriate temperature is difficult (*66*). ONT has been used to sequence full-length DENV from clinical samples with a coverage of about 1000x. In addition to allowing the sequencing of full-length genomes, this also proved to be a more cost effective compared to Illumina sequencing. Compared to Illumina paired-end reads, the ONT reads had a lower quality score, but pairwise similarity of consensus sequences was high. All MinION data were analyzed using Nano-Q, a bioinformatics tool to determine within-host variants (*67*). Another

study used the MinION to detect Ross River virus in mosquito as in-field surveillance and similarly had comparable results to sequencing to an Illumina MiSeq (*68*). The MinION has also been used to sequence DENV and other viruses in patient sera and urine in an effort to find less invasive methods of virus detection. Additionally, pan-flavivirus primers have been used for detection of 7 different flaviviruses in patient samples from Brazil and Vietnam using the ONT MinION (*69*). Indeed, the ability to generate and sequence full-length genome libraries and the size and portability of the ONT MinION (and now also Flongle) make this technology extremely useful in the study of ever-changing viral genomes.

In addition to the methods and platforms used to sequence viral genomes, the pipeline used to analyze the resultant data is crucial. All pipelines involve quality filtering and read trimming followed by alignment to a reference genome. However, parameters of alignment differ when comparing pipelines. When one is simply interested in building a consensus sequence, your pipeline must remove reads aligning to the host genome and compare your base-called data to the reference. Single nucleotide variants (SNVs) or point mutations are one form of viral evolution, but not the only form. Pipelines have been developed to detect other means of viral evolution, in particular recombination events. For example, *DI-tector* is a pipeline used to detect defective interfering (DI) genomes produced by deletion events (*70*). DI genomes have been shown to increase in serial passaging, but DI genomes have also been found in some clinical samples, and are thought to contribute to attenuation or persistent infection. Another pipeline, *VERSE* (virus integration through iterative reference sequence customization), is useful to detect virus integration into host genomes (*71*). There is also a pipeline for detecting copy-back RNAs called viral opensource DVG key algorithm or *VODKA* (*72*). Of note are Virus Recombination Mapper and Co-Variation Mapper (*ViReMa* and *CoVaMa*, respectively) (*73-75*). These detect deletions,

duplications, SNVs, insertions, and virus:host recombination events. Additionally, *CoVaMa* can be used to determine if any of these events coincide with one another. These methods have been used to study recombination and events that coincide with one another (or are mutually exclusive) in HIV, FHV, Sendai virus, and others. We will describe recombination events in these viruses in detail in Chapter 5.

Sequencing Host Transcripts

NGS has also allowed for in depth sequencing of total cellular RNA, including host transcripts during viral infection. The depth of sequencing allows for investigation of differential gene expression and post-transcriptional modifications to the host transcriptome. Here, we briefly describe how ZIKV and DENV infection results in changes to the host transcriptome in various cell types.

ZIKV infection has been shown to alter the host transcriptome in various cell types, in both mammals and mosquitos (*76-82*). The findings of many such studies suggest that lipid and ceramide metabolism and innate immune response are up-regulated following infection, especially in cells such as HEK-293. Interestingly for many cell types it is also seen that the development of embryonic tissues is perturbed. This of course is in line with the fact that ZIKV infection was associated with microcephaly caused by perinatal infections. Thus, it was important to look into infection of the placenta, which should serve as a barrier to infection for the fetus, as well as fetal neuronal tissue which has impeded growth due to ZIKV infection. Previously, a study compared ZIKV-elicited gene expression patterns in human induced pluripotent stem cells (hiPSC) from trophoblasts from dizygotic twins discordant for congenital Zika syndrome (CZS) (*76*). Their findings suggested that although interferon gene expression was not differentially expressed in the

CZS trophoblasts in response to ZIKV infection, there was a significant increase in IFNL1 secretion from the non-affected twins – with no increase observed in trophoblasts from CZS twins. This indicated that the CZS twins' trophoblasts had a lower ability to migrate, recruit immune cells, and control the viral infection. To investigate the impact of ZIKV infection in neuronal cells some groups have done transcriptomics studies in SH-SY5Y (neuroblastoma) cells, human neural stem cells (hNSCs) or human neural progenitor cells (hNPs) (*79, 81*). It has been shown that infection of neuroblastoma cells with the Puerto Rican (PR) strain of ZIKV results in the up-regulation of stress and DNA damage responses as well as regulation of cell migration and down-regulation of metabolic processes, signal transduction and apoptosis (*77*).

The transcriptomic changes are not limited to differential gene expression. Alternative splicing in response to ZIKV infection has been shown by many groups. A study using fragmented poly-A selected RNAs to generate RNAseq libraries for sequencing on an Illumina platform with replicate multivariate analysis of transcript splicing (rMATS) showed that for both PR and an African isolate (MR766) strains of ZIKV in SH-SY5Y cells over 50% of AS events are exon skipping events (*77*). Interestingly, there was little overlap in the AS events, with more occurring in response to PR than to MR. They validated exon skipping of a few genes, namely HNRNPDL and RBM39 which function in transcriptional regulation, and SRSF2 which is involved in splicing, using RT-PCR. Similarly, the MR strain used in that study was used in a study involving human neural progenitor cells (hNPCs) where they produced RNAseq data that was later analyzed to identify 229 AS events, again, mostly exon skipping (45-50%). They validated exon skipping in a number of genes with the RNAseq data, including HNRNPA2B1 which is also involved in splicing (*79*). Such changes appear to be attributed to both NS5 and subgenomic flavivirus RNAs (sfRNAs) produced by ZIKV. Numerous host proteins involved in splicing have been shown to be

sequestered by ZIKV sfRNA, including: SF3B1, PCBP2, and HNRNPK (*83*). Immunoprecipitation of NS5 from both ZIKV and JEV have shown interaction with HNRNPs and splicing factors (*84*).

Similarly, DENV has been shown to alter gene expression patterns in various mammalian and mosquito cell types. DENV-2 infection of neuroblastoma cells resulted in up-regulation of metabolic processes and down-regulation of the inflammatory response and extracellular structure organization (*77*). The up-regulation of metabolic processes was also found in clinical samples from DENV infected patients under 15 years of age that were hospitalized in Nicaragua September 2003 through February 2004 (*85*). For this study, microarrays were used to perform differential gene expression from venous blood samples. In mosquito cells (C6/36), RNAseq data from DENV-2 infection revealed 1133 up- and 106 down-regulated genes (*86*). Enrichment analysis identified up-regulation of various metabolic processes and membrane components and down-regulation of catabolic processes. Another study in human monocytes and trophoblast cells compared DENV-4 infection to other flaviviruses and suggested that DENV elicits a less robust antiviral innate cytokine but stronger interferon response compared to a ZIKV Asian strain (FSS13025) (*87*).

DENV has also been shown to elicit splicing pattern changes in host cells. A study comparing differential gene expression and alternative splicing in neuroblastoma cells between ZIKV and DENV-2 showed that DENV infection resulted in 94 AS events, of which 46% were exon skipping events, and only 32 of the 94 events were unique to DENV2 infection (*77*). RT-PCR was used to validate exon skipping in CHID1, SRSF2, HNRNPDL, and RBM39 which have been shown to be involved in pathogen sensing, splicing, splicing and transcriptional regulation, and transcriptional regulation respectively. In another study, DENV-2 infected HEK-293T (human embryonic kidney

cell line) and A549 (human lung adenocarcinoma epithelial cell line) there was increased inclusion of exon 4 of spermidine/spermine acetyl-transferase 1 (SAT1) upon infection (*88*). SAT1 is a known antiviral effector, and inclusion of exon 4 marks SAT1 transcripts for non-sense mediated decay because it introduces a frameshift. Splicing factor RBM10 is responsible for this AS event, and interacts with viral non-structural protein 5 (NS5). Overexpression of RBM10 rescued normal splicing of SAT1. DENV NS5 has also been shown to interact with CD2BP2 and DDX23, core components of the U5 snRNP (*89*). These interactions may account for some, but not likely all AS events induced by DENV infection.

It is interesting that although these viruses are closely related, they have entirely different effects on gene dysregulation and splicing. Even two strains of ZIKV (PR and MR) and DENV2 infecting SH-SY5Y cells show very little overlap in regards to differentially expressed and alternatively spliced genes (*77*). Although it might be easy to explain the small overlap in ZIKV to DENV comparisons, it is also likely that there are other transcriptomic changes worth investigating that might aid in understanding the pathogenesis of these viruses.

Investigating RNA-protein interactions

Another way that NGS has been used to study viral pathogenesis is its use in new methods developed to identify RNA-protein interactions. This can be approached from two ways: viral proteins and the RNAs they bind or host proteins that interact with the viral genome. Here we will discuss a few examples of both approaches and the methods used to accomplish these tasks.

Crosslinking of protein to RNA can be done chemically using formaldehyde or with UV irradiation. Application of this method to a cell will crosslink all proteins to all RNAs, therefore enrichment of your protein or RNA of interest is necessary. Typically, this enrichment comes from

immunoprecipitation. This method, first described as crosslinking immunoprecipitation sequencing (CLIP-Seq, also called HITS-CLIP), now has many derivatives including photoactivatable ribonucleoside crosslinking immunoprecipitation (PAR-CLIP), enhanced CLIP (eCLIP) and individual nucleotide resolution CLIP (iCLIP). These all involve UV-crosslinking. CLIP combines short wavelength UV irradiation and immunoprecipitation to identify RNA sequences interacting with a selected protein target (*90, 91*). This is possible because the aromatic rings in nucleobases are excited to a higher energetic state to exceed the ionization potential when UV irradiated, generating cation radicals. This can result in the formation of covalent bond with similar radicals in direct vicinity (such as UV-excited aromatic rings or other active side chains in amino acids). HITS-CLIP and PAR-CLIP have been used to pull down Argonaut and find miRNAs for viruses such as Epstein-Barr virus (EBV) (*92*), Kaposi's sarcoma associated herpesvirus (KSHV) (*93, 94*), and investigate the interaction between flavivirus hepatitis C virus (HCV) and miRNA-122 (*95*). These interactions are important for viral pathogenesis, perhaps because the production of miRNAs serve a similar function to DI genomes in persistent infection and attenuation.

Viral proteins typically of interest in the development of antiviral treatments include the viral enzymes as well as the structural proteins. Although viral polymerases are the most common target for antiviral drugs, other enzymes can also serve as valid targets. HIV-1 integrase (IN) is a multi-domain enzyme and one of the cleaved products of Pol polyprotein (*96*). IN mediates the integration of viral DNA into host chromosomes following the production of double-stranded proviral DNA from reverse transcription (*97, 98*). In addition to viral DNA insertion, IN has long been suggested to coordinate viral replication and virion maturation, as mutated IN can lead to the eccentric "exile" of ribonucleoprotein complexes (RNPs) outside of capsid shell, and ultimately

impairs virion maturation (*99-103*). Using PAR-CLIP, Kessl et al. identified IN-binding RNA targets in virions (*104*). IN showed strong binding preference for the trans-activation response (TAR) element (*105*) and RRE, but not packaging element ψ, suggesting IN and nucleocapsid (NC) have both shared unique roles in HIV-1 genome assembly and particle maturation (*104*). Structural proteins must be able to encapsidate and protect the viral genome. HITS-CLIP has been used to determine binding sites for nucleocapsid in influenza A virus (IAV) and influenza B virus (IBV) within virus particles (*106*). Another method that utilizes crosslinking but without immunoprecipitation is virus photoactivatable crosslinking (vPAR-CL) which we describe in detail below (*107*). This has been used to map capsid-vRNA binding in FHV. Gag is a structural protein involved in many facets of the HIV replication cycle that has been used as the target for PAR-CLIP studies. Kutluay et al., used PAR-CLIP to enrich Gag-RNA complexes from both cells and virions to investigate the global Gag-RNA interactome profile during and after Gag-orchestrated genome assembly (*108, 109*). This uncovered a surprising and drastic shift in profiles of Gag-interacting RNA during HIV-1 intracellular virion assembly (*109*). Arguably, understanding the interactions of flavivirus may aid in the design of effective antivirals that can prevent RNA binding.

SOLUBLE FLAVIVIRUS PROTEINS

As highlighted above, there are three viral proteins that are soluble in the host cell once co- and post-translationally cleaved from the flavivirus polyprotein: NS5, NS3, and capsid. Interestingly, each of these proteins have the ability to interact with double-stranded RNA (dsRNA). These features of these viral proteins are likely not coincidental. Here we will describe the structure of each protein, their main role in the replication cycle, and any known alternative

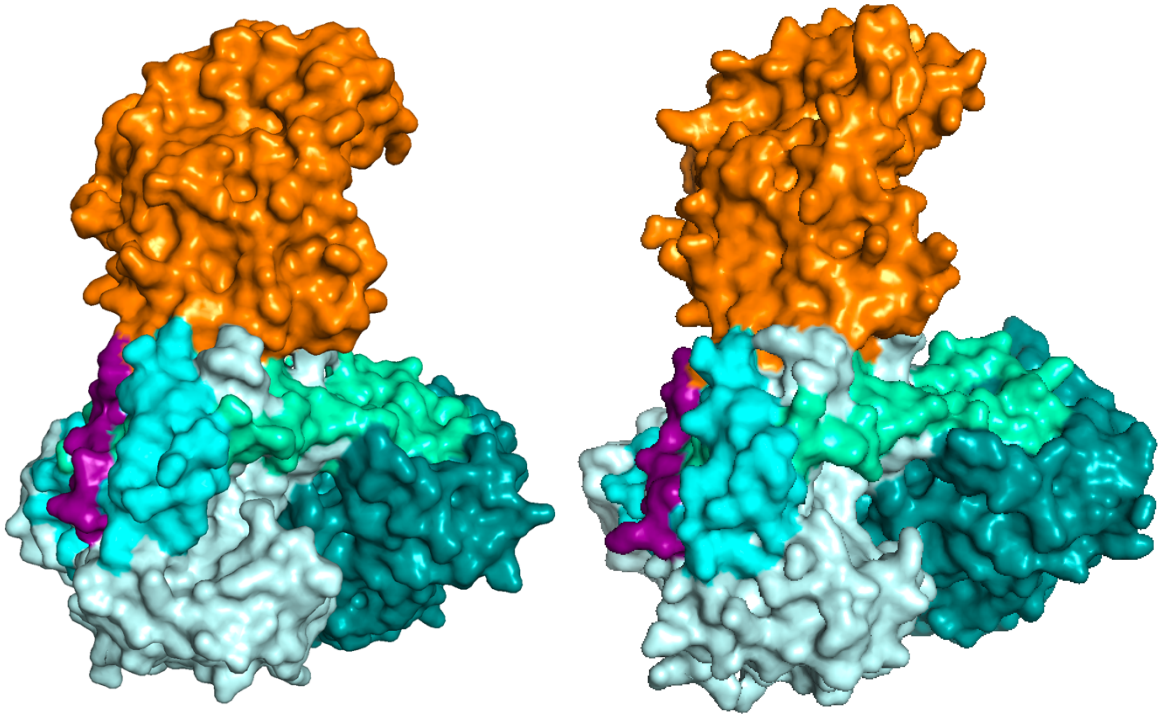
functions. Note that, to-date, both NS5 and NS3 proteins serve as targets for antiviral drugs, but usually capsid in enveloped viruses is not (*110*). This is targeting the enzyme activities of NS5 (polymerase) and NS3 (protease) is less challenging than the structural interactions of C protein.

RNA-dependent RNA polymerase with methyltransferase (NS5)

Flavivirus non-structural protein 5 (NS5) encodes a viral RdRP with a methyl transferase domain (*111-113*). It is the largest non-structural protein produced by flaviviruses, roughly 900 amino acids. The RdRP is not dissimilar to other RNA virus polymerases in that its' structure is referred to as a right hand, with a thumb, palm, and fingers domain, which is linked to the methyl-transferase domain (**Figure 1.5**) (*114*). The conservation of the NS5 protein, and similarity to other RNA virus RdRPs makes this protein a good target for antiviral drug design (*110, 115, 116*). The primary role of flavivirus NS5 is to synthesize viral genome of both positive and negative senses using the opposite sense as a template. Note that negative sense genome only serves as template for the production of positive sense genomes and cannot be translated or packaged. The methyltransferase of NS5 aids in the production of the 5' cap present in the viral genome which allows recognition by host proteins to enhance translation (*113*). As NS5 is not bound to the ER membrane, it can be found in the cytoplasm as well as the nucleus; in fact roughly 20% of flavivirus NS5 may reside in the nucleus.

Although producing viable viral genomes is important, NS5 also acts in other ways to promote pathogenesis and the replication cycle. Flavivirus NS5 from different viruses utilize different strategies to antagonize type interferon I (IFN-I) by preventing JAK-STAT signaling (*117*) (*118*). DENV-2 NS5 binds both STAT2 and UBR4 (ubiquitin protein ligase E3 component N-recognin), resulting in the degradation of STAT2 protein in the cytosol. ZIKV NS5 expression

Figure 1.5: Structure of DENV (5ZQK.pdb, left) and ZIKV (5U0B.pdb, right) NS5 (RdRP) protein. In these side views of the NS5 proteins we show the surface of the protein and color code each relevant domain: methyltransferase (orange), linker region (purple), index "finger" (green), palm (pale blue), and thumb (dark teal). Other regions shown in cyan. Image created by S. Sotcheff using PyMol under license.



alone is enough to result in degradation of STAT2, but using a different E3 ligase than UBR4 which has yet to be identified. However, IP of ZIKV NS5 revealed interactions with ubiquitin ligase CUL4-DDB1 which may have a role to play in preventing JAK-STAT signaling. Interestingly, a close relative of ZIKV, Spondweni virus (SPOV), actually inhibits the signaling pathway downstream of STAT2. SPOV NS5 translocates to the nucleus, where phosphorylated STAT1 and STAT2 are present, and prevents transcription of IFN-stimulated genes (*117*). It is interesting that ZIKV and DENV appear to inhibit the JAK-STAT pathways in similar fashion, even though SPOV is more closely related to ZIKV than DENV is.

An additional role NS5 plays in pathogenesis involves its effect on splicing which is briefly mentioned above. NS5 interactions with splicing factors and the spliceosome certainly alter splicing patterns in host cells (*84, 88, 89*). For example, DENV NS5 interactions with splicing factor RBM10 results in non-sense mediated decay of SAT1, an antiviral host protein, due to inclusion of an exon that introduces a frameshift (*88*). The viral NS5 protein has also been shown to interact with U5 snRNP components – such interactions with the spliceosome might imply broader, or less specific, AS changes. However, studies have shown this is not the case. Based on this, there are certainly more pieces to the puzzle describing NS5's impact on host splicing, and it is unlikely that all AS in response to flavivirus infection can be attributed to NS5 protein.

Viral protease (NS3)

Flavivirus non-structural protein 3 (NS3) is the viral protease with helicase activity (*115, 119*). This is necessary as flavivirus genomes have a single open reading frame that is translated into a viral polyprotein that requires cleavage into its 10 separate viral proteins. Note that not all cleavages are conducted by NS3, which is crucial as NS3 is not packaged into virus particles. In

replication factories in the ER membrane NS3 is found interacting with NS2B to form a multi-functional complex. The N-terminal domain of NS3 interacts with NS2B to form the viral protease and the C-terminal domain of NS3 functions as a helicase that aids in particle assembly and replication. These two domains are connected via a 10 amino acid linker region that does not appear to have conserved sequences across flaviviruses, but rather has flexibility that is conserved in this region (*120, 121*). The full-length NS3 is just over 600 amino acids with the first 169 comprising the protease domain (**Figure 1.6**). Of note, a number of antiviral drugs against NS3 have been FDA approved for the treatment of individuals with hepatitis C virus (HCV) (*122, 123*).

NS3 is not bound to the ER membrane when not in complex with NS2B. In fact, it has been shown that NS3 in both DENV and ZIKV interacts with the nuclear pore complex. This interaction results in the protease cleavage of various nucleoporins that can be prevented by serine protease inhibitors, TLCK and Leupeptin (*122-124*). Additionally, NS3 has been found to co-localize in the nucleus of infected cells with NS5 and newly synthesized viral RNA, suggesting a role in replication outside of the factories at the ER membrane.

Viral capsid (C)

Capsid proteins are the least genetically conserved of flavivirus proteins, but their structure and charge distribution are functionally well conserved. Capsid is a small (~12 kDa), highly positively charged protein comprising the first ~105 residues of the flavivirus polyprotein (**Figure 1.7).** Crystal structures for capsid have been solved for ZIKV, JEV, and WNV (*125-127*). The DENV C structure is similar to that of the other flavivirus capsids, as determined by nuclear magnetic resonance (NMR) (*128*). Each monomer is comprised of three to four alpha-helices, the first being the most flexible, consistently forming a right-handed bundle with the

Figure 1.6: Structure of DENV NS3 (helicase and protease) (2VBC.pdb, left) and the apo form of ZIKV NS3 helicase (6ADW.pdb). In this side view of DENV NS3 the surface of helicase is colored in yellow, the protease is blue and the flexible linker region is green. There are three domains of helicase, showcased in this top view of ZIKV NS3, domain 1 is colored yellow, domain 2 is yellow-orange and domain 3 is orange. Image created by S. Sotcheff using PyMol under license.

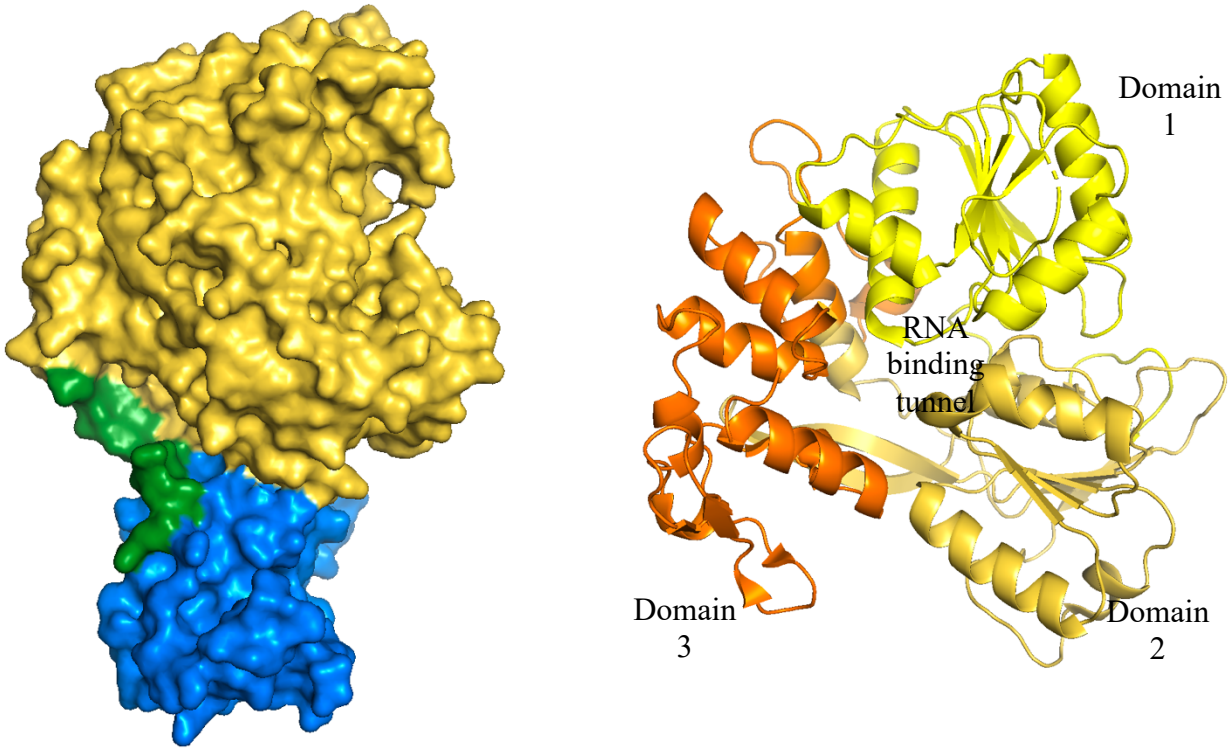


Figure 1.7: Bottom view of C (capsid) protein in homo-dimer for Zika (5YGH.pdb, left) and West Nile virus (1SFK.pdb, right) that the surface is primarily hydrophilic. From Sotcheff & Routh, *Pathogens*, 2020. Image created by S. Sotcheff using PyMol under license.

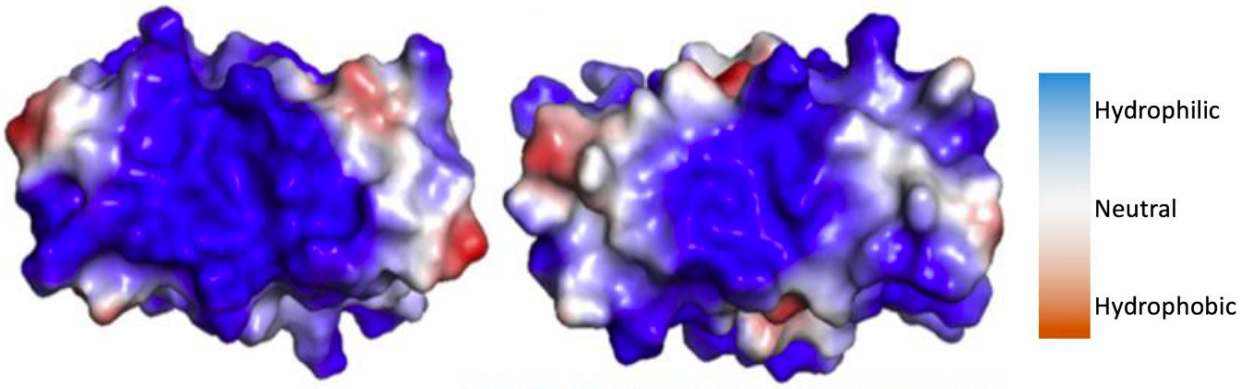
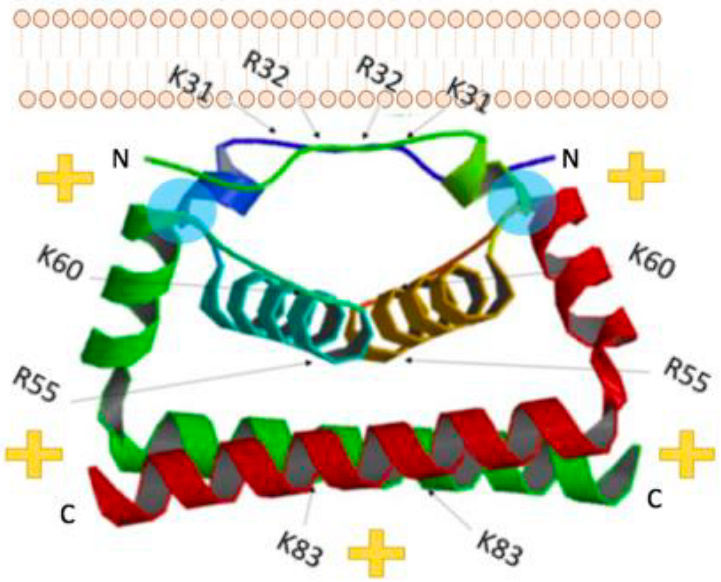


Figure 1.8: Interactions between flavivirus capsid dimers and lipid droplet membranes. Side view of ZIKV C dimer oriented to lipid bi-layer. Blue circles mark location of hydrophobic residues important for this interaction (L50 and L54). From Sotcheff & Routh, *Pathogens*, 2020. Image created by S. Sotcheff using PyMol and Powerpoint under license.



second helix (and third if there are four in total). Within dimers, these helices interact via hydrophobic interactions as illustrated in **Figure 1.8**. As depicted, the top (α1–α1′), bottom (α4–α4′), and core (α2–α2′) helices interact via hydrophobic interactions (*126*). In contrast, the first ~20 residues are intrinsically disordered in solution and, similar to the final helix which extends away from the monomer core, are highly basic (*129-131*). The N-terminus is expected to interact with the viral genome within virus particles. In contrast, the regions connecting α1–α2 and α1′–α2′ are relatively hydrophobic, allowing interactions with lipid bilayers (*131*). In reference to **Figure 1.8**, where this hydrophobic linker-region is labeled by a cyan circle, the charge distribution places the basic residues on the "bottom" of the dimer, leaving the "top" of the dimer relatively uncharged.

Based on the very basic surface of capsid proteins, it is reasonable to consider that capsid binds to the negatively charged phosphate backbone in all nucleic acids. In 2018, Shang et al. (*126*) performed an isothermal titration calorimetry (ITC) assay to measure the binding affinity of ZIKV C to four types of nucleic acid: 5′ UTR of the ZIKV genome (ssRNA), dsRNA, ssDNA, and dsDNA, which may be found in the nucleus. Interestingly their studies indicated that ZIKV C bound all nucleic acids with affinities in the nanomolar range. This high affinity for all nucleic acids lends itself to the dsRNA binding ability which prevents dicer activity in mosquitos, as noted previously (*132*), and others have shown that ZIKV C's ability to bind ssDNA is made possible by the positively charged surface of the protein (*133*). Although the affinity for all nucleic acids is high, the specificity appears to be low. The dissociation constant for DENV C is roughly 20 nM (*134, 135*). Considering the wide range of nucleic acid binding and the varied localization of flavivirus capsid, it is reasonable that there may be additional (potentially transient) nucleic acid interactions for the capsid proteins that have yet to be described.

Binding of viral RNA (vRNA) to capsid initiates particle formation by causing an aggregation of capsid. The aggregation of membrane-associated capsid into the nucleocapsid structure induces budding into the ER and the formation of immature virus particles. Capsid protein has been shown to bind multiple nucleic acid templates in a sequence-independent manner via electrostatic interactions with the negatively charged phosphate backbone (*52, 126, 134*). The coupling of replication and packaging within these ER membrane compartments prevents capsid from packaging host RNAs. Until recently, it was unclear how capsid within these replication factories interacts specifically with the (+) sense viral genome, as there is some amount of (−) sense template available within these compartments. It was proposed that the (−) sense associated with nascent (+) sense intermediate prevents capsid binding (*52, 126*)—however, this has not been definitively demonstrated. Recent studies demonstrating NS2A binding to 3′ UTR of genomic RNA and the subsequent localization to membrane-bound assembly factories suggest that this viral protein may nucleate or 'seed' the loading of capsid, thus providing the specificity in packaging of just the genomic RNA (*60*).

It has also been shown that capsid proteins co-localize with the nucleoli and lipid droplets within infected cells (*126, 131, 136*). Interestingly, the ability of capsid dimers to interact with lipid droplets is essential for efficient production of virus particles (*131, 134, 137, 138*). The result of interactions within the nucleus and nucleolus can result in either the activation or repression of various pathways with deleterious effects, including apoptosis or cell cycle arrest. Capsid has also been shown to enter the nucleus and wreak havoc on ribosome biogenesis and the host transcriptome (*139-144*). The function of these interactions within the nucleus and nucleolus in the context of the flavivirus life cycle is poorly understood. Additionally, various

studies have investigated binding partners of flavivirus capsid proteins which are summarized in **Table 1.1**. Certainly, all these interactions and the distribution of capsid proteins within infected cells, coupled with the indiscriminate binding of nucleic acids, suggests multiple roles for capsid in pathogenesis aside from packaging the viral genome.

A TALE OF TWO PROJECTS

In the studies below we utilize cutting edge next generation sequencing methodologies to study a) the impact of flavivirus (DENV and ZIKV) on gene expression/regulation in human placental cells and b) how opioid use alters gene expression in the mesolimbic pathway (involved in pleasure seeking and learning) of rats and how that may impact severity of disease in the context of a viral infection. This allows us to investigate two of the major tropisms of flavivirus infections: brain and placental tissue as well as take a closer look at how context of an infection could potentially alter clinical outcomes. Methods for the relevant experiments can be found in Chapter 2 (Materials and Methods) and results are included in their respective chapters.

Use Poly(A)-ClickSeq (PAC-seq) to investigate differential gene expression and alternative poly-adenylation in human placental cells as a result of flavivirus infection (Chapter 3)

In this study we are investigating the impact of ZIKV infection on the transcriptome of human placental cells (JEG3) at 16 hours post infection (hpi). This is in comparison to ZIKV infection of human liver cells (Huh7) and DENV infection of JEG3 cells at 16 hpi. Such studies are important to understand both how the infection of ZIKV may impact the fetus but DENV does not, as well as understand differences in infections of different tissues.

1. **Compare transcriptomic changes between ZIKV and DENV**

Table 1.1: Host proteins that interact with flavivirus C (capsid) protein.

Host proteins that interact with flavivirus C protein in different cellular compartments and how this antiviral treatments preventing this interaction may affect the virus. The table is color coded for easier viewing: ER: blue, nucleus: green, nucleolus: red, and lipid metabolism: orange. From Sotcheff & Routh, *Pathogens*, 2020.

Location.	Pathway	Protein	Result	Pro- or Anti-viral	Antiviral treatment could
Endoplasmic Reticulum	Vesicular Transport	YKT6 SCFD1 USE1 VAPB STX8 TMEM85 SAR1B	Movement of capsid from site of cleavage to lipid droplets or nucleus	Pro	Prevent capsid storage on lipid droplets, entry into nucleus/nucleolus, prevent particle assembly
	Signal Peptidase Complex	SPC52 SEC11A SPC53	Cleavage of capsid from flavivirus polyprotein		
	ER Stress Response	PARP16 DORGK1 UBXN8 RTN3 SLC33A1	Cell survival or cell death	Anti	
Nucleus	Importins	IPO7 IPO9	Targeting of viral capsid to nucleus	Both	Prevent capsid interaction & entry into host nucleus
	LINC Complex	SUN1 SYNE2	Interaction with nuclear membrane		
	NPC	NUP35 NUPL2 TMEM48 POM121	Entry of viral capsid into host nucleus		
	NMD Pathway	PABPC UPF1	Prevent degradation of viral transcripts, increase degradation of host transcripts	Pro	Allow degradation of viral transcripts by host cells
	Host Transcription	DAXX	Prevent binding of host transcription factors		Prevent changes in host gene transcription
		Core Histones	Change position of histones		
Nucleolus	p53 regulation	MDM2	p53 induced apoptosis	Anti	
	RNA methylation	NSUN2	Modified methylation of tRNAs, mRNAs, and ncRNAs	Not yet understood	Prevent changes in transcript methylation
	Ribosome biogenesis	NOL8 NOP16 NOC4L B23 NPM1 nucleolin	Ribosomal stress	Anti	
	Protein turnover	Jab1	Remove viral capsid protein		
Lipid Metabolism	Lipid droplets	AGPAT6 LPCAT2 AUP1 MBOAT2	Storage of accumulated capsid until particle assembly	Pro	Cause issues with particle formation; formation of VLPS
	Sphingolipid Metabolism	KDSR TEX2	Initial budding into ER		
	Lipid binding	GRAMD 1A GRAMD 3	Interaction with lipid membrane		
	Ceramide Metabolism	SMPD4 CERS2 GBA2	Initial budding into ER		

To determine differentially expressed genes (DEGs) associated with ZIKV infection, we are using a method developed by our lab, Poly(A)-ClickSeq (PAC-seq) (*145, 146*) which also allows us to look at alternative poly-adenylation (APA) patterns that may affect gene expression by altering stability and/or regulation of mRNA transcripts. This requires the use of an analytical pipeline called Differential Poly-A Cluster (*DPAC*) analysis that uses custom scripts and an R package (DESeq2) to look at differential expression of terminal exons, poly-A clusters (PACs), and genes. After enrichment analysis, the potential for AS was apparent in our PAC-seq data, therefore we decided to also use long read sequencing and full length alternative isoform of RNA (*Flair*) (*147*) analysis to identify AS events in our ZIKV infected JEG3 samples. Overall, our results suggest that additional interactions may be responsible for changes in splice patterns than the previously reported NS5:spliceosome interaction in related DENV.

2. Determine relevance of APA in response to flavivirus infection

To evaluate the relevance of APA as a result of flavivirus infection we wanted to determine which RNA binding proteins are able to bind the differentially regulated PACs. To accomplish this we first computationally extract a list of genes that have differentially regulated PACs from the DESeq2 output in *DPAC* by setting parameters for log2FoldChange and adjusted p-value. We generate two lists per comparison, an up- and a down-regulated. These lists can be used to extract all alignment information from the PAC bed files and subsequently a custom python script can pull sequences for these annotated PAC, within varying lengths of the PAC. These sequences can be put into online tools to determine motif enrichment and enriched motifs are cross-referenced against databases of RNA binding proteins and their known sequence specificities.

Use Poly(A)-ClickSeq (PAC-seq) and single nuclei RNAseq to investigate changes in gene expression in the mesolimbic pathway of mice upon opioid self-administration (Chapter 4)

In these studies, we are investigating the potential impact of oxycodone or fentanyl on disease severity as it pertains to viral infection. The ventral tegmental area (VTA) and nucleus accumbens (NAc) are the primary brain regions where addictive drugs act (*38*). The NAc is one of two parts of the ventral striatum. The VTA is located in the midbrain and consists of primarily glutamatergic, GABA-ergic and dopaminergic neurons (*148, 149*), whereas the NAc is primarily medium spiny neurons. Reward-seeking behavior elicited by use of opioids and other addictive substances or predictive uses (ex. seeing paraphernalia) is caused by release of dopamine in the ventral striatum from neurons originating in the VTA (*149-153*). These neurons are referred to as the mesolimbic pathway, which is involved in salience, motivation, and reinforcement learning (*38, 154*). Therefore, to investigate how opioid use may serve as an indication of severity of disease from viral infection, we conducted transcriptomic studies on nuclei from the NAc and total cellular RNA from both the NAc and the VTA. We have data for 12 rats for each of three experiments: rats immediately sacked after their last dose, rats sacked 24 hours after their last dose, and rats sacked 10 days after their last dose. The prolonged periods after their last dose are to mimic withdrawal. These transcriptomic studies aim to shed light on how viral infection may impact individuals suffering from OUD.

1. **Single nuclei RNAseq to investigate gene expression changes by cell type in the nucleus accumbens upon acute withdrawal from fentanyl use**

To determine if opioid use may impact severity of symptoms from a viral infection we aimed to identify differentially expressed genes in different types of cells within the NAc upon

acute withdrawal from fentanyl use via single cell sequencing. To accomplish this, we synthesized single nuclei libraries using the 10X Genomics 3' end NextGEM kit. We were able to generate libraries for 7 fentanyl self-administering and 5 saline self-administering rats. Data were analyzed using the Seurat package in R and ScDblFinder to find and remove any potential doublets (a doublet is a "cell" or "nuclei" in our data that actually represents 2 or more cells based on its expression profile). This allowed us to determine genes that were differentially expressed. We also looked for expression of genes involved in inflammation and the antiviral response.

2. PAC-seq to investigate gene expression changes in the nucleus accumbens and ventral tegmental area upon withdrawal from oxycodone use

To determine if opioid use may impact severity of symptoms from a viral infection we aimed to identify differentially expressed genes within the NAc and VTA immediately after use of or upon withdrawal from oxycodone. To accomplish this, we synthesized cDNA libraries of mRNA using PAC-seq protocols. These were sequenced on an Illumina platform and resulting data was analyzed using the *DPAC* pipeline. This uses DESeq2 (in R) which allowed us to determine genes that were differentially expressed in each condition. We also looked for expression of genes involved in inflammation and the antiviral response.

Chapter 2 Materials and Methods

In this chapter some sections are directly lifted from my published works. The proper citation for these works is provided below. For two of these works I am the first author and as such I wrote the majority of or all of the rough draft of the manuscript and had help with revisions from Dr. Andrew Routh. For the ZIKV transcriptomics (first) paper I extracted RNA, prepared RNA libraries, did all of the computational analyses, and prepared all of the figures. For the *ViReMa* (second) paper I conducted all computational analyses for libraries that were previously prepared and sequenced by or for collaborators and prepared all of the figures. These citations are also included in the references section. In the third manuscript I prepared both Tiled-ClickSeq and ARTIC libraries for comparative analysis. In the last manuscript I prepared single nuclei RNAseq libraries using 10X protocols, conducted all computational analyses and prepared all figures, and wrote the majority of the manuscript draft.

Sotcheff, S.; Elrod, N.; Chen, J.; Cao, J.; Kuymuycu-Martinez, M.; Shi, P-Y.; Routh, A. Zika virus infection alters gene expression and poly-adenylation patterns in placental cells. (*in Preparation*)

Sotcheff, S.; Zhou, Y.; Sun, Y.; Johnson, J.E.; Torbett, B.E.; Routh, A. *ViReMa*: A Virus Recombination Mapper of Next-Generation Sequencing data characterizes diverse recombinant viral nucleic acids. *BioRxiv*. 2022 Mar. https://doi.org/10.1101/2022.03.12.484090

Jaworski, E.; Langsjoen, R.; Mitchell, B.; Barbara, J.; Newman, P.; Plante, J.; Plante, K.; Miller, A.; Zhou, Y.; Swetnam, D.; **Sotcheff, S**.; Morris, V.; Saada, N.; Muchado, R.; McConnell, A.; Widen, S.; Thompson, J.; Dong, J.; Ping, R.; Pyles, R.; Ksaizek, T.; Menachery, V.; Weaver, S.;

Routh, A. Tiled-ClickSeq for targeted sequencing of complete coronavirus genomes with simultaneous capture of RNA recombination and minority variants. eLife. 2021 Sept. http://dx.doi.org/10.7554/eLife.68479

Sotcheff, S.; Zheng, J.; Stafford, S.; Routh, A.; Anastasio, N.; Mendoza, I.; Cunningham, K. Acute withdrawal from fentanyl use may exacerbate severity of COVID-19. (in Preparation)

CELL CULTURE

Human placental cells (JEG3) were maintained at 37^{o}C in DMEM (11965-092, Gibco) supplemented with non-essential amino acids (11140-050, Gibco), sodium pyruvate (58636, Sigma), 10% fetal bovine serum (FBS), 1% pen-strep (P/S) and HEPES (15630-080, Gibco). Human liver cells (Huh7) were maintained at 37^{o}C in DMEM supplemented with 10% FBS, 1% P/S and sodium pyruvate.

INFECTIONS AND PLAQUE ASSAYS

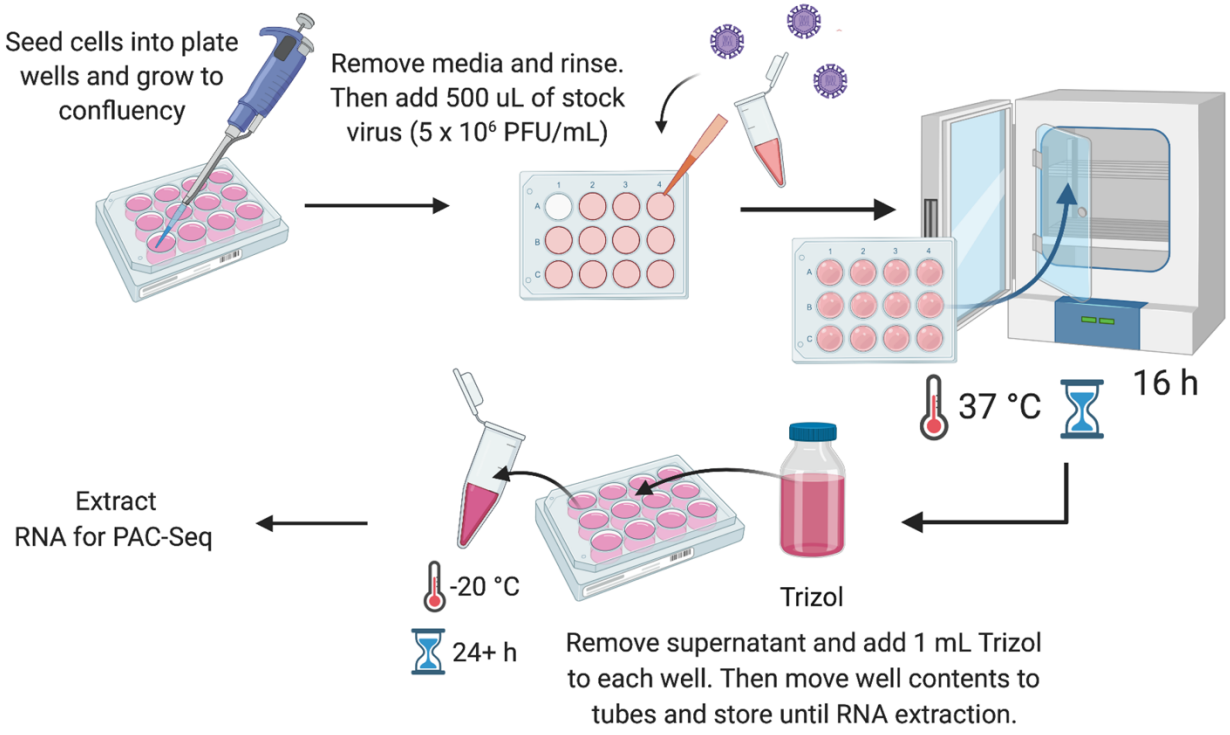
The ZIKV Puerto Rican strain (PRVABC59) was grown in African green monkey kidney (Vero) cells and stocks of virus were prepared by centrifuging the post-transfection media to remove cell debris, then aliquoting the supernatant. A DENV-2 strain (DENV2Y98P) was grown in Vero cells and stocks of virus were prepared by centrifuging the post-transfection media to remove cell debris, then the supernatant was aliquoted. Titers of each Passage were calculated by traditional plaque assay. JEG3 cells were either mock-infected with media or infected with either Passage 2 (P2) PRVABC59 or DENV2Y98P with a multiplicity of infection (MOI) of 3 plaque

forming units (pfu)/cell in order to ensure the infection of all cells to obtain a representative RNAseq profile of each condition. **(Figure 2.1)** Similarly, Huh7 cells were either mock-infected with media or infected with P2 PRVABC59 at an MOI of 3. After adsorption of virus for 1 h at 37°C, the virus-containing medium was removed and replaced with 2 mL fresh culture medium. Infected cells were then incubated at 37°C and harvested at 16 hours post-infection (hpi) by pipetting with 1 mL of TriZol reagent. Cell membranes were lysed and proteins degraded (in effect inactivating any remaining virus within cells) by storing samples in TriZol reagent at -20 ºC prior to RNA extraction. For vPAR-CL studies we used the Dakar strain of ZIKV, generated by *in vitro* transcription and electroporation in Vero cells. We generated P2 virus in Vero cells and used this to infect fresh Vero cells at an MOI of 1. We used plaque assays to compare the production of infectious particles (PFU) when 4-thiouridine (4SU) was supplemented in media to virion production without 4SU.

RNA EXTRACTION

Total cellular RNA was isolated from thawed cells in TriZol using Direct-zol RNA Microprep kits (Zymo Research). Buffers were prepared with recommended ethanol addition to the RNA PreWash and Wash Buffer and DNase I was reconstituted with the recommended amount of water. 500 µL of TriZol Reagent was added to each well from 6 well plates of either infected or mock infected human cells, where media had previously been removed. An equal volume of 96% ethanol was added and mixed, prior to Zymo-Spin IC columns in collection tubes for centrifugation with retention of the nucleic acids in the column. After DNase treatment at room temperature for 15 minutes, the RNA PreWash buffer is added to the column prior to centrifugation – allowing digested DNA to flow through the column, but keeping RNA in the column. RNA in

Figure 2.1: Schematic of infections for flavivirus transcriptomics studies. JEG3 or Huh7 cells were infected with ZIKV or DENV at an MOI of 3, and allowed to incubate at 37ºC for 16 hours prior to the addition of TriZol reagent and subsequent RNA extraction. Image created by S. Sotcheff using BioRender under license.



the column is washed with the RNA Wash Buffer and centrifuged once more, prior to elution with DNase/RNase Free water. RNA samples were quantified on a spectrophotometer (NanoDrop ND-1000; Thermo Scientific), with attention to the A260/A280 and A260/A230 ratios for quality. This RNA is suitable for use to prepare PAC-seq libraries.

To prepare RNA to generate long-read cDNA libraries for the ONT MinION, RNA was poly(A) selected by using dT oligonucleotides attached to magnetic beads with the Poly(A) mRNA Magnetic Isolation Module (NEBNext). Magnetic Oligo $d(T)_{25}$ Beads were washed with RNA Binding Buffer twice before adding dilute 1 μg of the total cellular RNA extracted above. RNAs were denatured at 65 ºC for five minutes then held at 4 ºC. These were mixed thoroughly and allowed to incubate at room temperature to allow binding of the poly-A RNA to the beads. The beads were pelleted on a magnetic rack and washed twice with Wash Buffer. Poly-A RNA was eluted with Tris Buffer by placing the tubes in thermocycler set to 80 ºC for 2 minutes, then brought to room temperature. Binding buffer was added again to allow the RNA to bind the beads a second time. The wash and elution steps are completed again, and purified poly-A RNA quality is assessed using a spectrophotometer.

LIBRARY PREPARATIONS

We utilize a number of different RNA library preparation methods to investigate different aspects of viruses. The overall method is largely similar across the board including reverse transcription, adapter ligation, polymerase chain reaction (PCR) amplification and size selection. Although most of the libraries produced were sequenced on an Illumina platform, we will note that the long-read cDNA and ARTIC libraries were sequenced on an ONT MinION. The

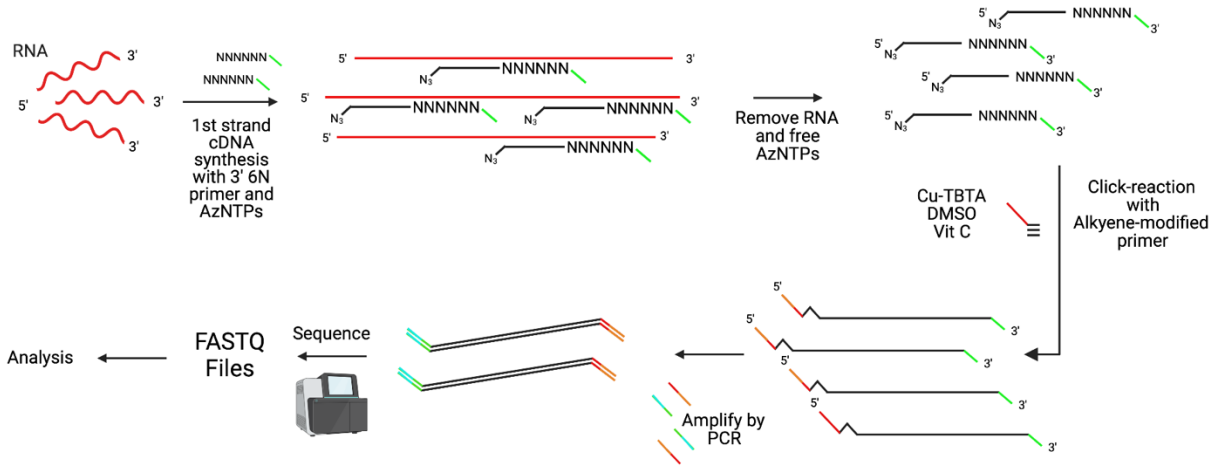
differences between the various methods of library preparations and their uses will be outlined below.

ClickSeq

ClickSeq (*155, 156*) is a method developed by my mentor Dr. Andrew Routh, that is largely similar to traditional RNAseq in that reverse transcription from the sample RNA is conducted with random primers (3' – 6N: see Appendix A). **Figure 2.2** outlines the protocol for this library preparation method. However, traditional RNAseq requires a fragmentation step prior to ligation of a sequencing adaptor and sub-sequent PCR amplification. The reverse transcription step in ClickSeq (with SuperScript III reverse transcriptase) utilizes the incorporation of azido-NTPs (deoxyribonucleotides with an azido group in place of the 3' hydroxyl group) at a 1:35 ratio. These azido-NTPs are randomly incorporated into the newly synthesized complementary DNA (cDNA) and cause stochastic termination of the reverse transcription reaction because they lack the 3' -OH group necessary for addition of another nucleotide to the strand. This azido group also serves as the site of a click reaction where the p5 adaptor can be ligated. This is a copper catalyzed reaction between an azide and alkyne group resulting in the formation of a 5-membered triazole ring linkage (Copper-catalyzed Alkyne Azide Cycloaddition or CuAAC). This ring does not perturb subsequent PCR amplification that also incorporates our p7 adaptor (index). We use unique molecular identifiers in the p5 adaptor and a low number of PCR cycles, about 17 to 20, to mitigate PCR duplication. An agarose gel is used post-PCR to select for libraries that are 400-700 bp in length. These libraries can be sequenced on an Illumina platform.

Poly-A ClickSeq

Figure 2.2: ClickSeq library preparation method. Reverse transcription is done with random primers with a 3' 6N sequence and SuperScript III with a 1:35 ratio of azido-NTPs (az-NTPs) to regular NTPs. The cDNA complexed with RNA is treated with RNase H to remove RNA. We clean the cDNA to remove RNA and free nucleotides and az-NTPs. An i5 adapter with a alkyne group is added by a copper-catalyzed click cyclo-addition, then libraries are PCR amplified which adds an i7 index. PCR amplified libraries are gel purified (fragments 400-700 nts long) and then pooled and sequenced on an Illumina platform. Image created by S. Sotcheff using Powerpoint and BioRender under license.

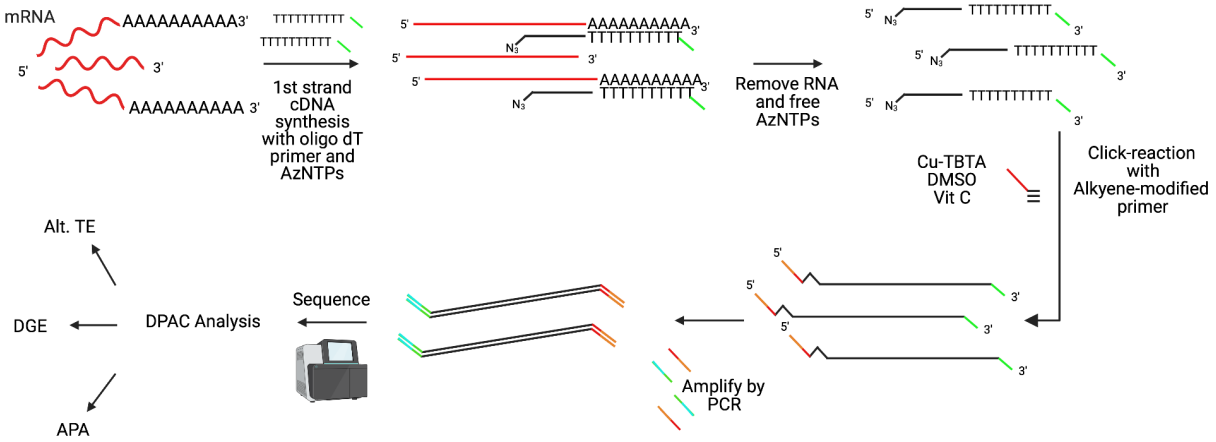


Others use traditional RNAseq to produce libraries for differential gene expression analysis. This requires the fragmentation that ClickSeq (above) circumvents as well as a very large number of reads per sample (~30-40M) to cover the entire transcriptome, or to perform a poly-A selection or ribo-deplete on the extracted total cellular RNA. Using Poly-A ClickSeq (PAC-seq) we use an oligo-dT (see Appendix A) primer in the reverse transcription (RT) step along with azido-VTPS (azido-NTPs excluding azido-TTP) at a 1:5 ratio with regular NTPs (*145, 146*). This saves both time and money, removing a selection or depletion step and minimizing the number of reads (~10M) needed per sample as only the 3' end is sequenced of each mRNA. We begin with ~1 μg (but can use as little as 100 ng) total cellular RNA in the RT reaction. Then we use a copper catalyst to click ligate the p5 adapter before PCR amplification that adds the p7 adapter. We use 17-20 cycles and unique molecular identifiers (UMIs) to minimize PCR duplication and be able to remove them computationally (respectively). These libraries are run on agarose gels and size selected for 200-400 base pairs. Libraries are pooled and run on an Illumina platform. The PAC-seq protocol is outlined in **Figure 2.3.**

Oxford Nanopore cDNA PCR barcoded long-read libraries

Poly-A selected total cellular RNA was used to synthesize libraries of full-length transcripts using the Oxford Nanopore cDNA PCR barcoding kit. This entailed reverse transcription of 50 ng poly-A selected RNA using SuperScript IV reverse transcriptase and a strand switching primer. The amplification via PCR is accomplished with a LongAmp Taq master mix and a barcoding primer. The resulting cDNA is cleaned using AMPure XP beads and 70% ethanol, eluted in the ONT rapid annealing buffer (RAB). The ONT cDNA adapter mix (CAM) is added to the amplified library, and ligation occurs at room temperature over 5 minutes.

Figure 2.3: Poly(A)-ClickSeq (PAC-seq) library preparation protocol. Reverse transcription is done with oligo-dT primers, SuperScript III, azido-NTPs (az-NTPs) at a ratio of 1:5 to regular NTPs. The cDNA complexed with RNA is treated with RNase H to remove RNA. We clean the cDNA to remove RNA and free nucleotides and az-NTPs. An i5 adapter with an alkyne group is added by a copper-catalyzed click cyclo-addition, then libraries are PCR amplified which adds an i7 index. PCR amplified libraries are gel purified (fragments 200-400 nts long) and then pooled and sequenced on an Illumina platform. Image created by S. Sotcheff using Powerpoint and BioRender under license.



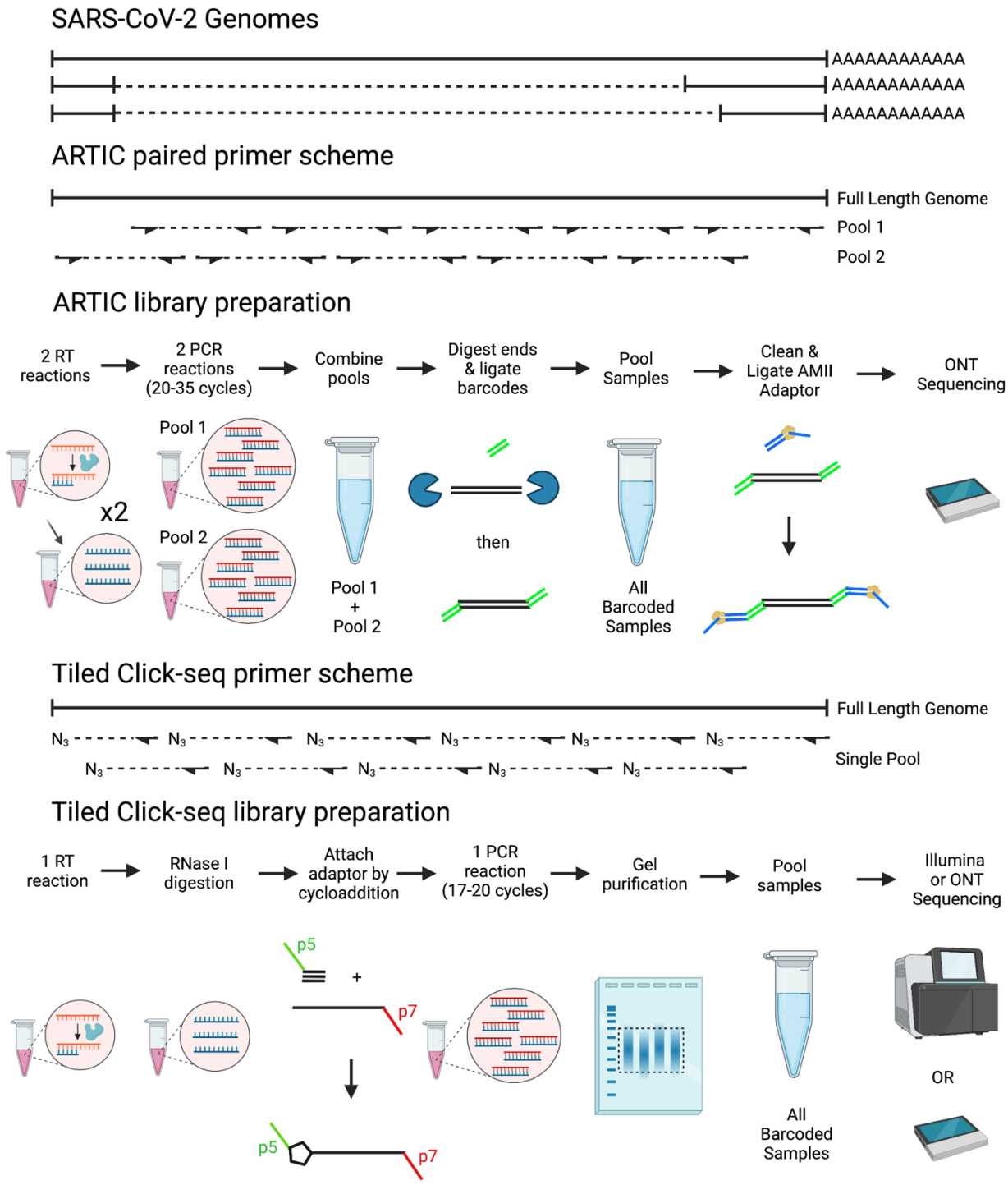
The adaptor in this case contains a motor protein that will aid with sequencing on an ONT platform. Adaptor ligated libraries are cleaned with the use of AMPure XP beads and ethanol an additional time, eluted in the ONT elution buffer (ELB or EB). To sequence, protocols for priming the MinION flowcell (which is plugged into the USB of a computer running MinKNOW) were followed. Buffer and loading beads were added to the library prior to loading the flow cell, allowing reads to be sequenced for 24 to 48 hours depending on flow cell quality. The resulting reads were base-called and de-multiplexed after sequencing was complete. These reads were used for AS analysis and to validate APA events found in the PAC-seq data.

ARTIC

The ARTIC network has published a protocol for sequencing SARS-CoV-2 in patient samples even when undetectable by standard q-RTPCR (*65*). First reverse transcription takes place with two sets of SARS-CoV-2 specific primers (**Figure 2.4,** see Appendix B). Forward and reverse primers are placed roughly 300 bases apart and these cover the entire genome in each set, with overlap between the first and second set. After reverse transcription amplicons are generated then cleaned using SPRIselect beads. The ends are repaired and an A-tail is added before adaptor ligation. Next, indexing and amplification is accomplished via PCR. It is important to note, that a large number of PCR cycles is required (~35), meaning that PCR duplication may make it difficult to reliably identify any recombination events or single nucleotide poly-morphisms. These can be sequenced on an Illumina platform with paired end capability. A modified protocol can result in libraries that can be sequenced on the ONT MinION.

Tiled-ClickSeq

Figure 2.4: Library preparation methods using sequence specific primers for SARS-CoV-2. (Top) ARTIC and (bottom) Tiled-ClickSeq both use sequence specific primers to reduce the overall reads necessary to get adequate coverage to detect single nucleotide variants and recombination events in SARS-CoV-2. Image created by S. Sotcheff using BioRender under license.



In summary, this method combines the ClickSeq and ARTIC protocols described above and is compared to ARTIC in **Figure 2.4**. Like ARTIC, this protocol uses primers for the reverse transcription step that are specific to a virus of interest. In this dissertation we are comparing ARTIC to Tiled-ClickSeq, and therefore the primers used were specific to SARS-CoV-2.

Different from ARTIC, we are only using a reverse primer (*63*) – no forward primer. The primer sequences were generated by taking the "right" primer sequence generated by *primalseq* (*157*) and including a short linker between it and the Illumina p7 adaptor (e.g. GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT + NNNN + TGTCTCACCACTACGACCGTAC). Primer sequences can be found in Appendix C. Other than the use of the specific primer, this protocol is identical to ClickSeq: reverse transcription includes the incorporation of azido-nucleotides for stochastic termination, cDNA is purified on magnetic beads, a click reaction is used to ligate the adapter before another round of bead purification and finally PCR.

Single nuclei RNAseq (snRNAseq) with 10X

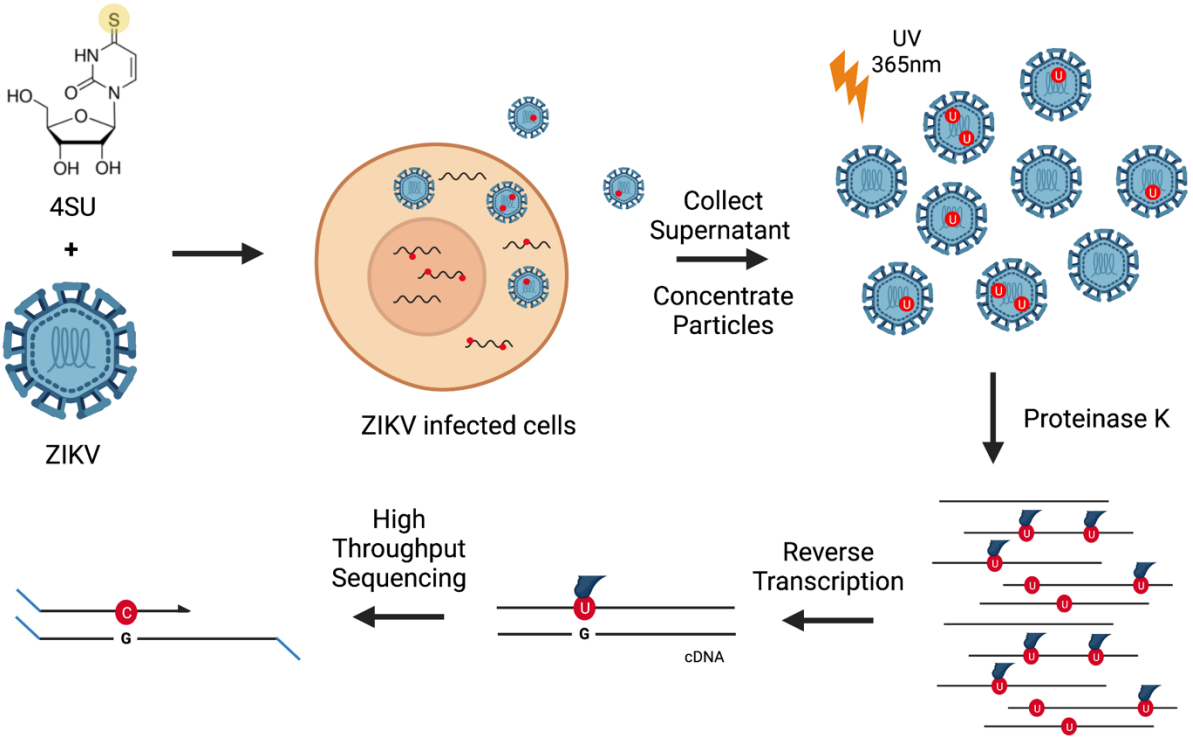
10X Genomics provides a means to generate single cell and single nuclei sequencing. To do this we start by partitioning cells, or nuclei in this case, into nano-liter scale Gel Beads-in-emulsion (GEMs) so that all generated cDNAs share a common 10X barcode. GEMs are created by loading a partitioning oil, a master mix for the RT reaction containing the nuclei, and the gel beads onto a Chromium NextGem Chip G. Each gel bead is coated in an oligo(dT) primer to capture the 3' of mRNAs as well as two additional primer sequences used for capturing and priming of the 10X Feature Barcoding technology. We loaded ~7000 cells per sample. After the GEMs were generated we transferred the emulsions to tubes for placement into the thermocycler

to dissolve the gel beads releasing the primers and lysing any co-partitioned nuclei. After reverse transcription GEMs were broken and magnetic beads were used to purify cDNA from pooled fractions for amplification via PCR. The amplified DNA is then fragmented and adaptors are ligated prior to a final round of PCR amplification. Quality of the libraries was assessed using an Agilent BioAnalyzer High Sensitivity Kit and the libraries were sequenced at Baylor College of Medicine in Houston, TX. They were sequenced on an Illumina NovaSeq, producing ~400M reads per sample.

Virus Photo-activateable ribonucleoside crosslinking (vPAR-CL)

vPAR-CL, mentioned above, has previously been used by our lab to identify capsid-RNA interactions in FHV (*107*). P1 particles of PRV59ABC were used to infect Vero cells and the ZIKV infected cells were be treated with 4-thiouridine (4SU), a nucleoside analog that can be converted to 4S-UTP and incorporated into newly synthesized RNA. Virus particles were purified and concentrated from the media 48 hours post infection. These analogs crosslink with lysine and aromatic amino acid side chains when subject to 365 nm UV. Viral structural proteins were digested and these RNAs were subject to RT with azido-NTPs to stochastically terminate the reactions **(Fig. 2.5)** (*107*). Adapters were ligated to the cDNA fragments and PCR was used to amplify the libraries to be sequenced using an Illumina platform, following the ClickSeq protocol.[155] The incorporated 4SU nucleotides retain a piece of the crosslinked amino acid side chain causing it to pair incorrectly during reverse transcription. Therefore, NGS reveals T to C mutations **(Fig. 2.5).** We used custom scripts (noted below) for analysis. Our negative controls were cells treated with 4SU+/UV-, 4SU-/UV+, and 4SU-/UV-.

Figure 2.5: Protocol for getting 4SU incorporated flaviviral RNA in virions for virus photoactivateable ribonucleoside crosslinking (v-PAR-CL). 4-thio-uridine (4SU) is supplemented in media for Vero cells infected with ZIKV Dakar. The cells convert 4SU to 4S-UTP and this can be incorporated into nascent RNAs – including newly synthesized viral RNA. This is packaged into particles which we purify by PEG precipitation. We UV crosslink the capsid protein to the 4SU incorporated into the viral genome and then treat with proteinase K prior to generating ClickSeq libraries. 4SU in the viral genome that crosslinked to protein will have an adduct that causes incorrect base pairing during library preparation which results in NGS data with T->C at those nucleotides. Image created by S. Sotcheff using BioRender under license.



DATA ANALYSIS

The next generation sequencing (NGS) data generated by sequencing the libraries produced with all of the methods listed above requires extensive data analysis. Typically sequencing platforms can de-multiplex for you, but it is also necessary to trim the reads in the resulting FASTQ files and quality filter them. Then these reads have to be aligned. Parameters for alignment vary for different pipelines. We detail below the pipelines necessary to accomplish the works described in this dissertation and their uses in their respective studies.

Differential Poly-A Cluster (*DPAC*) analysis

Differential Poly-A Cluster (DPAC) was used to analyze the data from the short-read Illumina libraries.[158] *DPAC* can identify changes in overall expression, poly-adenylation or poly-A cluster (PAC), and terminal exon. The overall pipeline has been previously described. The command ran for this data set was:

```
~/DPAC -p PMCDB -t 4 -x [flattened_annotations] -y [reference_names] -g [genome]
-n 3 -v human,hg19 [metadata_file]  [experiment_name] [output_directory]
```

Here -p indicates processes to run, in this case P (perform data pre-processing), M (map data), C (force new PAC cluster generation), D (perform differential APA analysis), and B (make individual bedgraphs), -t indicates how many threads were to be used, and -n indicates number of replicates. Before running D, it is necessary to generate a text file (metadata) that indicates which .fastq files are associated with a particular treatment. For my dissertation project these studies we used annotations, gene names, index for human and mapped to the human genome (hg19). In

chapter 4 we also highlight the use of *DPAC* to analyze transcriptomic changes in rat brains using the corresponding annotations, gene names, index and genome (rn6).

Extraction of sequences surrounding poly-A clusters and motif enrichment analysis

First, the PACs DE-Seq2 output is filtered for |log2fold change| > 0.585 and adjusted p-value of 0.1, where PACs with a log2foldchange values greater than 0.585 are up-regulated and those with log2foldchange values below -0.585 are down-regulated. This indicates either a 50% increase or decrease of the usage of a particular PAC respectively. From here lists of up- and -down-regulated PACs were generated for each experiment (ZIKV Huh7, ZIKV JEG3, and DENV JEG3 each vs. mock infection at 16 hpi). This was done using an awk command as shown below.

```
awk -F, '{if($4 > 0.585 && $8 < 0.1) print $2}' [Condition1]-vs-[Condition2]_[Experiment]_PACs_output_PAC_DESeq2-results.csv > [Condition1]-vs-[Condition2]_[Experiment]_PACs_up.txt
```

We then gather the alignment information from the PACs.bed file for only genes in their respective lists using the following command in terminal:

```
while read line; do echo $line; grep $line [Experiment]_PACs_output_PACs.bed > up_PACs.bed; done < [Condition1]-vs-[Condition2]_[Experiment]I_PACs_up.txt
```

We can then use a python script to extract the sequence information of those alignments within a user-defined number (X) of nucleotides of the PAC.

```
python2 Extract_nts.py up_PACs.txt [genome].fa [output_name].txt X
```

This writes sequences that cover X nts upstream to X nts downstream of the annotated PAC for each gene that has differentially regulated PAC according to the PAC_DESeq2-results.csv output of *DPAC*. We chose to pull sequences using X values of 100, 250, 500, 1000, and 2000. These sequences were all put into an online motif enrichment tool called *DREME* (*159*) to identified enriched sequences. Enriched sequences were compared to RBPDB (*160*) to determine potential RNA binding proteins that bind the enriched motifs.

***Full length alternative isoform of RNA (Flair)* analysis**

A python program, *Flair*, was used to identify alternative splicing (AS) events across the MinION long-read samples.[147] First, .fastq files were aligned to the human genome (hg19). The resulting bed files were corrected, converted to .psl, and concatenated.

```
python3 [path]/flair.py align -r $File -g [genome].fa -o [Root] -m minimap2
python3 [path]/flair.py correct -q [Root]'.bed' -g [genome].fa -o
'Corrected_'[Root] -f [genome]_NCBIRefSeq.gtf
python3 [path]/bin/bed_to_psl.py [genome].chrom.sizes
'Corrected_'[Root]'_all_corrected.bed' [Root]'_corrected.psl'
cat *.psl > [Root]'_All_Corrected.psl'
cat *.fastq > [Root]'_All.fastq'
```

The collapse parameter was used with the Ensembl mRNA annotations from the UCSC Genome Browser to generate tables of used isoforms.(*161*)

```
python3 [path]/flair.py collapse -g [genome].fa -r [Root]'_All.fastq' -q [Root]'_All_Corrected.psl' -f [genome]_Ensembl.gtf -m minimap2
```

The .psl files were converted to .bed files and we used flair-quantify and flair-diffSplice to generate a matrix of the uses of all splice isoforms within each sample and diff-splice to identify AS events. This produces 4 tab-delimited files, one for each type of event detected by the pipeline: exon skipping, intron retention, alternative 3' and alternative 5'.

```
python3 [path]/flair.py diffSplice -i flair.collapse.isoforms.psl -q [Root]'_counts_matrix.tsv' -t 4 --conditionA $cond1 --conditionB $cond2 -o [Root]'_diffSplice_'$cond1'_vs_'$cond2
```

Command line was used to find and replace annotations with gene names. The *Flair* package also includes a python script (plot_isoform_usage) that uses R to generate plots of differential isoform usage which we used to depict AS events using gene names.

```
Python3 [path]/plot_isoform_usage.py [isoforms.psl or isoforms.bed] counts.matrix.tsv [gene_name]
```

Two batch scripts were prepared to accomplish these tasks, they are included in Appendices D and E. The first is for pipeline steps on individual sample sets and the second is to merge the sets into one large set and continue with AS analysis. Note that plot_isoform_usage.py is used for individual genes and therefore not included in the batch scripts.

Seurat

Seurat is a package in R that can be used to analyze single cell sequencing data (*162*). Cell Ranger, a tool for converting 10X fastq files to a counts matrix, gives an output of several .csv files as opposed to a full matrix to save disk space by saving data with the least number of rows or columns possible. These files include information on gene counts, features or genes, and individual cell barcodes. For each sample all of these files are input into R using Seurat. Before further analysis we filter the data to control for quality of the data. This includes excluding cells that are expressing a high percentage of mitochondrial genes, cells that appear to express either too few or too many genes, and cells that have a high probability of actually representing more than one cell (based on a doublet score – calculated using R package scDblFinder (*163*)). The parameters we used for this particular data set are noted in the R code below:

```
SAMPLE=Read10X(data.dir="~/PATH/SAMPLE/filtered_feature_bc_matrix/")
SAMPLE_so=CreateSeuratObject(counts=SAMPLE,project="SAMPLE", min.cells = 10,
min.features = 800)
SAMPLE_so[["percent.mt"]]=PercentageFeatureSet(SAMPLE_so, pattern = "^Mt-")
SAMPLE_so[["doublet.score"]]=FindScores(SAMPLE, methods)
SAMPLE.qc=subset(SAMPLE_so, subset = nFeature_RNA > 700 & nFeature_RNA < 2500
& percent.mt<5 & doublet.score<0.2)
```

This code indicates that we are choosing to keep cells that express a minimum of 700 and a maximum of 2500 genes (this higher number than standard scRNAseq minmum of 200 is because we are using nuclei as opposed to entire cells), a gene or feature is only kept in the data set if it is expressed in 10 or more cells, cells where more than 5% of the genes expressed are mitochondrial or have a doublet score of greater than 0.2 are excluded. This quality filtering is conducted for each sample individually before combining the data sets. After the data is

combined it is scaled and integrated before UMAP analysis is ran and clusters are found. UMAP plots indicate clustering of cells based on similar expression profiles where cells can be labelled by cluster or sample identification. From expression of marker genes in each cell is conducted by generating feature plots. This, in addition to a dotplot showing expression of all markers of interest across each cluster, aids in the determination of cell type for each cluster (code shown below).

```
DimPlot(INTEGRATED_DATA, pt.size = 0.5, label=TRUE, label.size=5) +
NoLegend()
MARKER_GENE=FeaturePlot(INTEGRATED_DATA,c("MARKER_GENE"))
MARKER_GENE+ggtitle("MARKER_GENE”)
DotPlot(INTEGRATED_DATA,assay="RNA",LIST_OF_MARKERS)+ theme(axis.text.x =
element_text(angle = 90, vjust = 0.5, hjust=1)) + ggtitle("Marker Genes")
```

Once cell determination is complete, clusters are re-labelled with their cell type and the data can be subset by cell type. Further subsetting by treatment allows labelling of cells with a treatment group, then the data for a given cell type can be recombined and re-scaled. At this point differential expression between groups can be completed. We are able to generate lists of differentially expressed genes (DEGs) and produce heatmaps and dotplots for relevant DEGs.

```
CELL_TYPE_MarkersbyTreament=FindMarkers(CELL_TYPE.combined,ident.1=
"TREATMENT1", ident.2="TREATMENT2")
dittoHeatmap(CELL_TYPE.combined,DEGs,annot.by=
c("seurat_clusters"),cluster_rows=FALSE, annotation_colors=annotation_colors,
main="CELL_TYPE")
```

The entire R script for this analysis can be found in Appendix F.

Virus Recombination Mapper (*ViReMa*)

Virus recombination mapper (*ViReMa*) is a previously published method that uses small read mappers such as *bowtie* and *bwa* starting at the left-most position and reading until a qualifying mis-match occurs – presumably the junction of a recombination event. Due to base-calling errors and biological variation, it is not uncommon for a mis-match to occur. Therefore, 0 to 2 mismatches are tolerated in a mapped read before a junction is inferred. Mismatches are not tolerated within 'X' nucleotides of a junction, where the default value of 'X' is 5, but can be specified in command line. With longer reads (300 nt) the limitation of 2 mismatches per read appeared arbitrary, so an 'Error-Density' parameter was introduced to the pipeline, allowing for 1 mismatch in a single stretch of 25 nt by default, but allowing for changes in this parameter in command line. This newer version of *ViReMa* also produces outputs in canonical SAM files with associated SAM tags and appropriate CIGAR 1.4 scores. This is user friendly in that it allows for easy loading into visualization tools such as Tablet or Integrative Genomics Viewer (IGV) as well as allowing the incorporation of *ViReMa* into many analysis pipelines. *ViReMa* can be used to detect various recombination events: deletions, duplications, copy-backs, and virus-host recombination events.

For many of the case studies provided we ran *ViReMa* using the standard parameters, command line shown below. Note that if a parameter is not listed (ex. error density) then the default value is used.

```
--X 3 --Defuzz 0 --Host_Seed 25 --p 4 -BED --MicroInDel_Length 5
```

To investigate how changing the error density parameter affected permissibility of *ViReMa* we altered that parameter to 3 mismatch in a 100 nt window, 1 in 100, 1 in 10, 1 in 15 and 2 in 25. Where 1 in 100 is the least permissive, and 1 in 10 is the most permissive. When investigating the error density parameter, all other parameters were left the same.

v-PAR-CL scripts

The purpose of v-PAR-CL is to determine where protein crosslinked to an RNA of interest without the need to immuno-precipitate. This can be accomplished in virus particles where only one protein should be in contact with the viral genome. This method requires the incorporation of 4SU (converted to 4S-UTP) into nascent RNAs and photo-crosslinking. When the protein is degraded a small adduct remains covalently attached to the RNA at the U positions where the analog was incorporated. These adducts cause mismatching during reverse transcription that results in specifically T -> C mutations in NGS data.

A batch script was used to trim the adapter sequence AGATCGGAAGAGC from each read and retaining those reads that are a minimum of 40 nt in length after trimming using fastp. The batch script also uses *bowtie* and *samtools* for alignment to the genome of interest, in this case ZIKV Dakar (named after the city it was first isolated in, the capital of Senegal) and indexing of reads. A pileup file was generated using *samtools* mpileup with a maximum read depth of 1000000.

```
#!/bin/bash
File=$1
Root=${File##*/}
```

```
Root=${Root%%_S*}
echo "Working on "$Root" now....."
fastp -a AGATCGGAAGAGC -l 40 -w 4 -U --umi_loc read1 --umi_len 14 --
umi_prefix umi -i $File -o $Root'_prep.txt'
bowtie -v 2 --best -p 4 -S ZIKVab $Root'_prep.txt' | samtools view -buSh - |
samtools sort -@ 4 - -o $Root'_mapping_sort.bam'
samtools index $Root'_mapping_sort.bam'
samtools mpileup -f ZIKV_DKR.txt -d 1000000 $Root'_mapping_sort.bam' >
$Root'_pileup.txt'
```

From here two python scripts, both written by Dr. Yiyang (Tommy) Zhou in Dr. Routh's lab, determine mis-matches and error rates within the data using mpileup in *samtools*, described previously (*63*). This method has been published in Nucleic Acids Research (2019) where validation of the error rates for other transitions/transversions compared to the rate of T to C mutations is presented.

Chapter 3 Investigating changes in poly-adenylation patterns and flavivirus packaging using PAC-seq and vPAR-CL respectively

This chapter is largely lifted from a published study, cited below I wrote the rough draft of the manuscript and had help with revisions from Dr. Andrew Routh. I extracted RNA, prepared RNA libraries, did all of the computational analysis, and prepared all of the figures. The cell culture and infections were carried out by John Chen in Dr. Pei-Yong Shi's lab at UTMB.

Sotcheff, S.; Elrod, N.; Chen, J.; Cao, J.; Kuymuycu-Martinez, M.; Shi, P-Y.; Routh, A. Zika virus infection alters gene expression and poly-adenylation patterns in placental cells. (*in Preparation*)

Although many flavivirus infections are asymptomatic, symptoms can range from febrile illness, to hemorrhagic fever, encephalitis or even death. These viruses are an increasing public health risk with geographic expansion of Aedes mosquitoes caused by global warming (*3, 4, 164*). In addition to being transmitted via mosquito bite, ZIKV can be transmitted both sexually and from mother to child in the first trimester of pregnancy (*9, 165*). This may result in congenital Zika syndrome and, in some cases, the severe impairment of neural development and microcephaly (*27, 33, 79, 166*). Microcephaly in infants born to infected mothers was associated with the 2015-2016 outbreak in South America (*2, 29*). Additionally, studies after this outbreak have shown that ZIKV that can persistently infect the testes, meaning that males can pass the virus on to their partners for months after infection (*29*). Although ZIKV does not infect as many individuals as even other contemporary flaviviruses, it is important to consider the additional transmission routes and how this might affect public health in the future. Many have begun trying to understand how ZIKV might be able to cross the maternal-fetal interface (*33, 35, 76*). These studies may provide insight

into flavivirus pathogenesis that may aid in development of vaccines or antivirals in the future, not just against ZIKV, but perhaps other flaviviruses as well as others have also been shown to infect placental tissue.

STUDY EXAMINING TRANSCRIPTOMIC CHANGES IN RESPONSE TO ZIKV INFECTION

We infected either Huh7 or JEG3 with ZIKV or DENV as highlighted above and used total cellular RNA to generate PAC-seq libraries. These short-read libraries were sequenced using an Illumina NextSeq 550 and each sample had ~12M reads **(Table 3.1)**. After de-multiplexing, the data was analyzed using *DPAC (158)*, which uses *DESeq2* to analyze changes in gene expression and simultaneous investigation of alternative poly-adenylation (APA) as a result of flavivirus infection (*158*). The *DESeq2* output provides a table including gene names, log2FoldChange, p-value and adjusted p-value, and normalized gene counts across the samples in treatments being compared. To consider a gene as differentially expressed we used the cut-offs of adjusted p-value (we use adjusted because there are multiple hypotheses across the dataset) of less than 0.1 (<0.1) and an absolute value of the Fold Change greater than 1.5x (log2FoldChange greater than 0.585). A positive log2FoldChange indicates up-regulation and a negative down-regulation, the cut-off of 0.585 indicates a 50% increase or decrease in expression respectively. Additionally, the *DPAC* output tables indicated the percent distal usage (PDU) of poly-A clusters (PACs) or the percent of reads for a particular gene that use the furthest poly-A cluster(PAC, resulting in a longer 3' untranslated region or UTR). We also can distinguish up- or down-regulated PACs using *DPAC*. We extract the sequences for regions surrounding these PACs to perform motif enrichment analysis.

Table 3.1: Reads per PAC-seq samples for ZIKV and DENV infections of Huh7 or JEG3 cells.

Replicate	Cell Type	Virus	Time (hr)	Infected Y/N	Total Reads
1	Huh7	ZIKV PRVABC59	16	N	8,754,648
2	Huh7	ZIKV PRVABC59	16	N	8,190,547
3	Huh7	ZIKV PRVABC59	16	N	9,993,595
1	Huh7	ZIKV PRVABC59	16	Y	9,935,138
2	Huh7	ZIKV PRVABC59	16	Y	14,141,682
3	Huh7	ZIKV PRVABC59	16	Y	5,519,851
1	JEG3	ZIKV PRVABC59	16	N	18,762,126
2	JEG3	ZIKV PRVABC59	16	N	16,704,921
3	JEG3	ZIKV PRVABC59	16	N	15,108,156
1	JEG3	ZIKV PRVABC59	16	Y	12,888,212
2	JEG3	ZIKV PRVABC59	16	Y	16,086,782
3	JEG3	ZIKV PRVABC59	16	Y	15,593,261
1	JEG3	DENV2	16	N	10,123,821
2	JEG3	DENV2	16	N	9,931,289
3	JEG3	DENV2	16	N	9,887,112
1	JEG3	DENV2	16	Y	12,109,949
2	JEG3	DENV2	16	Y	14,181,345
3	JEG3	DENV2	16	Y	10,741,610

ZIKV INFECTION OF HUMAN LIVER (HUH7) CELLS

PAC-seq data after *DESeq2* analysis yielded 11,215 genes to compare mock to ZIKV infected Huh7 cells. Of those 83 genes were differentially expressed (40 down- and 43 up-regulated) with a p-adj <0.1 and |log2FC| greater than 0.585 and the distributions of genes based on their p-adjusted and log2FC values are shown in the volcano plot in **Figure 3.1**. We confirmed the data quality was sufficient by generating and loading bedgraph files (a format that allows display of continuous-valued data, or coverage, in a track in genome browsers) into the UCSC Genome Browser (*31*) as shown in **Figure 3.2** for genes encoding CHD1 and STK16, a chromatin remodeling protein and serine/threonine kinase, respectively. *Enrichr* (*32-34*) was used to identify enrichment of different pathways, ontologies, or phenotypes in the 83 genes differentially expressed (**Figures 3.3, 3.4**). The results suggest that ZIKV infection in Huh7 cells hinders matrix metallo-proteinases and the electron transport chain and promotes the interferon alpha signaling pathway and the use of the internal ribosome entry pathway.

DPAC identified 1,592 genes with multiple PACs in the ZIKV infected Huh7 cells. **(Fig. 3.5)** For our analysis we focused on changes of at least 20% in PDU. To confirm we uploaded the generated bedgraph files into the genome browser and viewed individual genes that were identified has having APA.(*161*) Bedgraph files confirmed mapping to the human genome at the 3' end of genes and indicated that U2SURP (U2 snRNP associated protein involved in splicing) and GNS (an acetylglucosamine sulfatase) undergo shortening and lengthening of their 3' UTRs respectively. **(Fig. 3.6)** Our analysis found 182 3' UTRs shortened and 198 lengthened 3' UTRs. As evidenced by the *Enrichr* Scatterplot plots in **Figure 3.7** (blue) there is shortening of transcripts involved in cell adhesion, cell cycle, and pre-mRNA processing.(*167-169*) In addition, these scatterplots show

Figure 3.1: Volcano plot highlighting differentially expressed genes (DEGs) in human liver (Huh7) cells in response to ZIKV infection. Each point is a gene. Up- (orange) and down-regulated (purple) genes have a p-adjusted value <0.1 and log2FoldChange of >0.585 or <-0.585 respectively. DEGs are labelled as space allows. Plot created by S. Sotcheff using R under license.

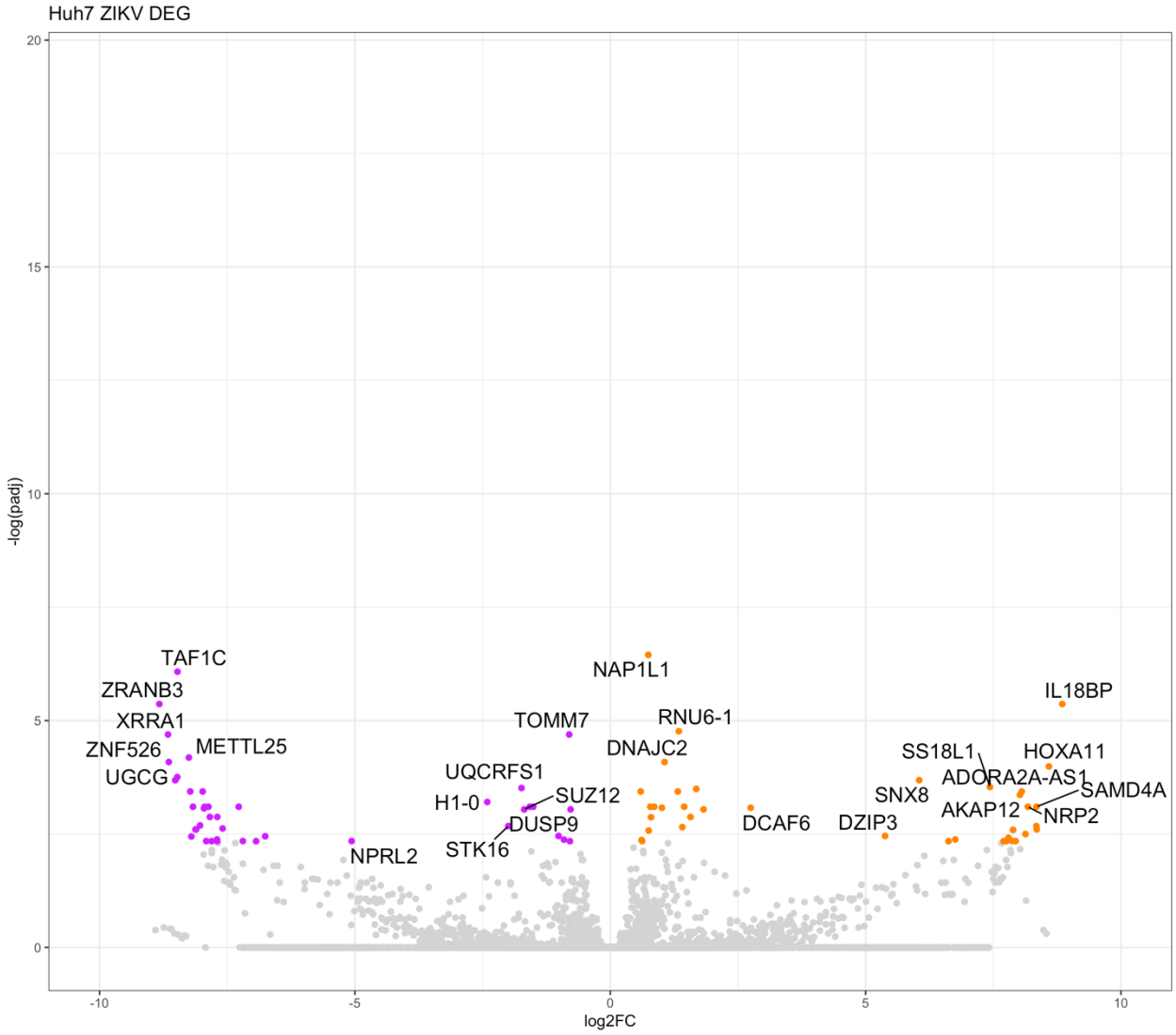


Figure 3.2: Alignment of bedgraph files for representative samples the ZIKV Huh7 data at one up-regulated (CHD1) and one down-regulated (STK16) gene. As PAC-seq primes from the poly-A tail and cDNA was stochastically terminated in the reverse transcription step, we expect coverage to be highest at poly-A cluster (PAC) and to taper off as we extend into the gene.

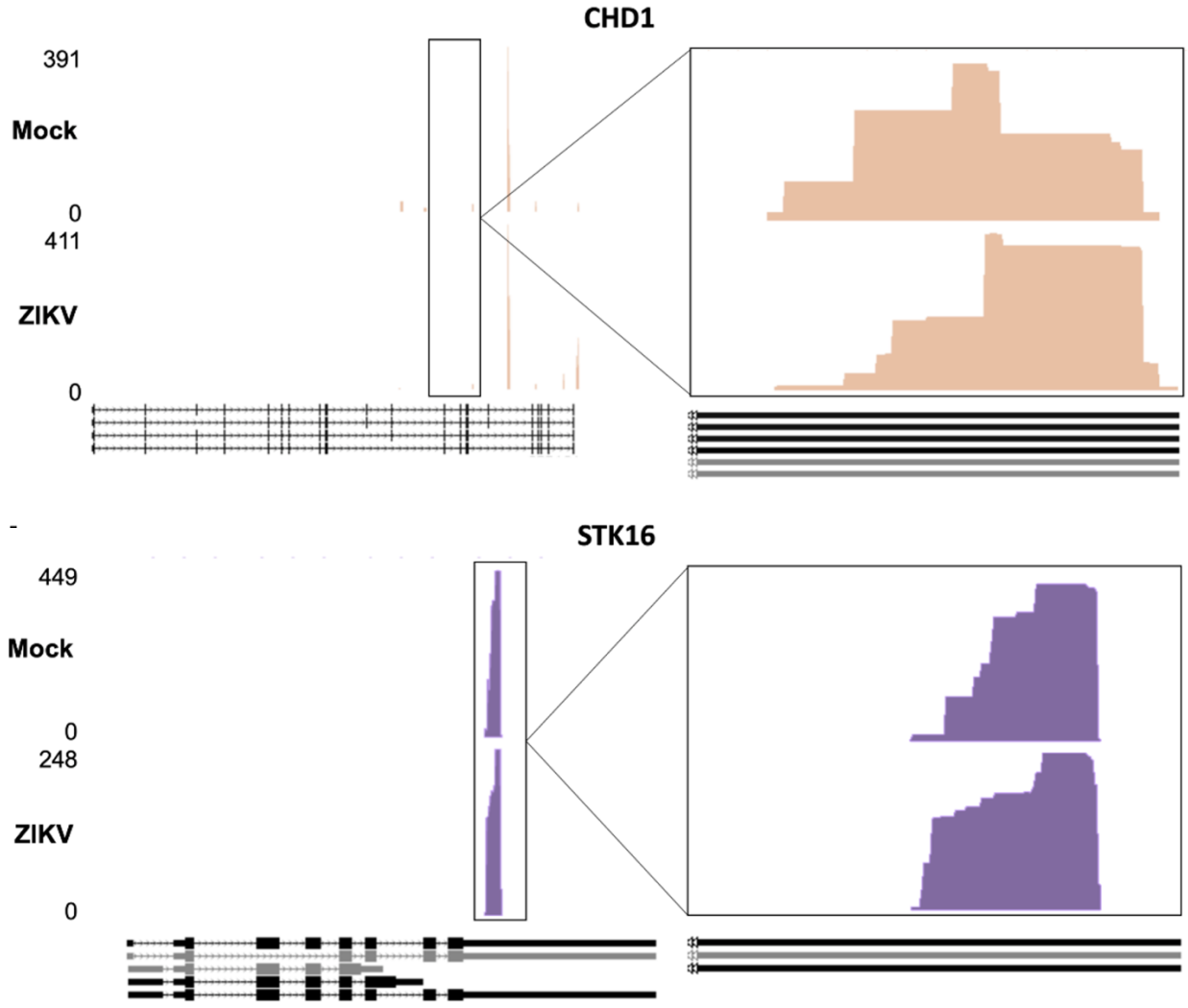


Figure 3.3: *Enrichr* (*167-169*) scatterplot showing up-regulated pathways (orange). Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Up-regulated pathways (p-value <0.05) are labelled as space allows.

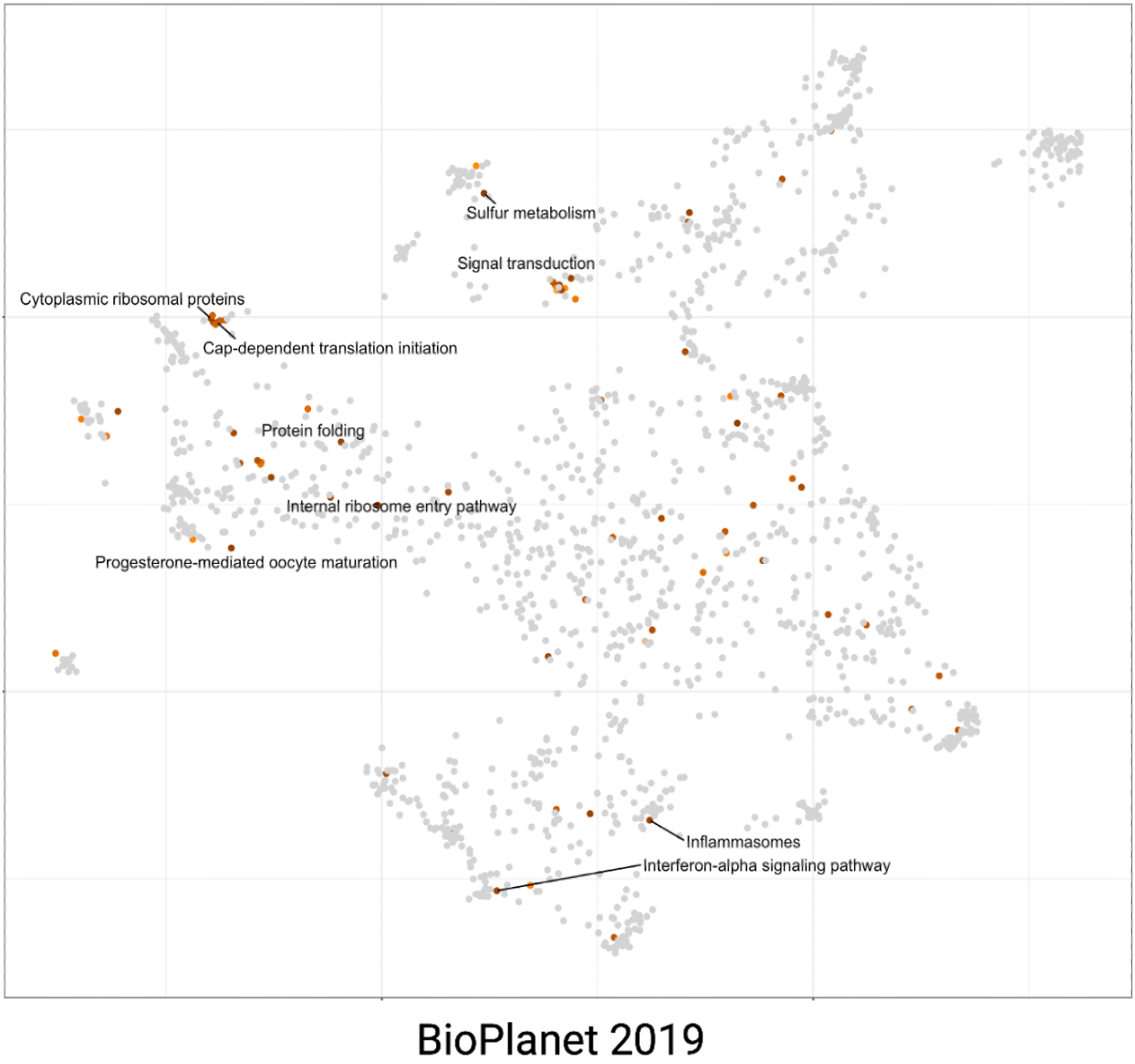


Figure 3.4: *Enrichr* (*167-169*) scatterplot showing down-regulated pathways (purple). Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Down-regulated pathways (p-value <0.05) are labelled as space allows.

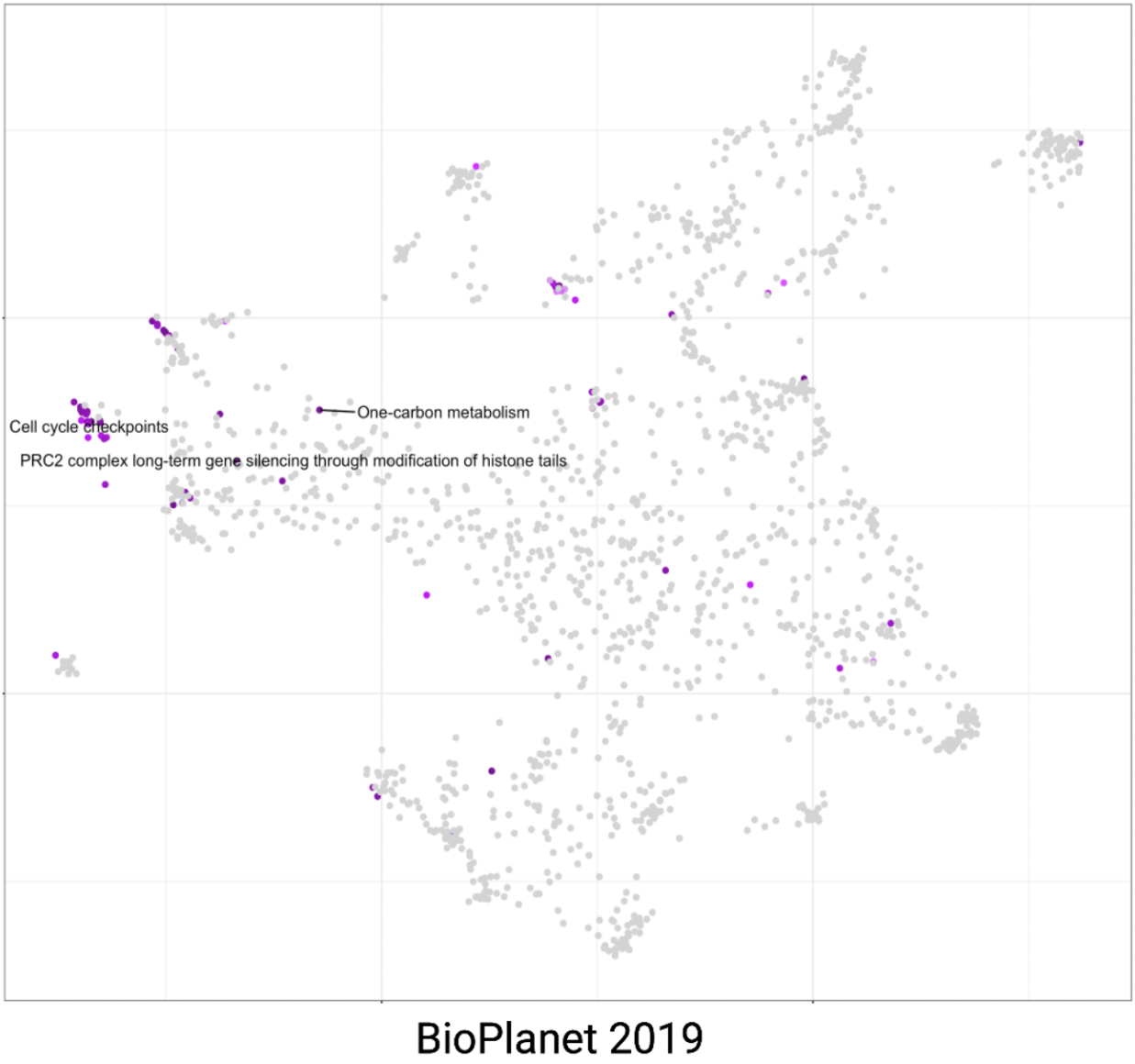


Figure 3.5: Plot depicting changes in percent distal usage (PDU) of poly-A clusters (PACs) between the mock-infected and ZIKV-infected human liver (Huh7) cells. Each point is a gene, those colored in yellow have an increase in PDU of 20% or more and those in blue have a decrease in PDU of 20% or more in response to ZIKV infection. These colored points (genes) have alternative poly-adenylation (APA) and are labelled as space allows. Plot created by S. Sotcheff using R under license.

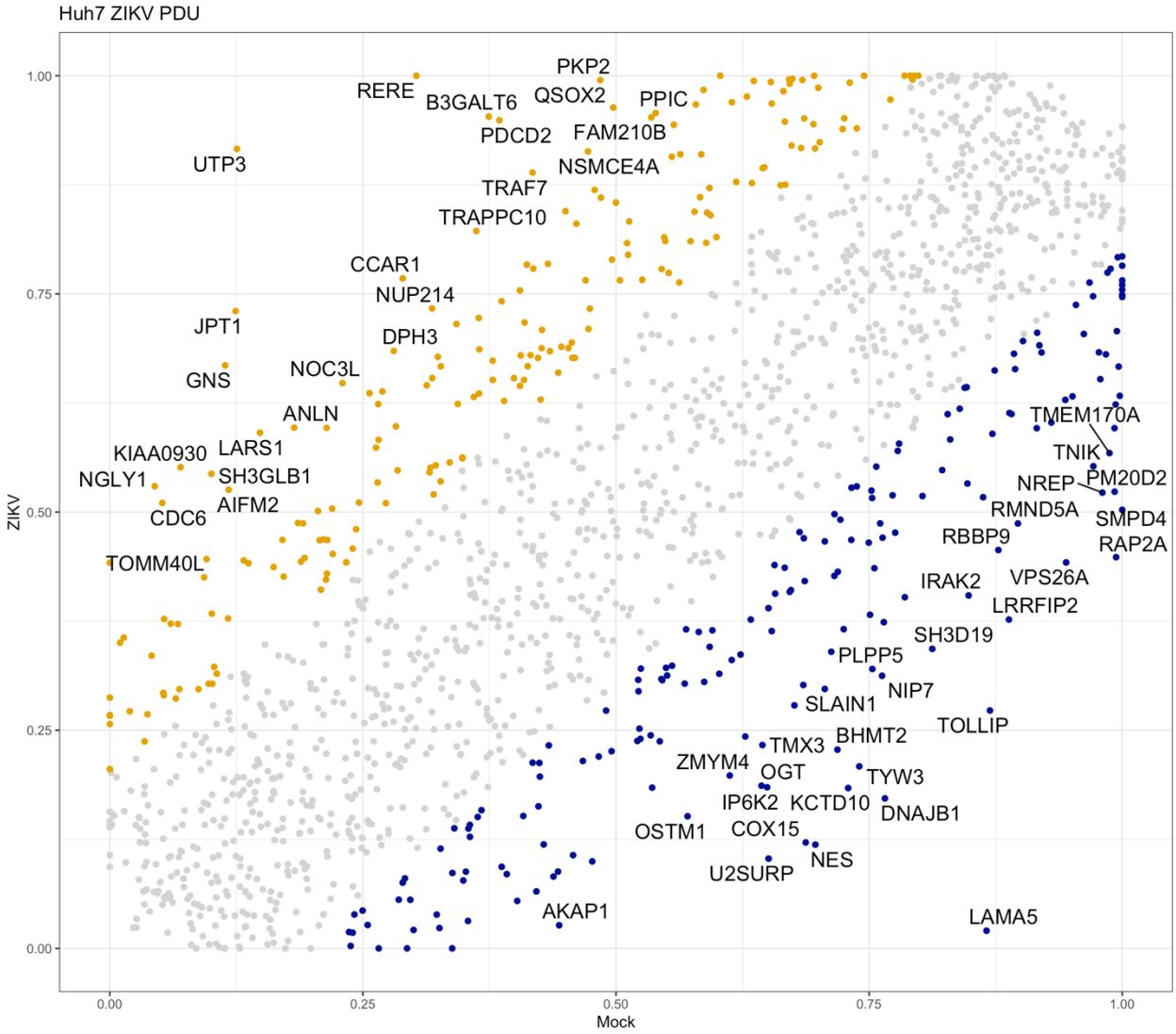


Figure 3.6: Alignment of bedgraph files for representative samples the ZIKV Huh7 data at one gene with a lengthened 3' UTR (GNS) and one with a shortened 3' UTR (U2SURP).

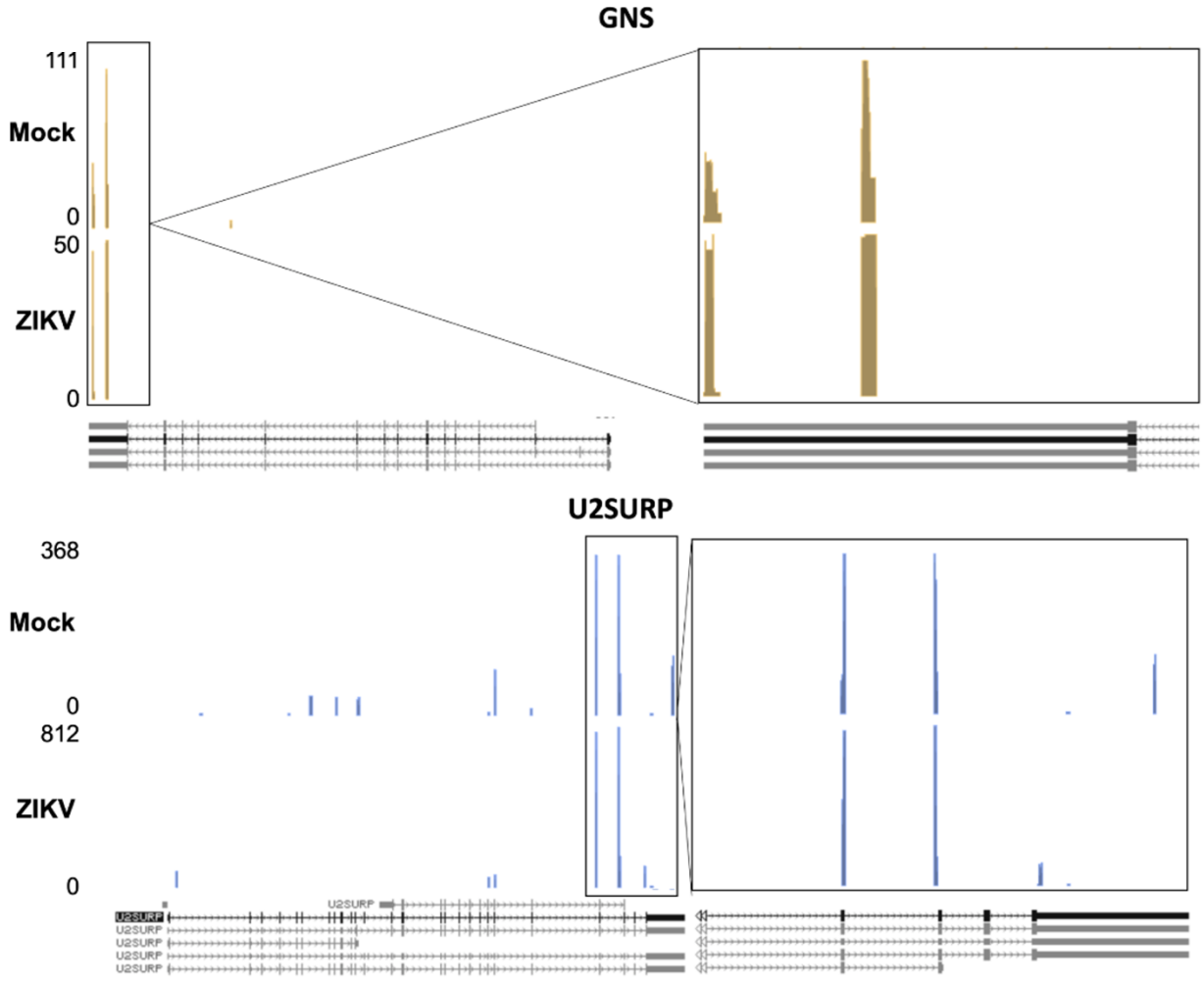
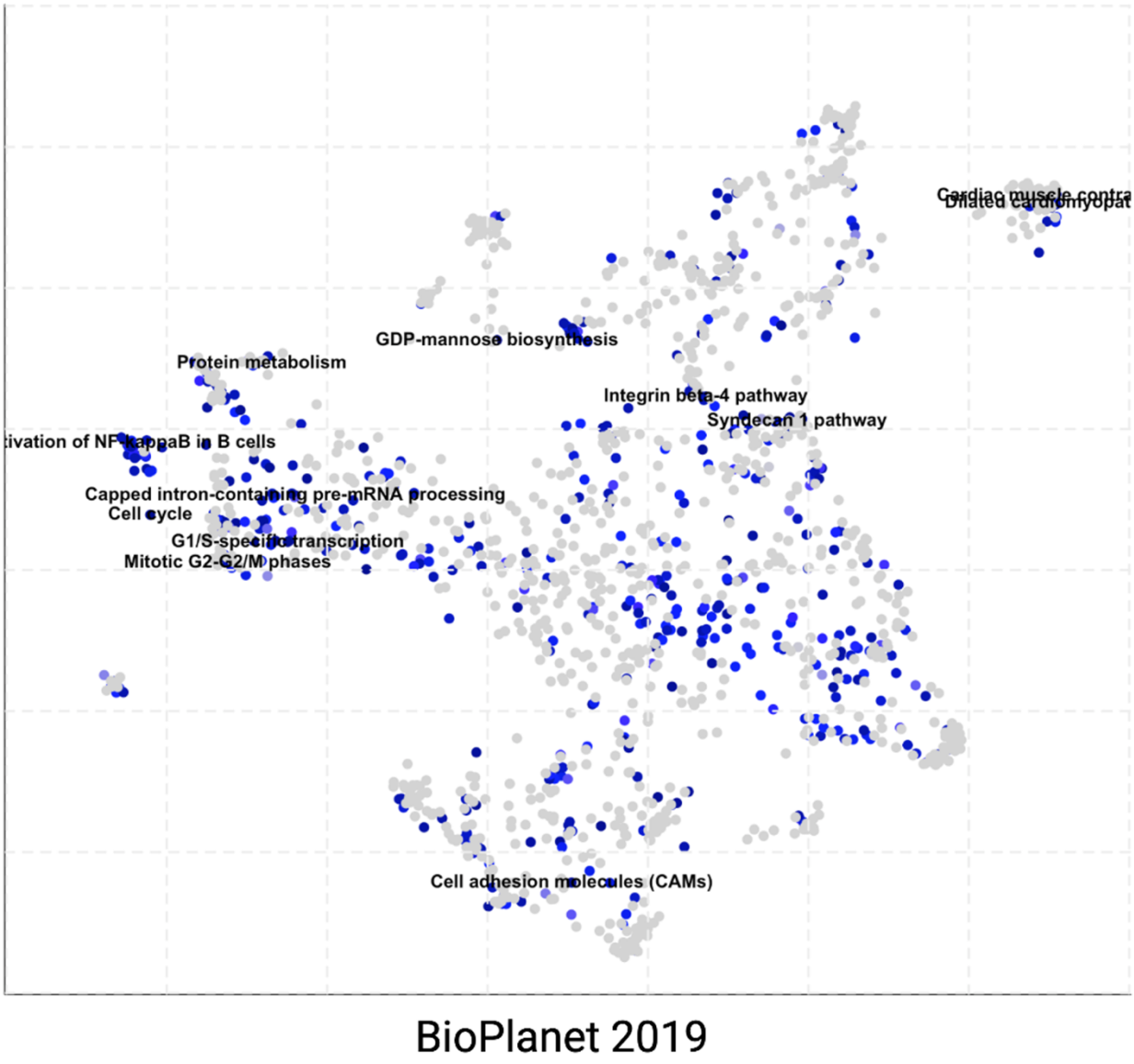


Figure 3.7: *Enrichr* (*167-169*) scatterplot showing pathways that were enriched for shortened 3' UTR (blue) in Huh7 upon ZIKV infection. Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Pathways enriched for 3' UTR shortening (p-value <0.05) are labelled as space allows.



lengthening (yellow, **Figure 3.8**) of transcripts involved in mRNA degradation and splicing (minor pathway).

We extracted sequences of varying lengths at up- and down-regulated PACs as described in Chapter 2. DREME (*159*), an online tool for identifying enriched motifs, was used to determine if any short sequences were enriched at these PACs. These enriched motifs are shown in **Table 3.2.** Further, these motifs were cross-referenced with RBPDB (*160*) to determine RNA binding proteins that may interact near these up- or down-regulated PACs. Interestingly, there were no enriched motifs within 100 nts of either the up- or down-regulated PACs, this trend extends to 250 nts for the up-regulated PACs. Within 500 nts of the up-regulated PACs we found AGGCCGAG and CAGCCAC, recognized by YBX and cytoplasmic poly-A binding protein 1 (PABPC1) respectively. When we extend to 1000 and 2000 nts from the up-regulated PACs we also found enrichment for AGCCACG (recognized by PCBP1), GAGACAG (recognized by PCBP2) as well as motifs identified by SRSF proteins 5 and 9. Similarly, when we extended the parameters for the down-regulated PACs we found enrichment of motifs recognized by PTBP1, PCBP2, and SRSF7. These results suggest that Huh7 genes with up- or down-regulated PACs in response to ZIKV infection are roughly equally likely to interact with splicing factors.

ZIKV INFECTION OF HUMAN PLACENTAL (JEG3) CELLS

DESeq2 analysis yielded 11,697 genes to compare mock to ZIKV infected JEG3 cells. Of those 126 genes were differentially expressed (28 down- and 98 up-regulated) with a p-adjusted value (p-adj) of <0.1 and an absolute value of log2 fold change (|log2FC|) greater than 0.585 (or minimum of 50% increase/decrease) and the distributions of genes based on their p-adj and log2FC are shown in the volcano plots in **Figure 3.9**. We confirmed the data quality was sufficient by

Figure 3.8: *Enrichr* (*167-169*) scatterplot showing pathways that were enriched for lengthened 3' UTR (yellow) in Huh7 cells upon ZIKV infection. Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Pathways enriched for 3' UTR lenthening (p-value <0.05) are labelled as space allows.

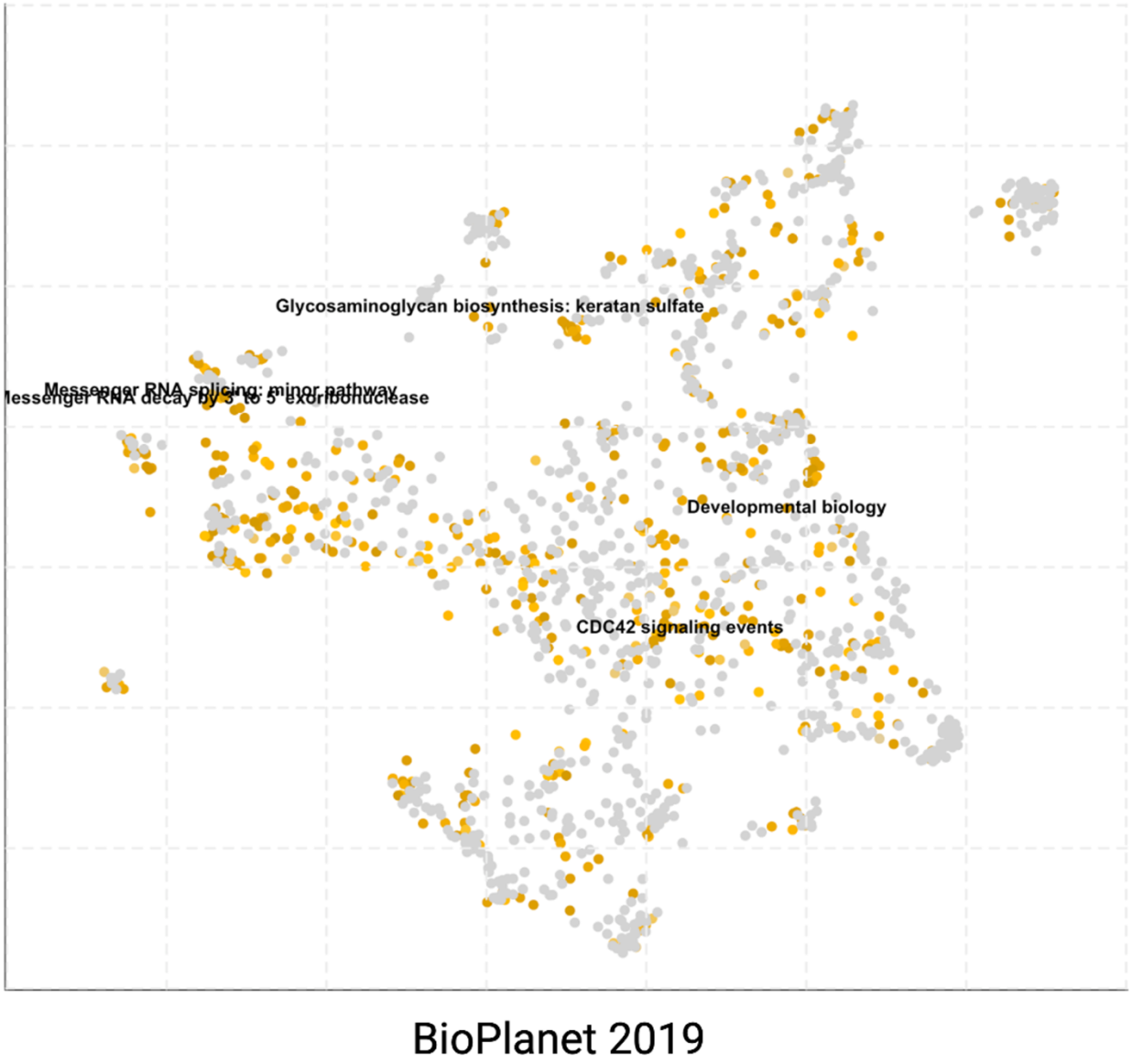
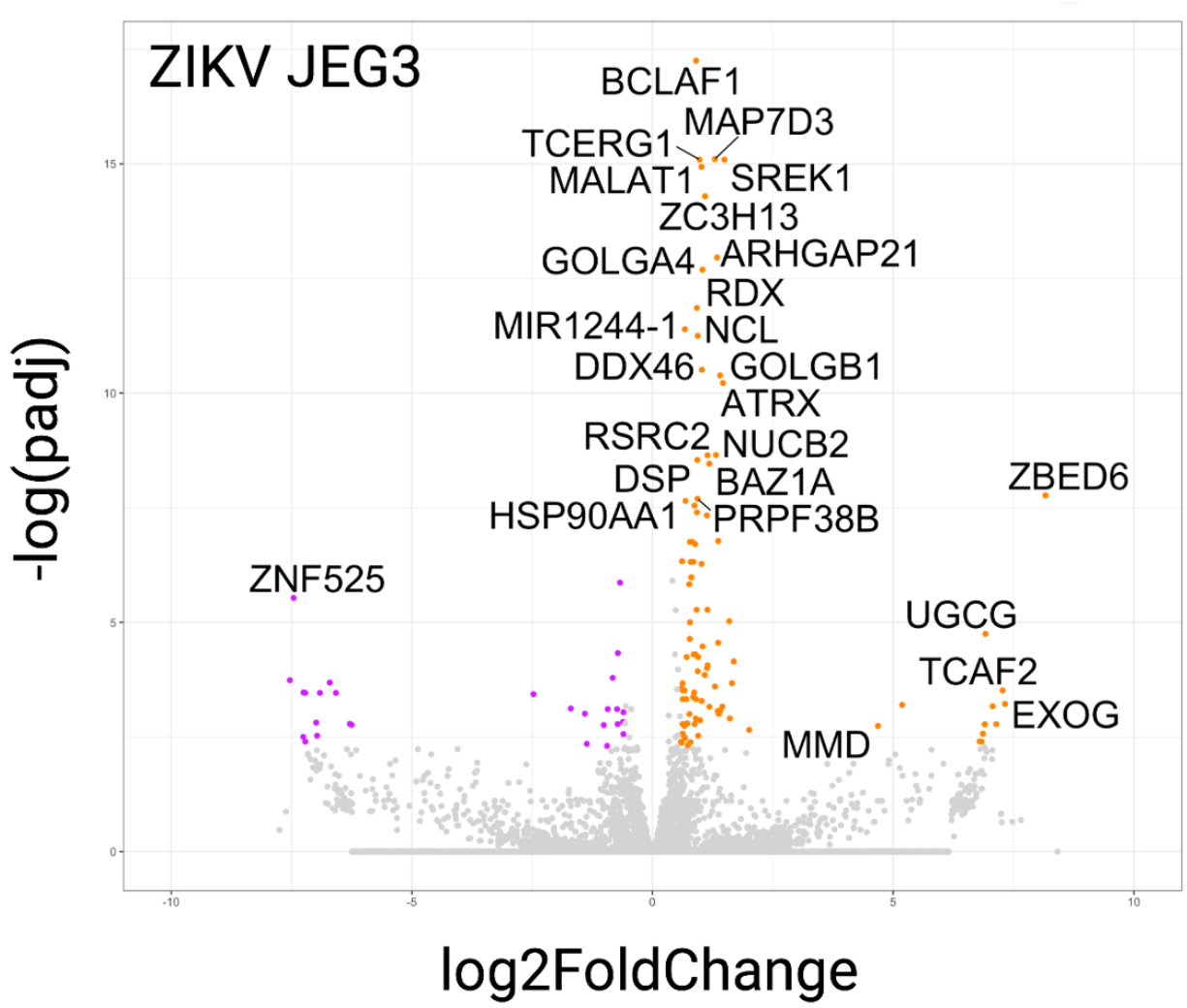


Table 3.2: Motifs enriched with 'X' nucleotides of up- or down-regulated poly-A clusters in human liver (Huh7) cells upon ZIKV infection.

Condition/ Nts from PAS	100	250	500	1000	2000		KEY
UP	None	None	CAGCCWC	AGCCTGGG	CCYCAGCC	AAAATTAG	R=purine (AG)
			GKCAGGA	AGCCACHG	CCAGYTAC	CTTGAACC	Y= pyrimidine (CT)
			CAGSCTGG	TACAGGCR	CAGGCTGG	AGCAAGAC	W = weak (AT)
			AAAATADT	CTGAGGCA	CRGTGAGC	CCACCTCM	S = strong (CG)
			ATCCCAGC	CTCACTGC	CTGTAATC	ATATATAM	K = keto (GT)
			AGGCSGAG	CCTCCRCC	GAGAYSGA	TACAGGCR	M = amino (AC)
			ARAAAWTA	GAGACAS	AGCCACYG	CWGCCTCC	U/T interchangeable
				GACCAKCC	AAWATACA	ACTGTGYT	D= not C
				CACTTTGG	AGGCSGAG	TATTTTTW	V = not T
				AAAATACA	GTGGTGGY	GAGAWGAA	H = not G
				CTCCTGAC	AAACCCCR	CAGGAGAA	B = not A
				CCCAGCTA	CTGTCDCC	GTGATCCR	N = any base
				CACTGCAC	CYGACCTC	GTTCGAGA	X = ACGT = .
				GGTTTCAC	AAGTGCTG	RGAGAGAA	
				CCACCWCG	GCYGGGCR	AAATAAAW	
				AAAAHAAA	TGCAGTGR	CCTSTCTC	
				TTTTAAAA	TATAYATA	GSCCAGGM	
				ATTTATW	ACAYGGTG	CGAGATCR	
				CTCCCRAG	CTCTACTA	AAATAYTT	
				ATCTCRG	AGHGAGAC		
				CCTCCTS	AAWAAAAA		
DOWN	None	ATTACAR	GCTGGGW	CTGTAATC	GTKAGCCA	AGTAGCTR	
			ACTGGR	AYTCCAGC	CKCCTGCC	CGTGAKC	
				CYGTCTC	GGAKTACA	ATATGTAY	
				GCCSAGGC	CAGRCTGG	TCACTGCA	
				TTTTAAWA	GAGGCVGA	CACYGCAC	
				CCACCKC	AAWATACA	AGCYTCCC	
				GCAGTGGC	GCCACCAY	CWCCTGAC	
				GCTGGGW	AGAGACR	AGAGRTGG	
				CAGTGAK	AGAGMAAG	GTGTGTGK	

Figure 3.9: Volcano plot highlighting differentially expressed genes (DEGs) in human placental (JEG3) cells in response to ZIKV infection. Each point is a gene. Up- (orange) and down-regulated (purple) genes have a p-adjusted value <0.1 and log2FoldChange of >0.585 or <-0.585 respectively. DEGs are labelled as space allows. Plot created by S. Sotcheff using R under license.



generating and loading bedgraph files into the UCSC Genome Browser.[161] We expected that reads would 1) map to the 3' end of genes and 2) reads would trail off as they were further from the 3' end because different reads would be of varying lengths (a block would potentially indicate PCR duplication of a single strand/read). These criteria were met as shown in **Figure 3.10** for genes encoding NFX1 and NPIPA2, a nuclear transcription factor that binds x-box motifs and a nuclear pore complex interacting protein, respectively. *Enrichr* (*167-169*) was used to identify enrichment of different pathways, ontologies, or phenotypes in the 126 genes differentially expressed (**Figures 3.11 & 3.12**). The results suggest that ZIKV infection in JEG3 cells hinders cell cycle control, the ERAD pathway and IL-7 interactions and promotes mRNA decay and processing (as well as splicing) and the IL-2 pathway.

DPAC identified 3,356 genes with multiple PACs in the ZIKV infected JEG3 cells. **(Fig. 3.13)** For our analysis we focused on changes of at least 20% in PDU. To confirm we uploaded the generated bedgraph files into the genome browser and viewed individual genes that were identified has having alternative poly-adenylation (APA), determined with our 20% PDU change cut-off.[161] Bedgraph files confirmed mapping to the human genome at the 3' end of genes and indicated that PTER (a gene producing a phosphotriesterase related protein) and SURF4 (an integral membrane protein that interacts with ER-golgi intermediate compartment proteins) do indeed undergo shortening and lengthening of their 3' UTRs respectively. **(Fig. 3.14)** Our analysis found 269 3' UTRs shortened and 229 lengthened 3' UTRs.

Generally, shortening of the 3' UTR reduces availability of functional elements for RNA binding proteins, micro-RNAs, and long non-coding RNAs to bind.. After plugging the lists of genes with PDU changes of greater than +/- 20% it is evident that there are a number of genes that are shortened that are involved in post-transcriptional modifications of mRNA. As evidenced by

Figure 3.10: Alignment of bedgraph files for representative samples the ZIKV JEG3 data at one up-regulated (NFX1) and one down-regulated (NPIPA2) gene. As PAC-seq primes from the poly-A tail and cDNA was stochastically terminated in the reverse transcription step, we expect coverage to be highest at poly-A clusters (PACs) and to taper off as we extend into the gene. This is indeed the case.

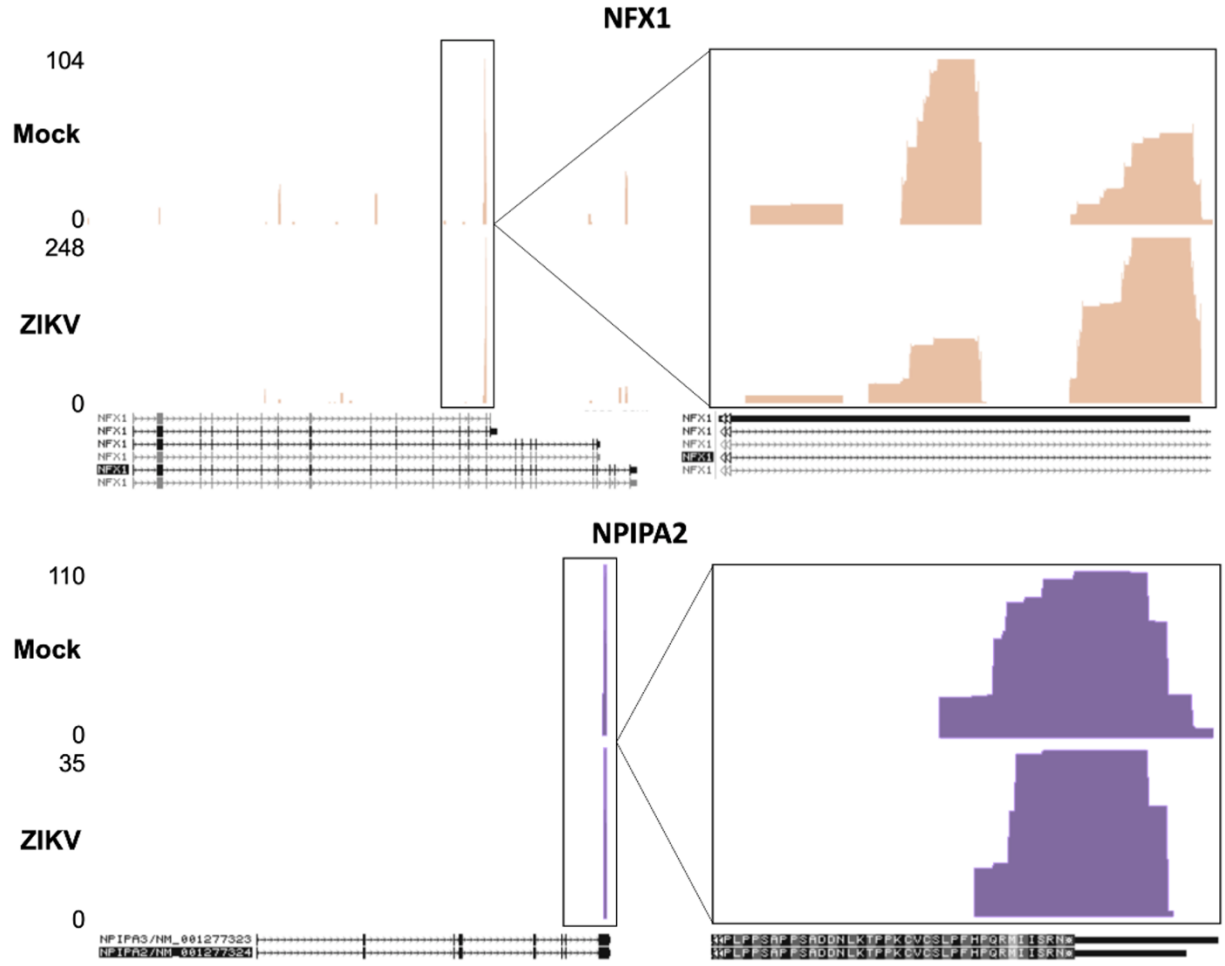


Figure 3.11: *Enrichr* (*167-169*) scatterplot showing up-regulated pathways (orange) in JEG3 cells in response to ZIKV infection. Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Up-regulated pathways (p-value <0.05) are labelled as space allows.

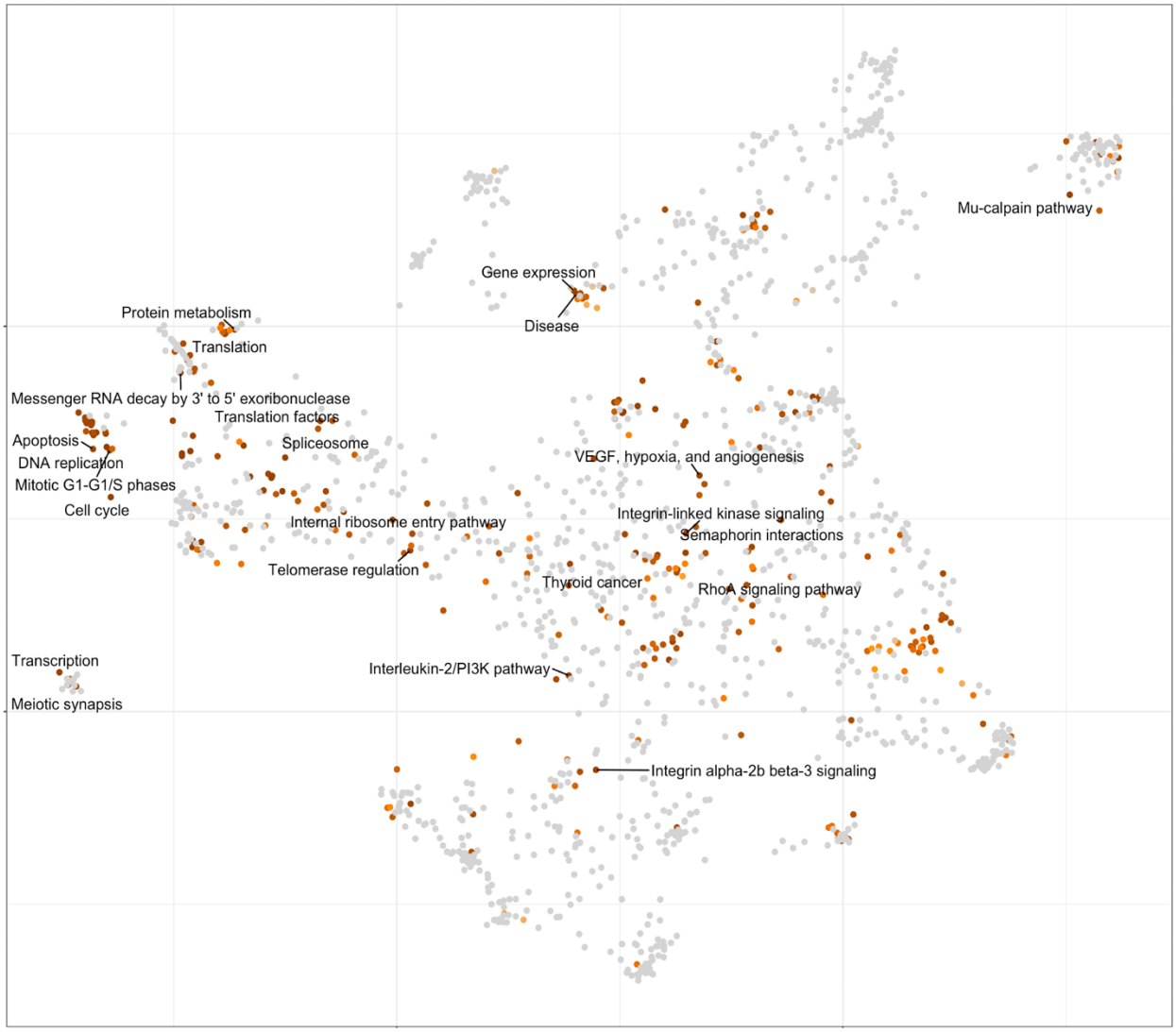


Figure 3.12: *Enrichr* (*167-169*) scatterplot showing down-regulated pathways (purple) in JEG3 cells in response to ZIKV infection. Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Down-regulated pathways (p-value <0.05) are labelled as space allows.

Figure 3.13: Plot depicting changes in percent distal usage (PDU) of poly-A clusters (PACs) between the mock-infected and ZIKV-infected human placental (JEG3) cells. Each point is a gene, those colored in yellow have an increase in PDU of 20% or more and those in blue have a decrease in PDU of 20% or more in response to ZIKV infection. These colored points (genes) have alternative poly-adenylation (APA) and are labelled as space allows. Plot created by S. Sotcheff using R under license.

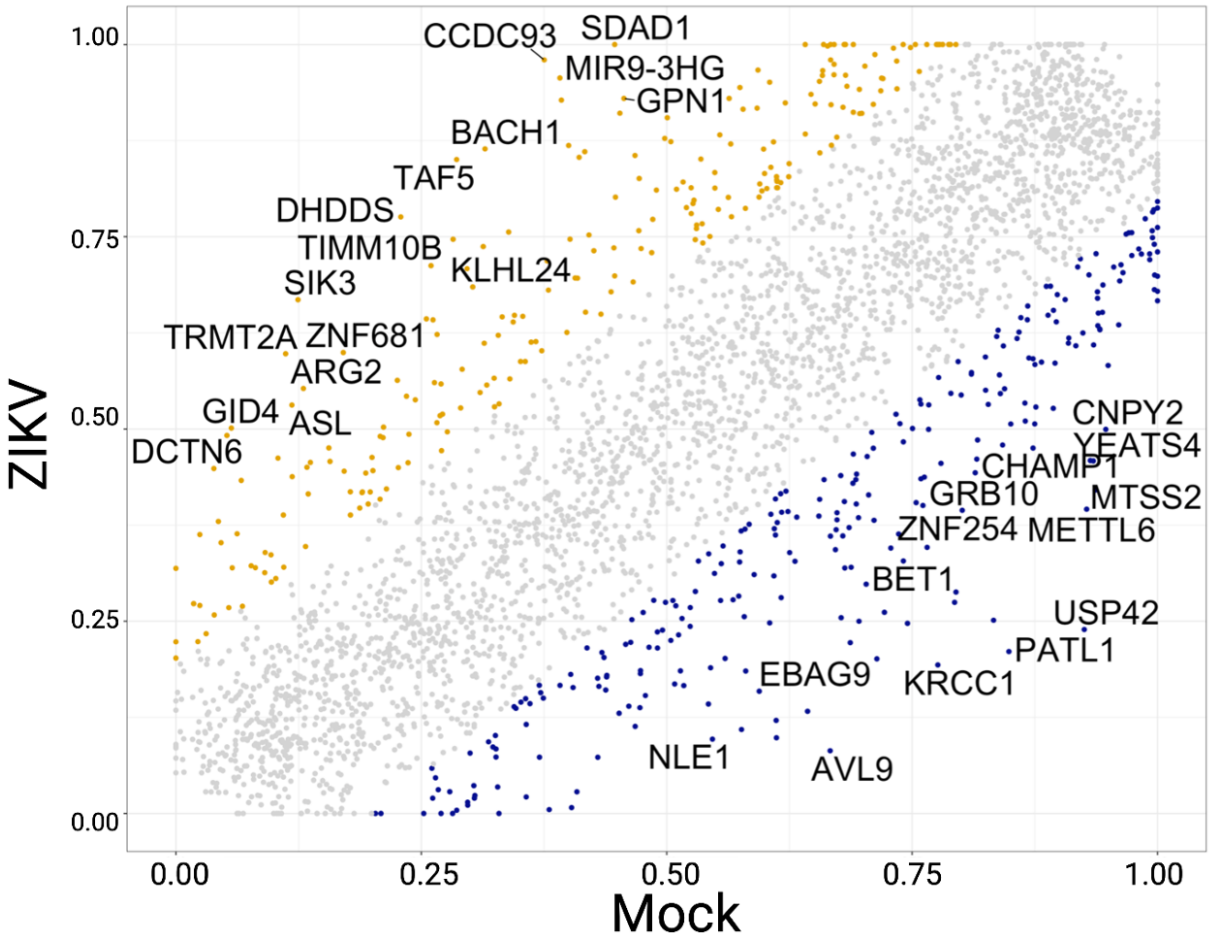
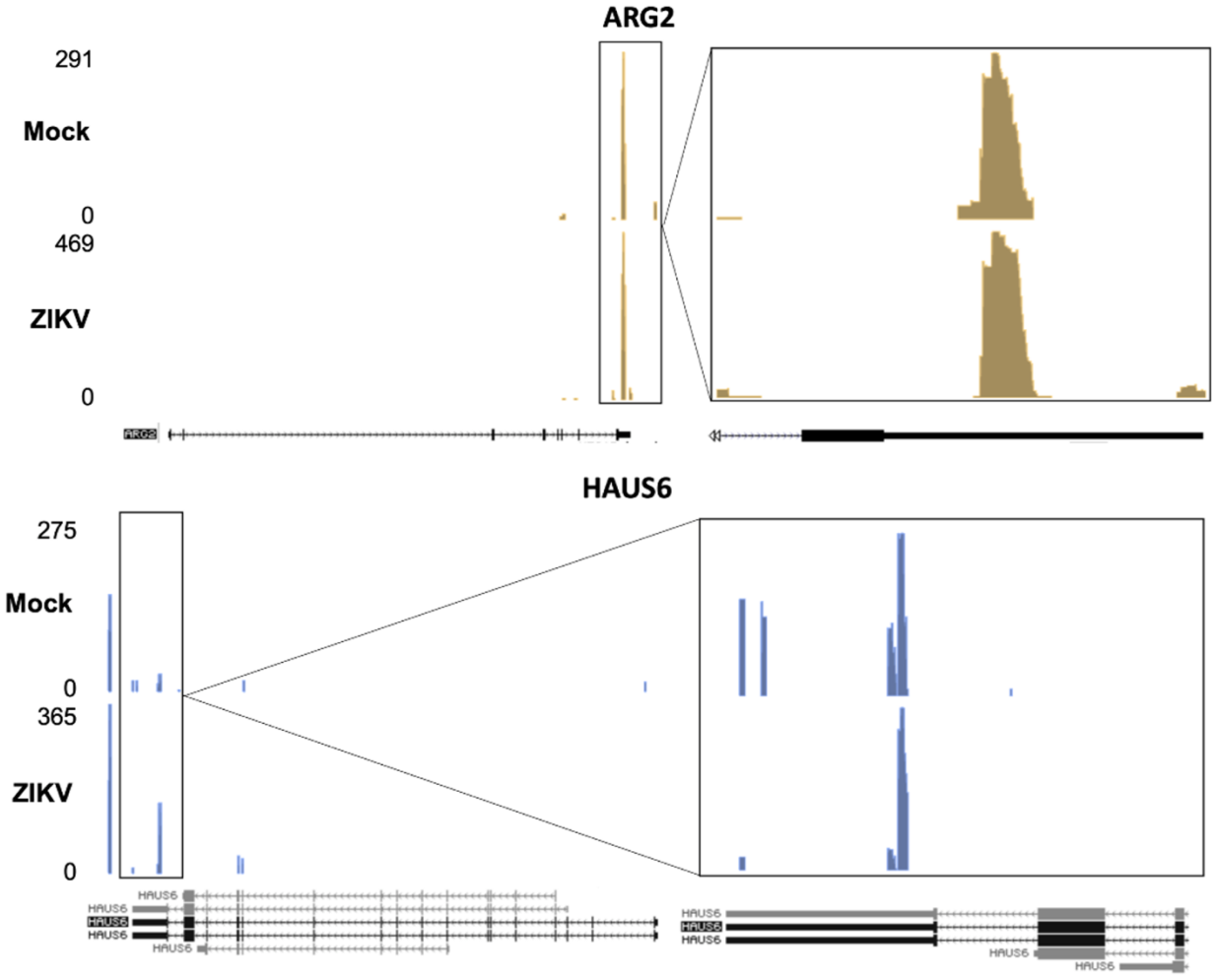


Figure 3.14: Alignment of bedgraph files for representative samples the ZIKV JEG3 data at one gene with a lengthened 3' UTR (ARG2) and one with a shortened 3' UTR (HAUS6).



the *Enrichr* Scatterplot plots (where each point is a gene set, and colored points are gene sets enriched with a p-value < 0.1) in **Figure 3.15** (blue) there is shortening of transcripts involved in siRNA biogenesis.[167-169] In addition, these scatterplots show lengthening (yellow, **Figure 3.16**) of transcripts involved in differentiation, hedgehog and Wnt signaling, and B cell receptor signaling.

Additionally, we wanted to investigate if any RNA binding protein motifs were present in proximity to the differentially used PACs. To do this we first extracted the sequences at various distances (100, 250, 500, 1000, and 2000 nts) surrounding PACs that were up-regulated or preferentially used and down-regulated for preferentially not utilized upon ZIKV infection in these JEG3 cells. Then we input these sequences into Discriminative Relative Expression Motif Elicitation (DREME) (*159*) to determine if there were any motifs enriched in these sequences. Further, these motifs were cross-referenced with RBPDB (*160*) to determine RNA binding proteins that may interact near these up- or down-regulated PACs. Within 100 nt of the up-regulated PACs we found the motif GGAAGAA which is part of the motif for various splicing factors and HNRNPs, but we did not find enrichment of the canonical motif for poly-A binding protein 1 (PABP1, motif AAUAAA). This motif, however, was found enriched in the sequences within 100 and 250 nt of the down-regulated PACs, about ~20 nt upstream of the PAC as expected (**Figure 3.17A**) as visualized using CentriMo (*170*). CentriMo is part of the MEME suite, and shows enrichment not just of specific sequences but also in a specific location within a set of sequences. In the plot we also see some enrichment, although not to the same extent, of the KKKKKK motif (a GU rich region) just downstream of the PAC. This indicates that the down-regulated PAC are using canonical poly-adenylation signals. Neither of these sequences are enriched in region surrounding the up-regulated PACs (**Fig. 3.17B**), suggesting an alternative mode for poly-

Figure 3.15: *Enrichr* (*167-169*) scatterplot showing pathways that were enriched for shortened 3' UTR (blue) in JEG3 cells upon ZIKV infection. Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Pathways enriched for 3' UTR shortening (p-value <0.05) are labelled as space allows.

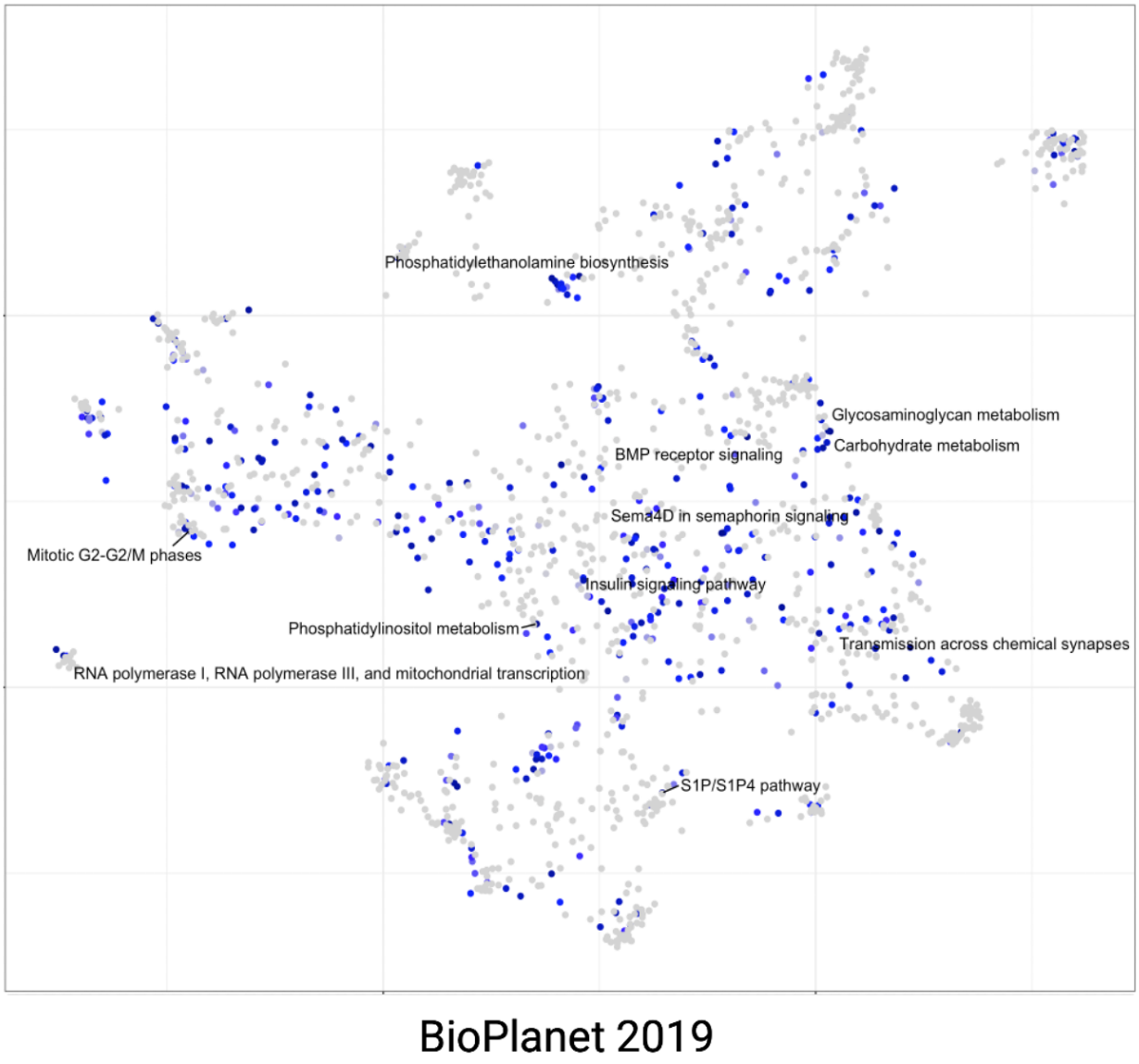
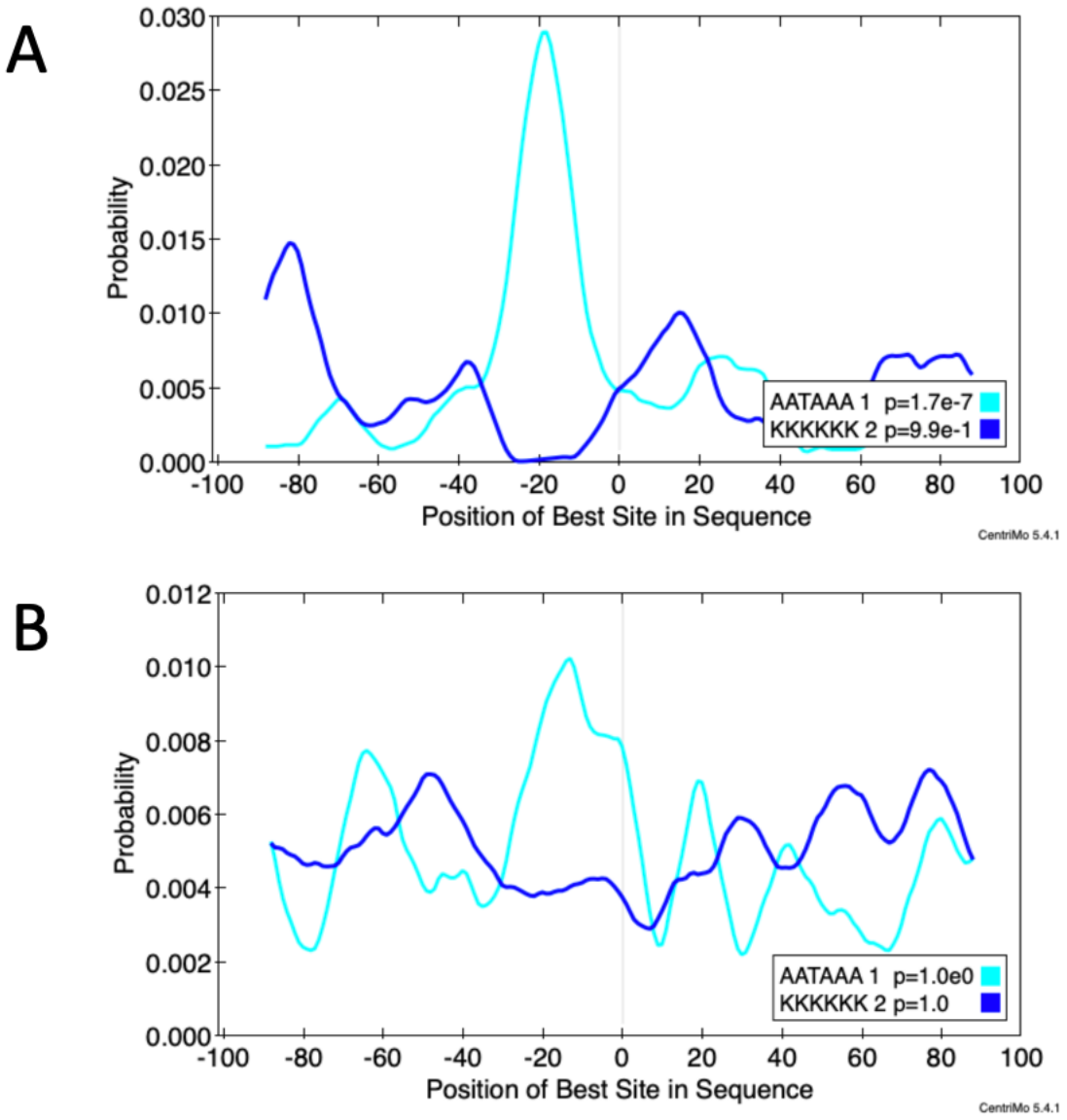


Figure 3.16: *Enrichr* (*167-169*) scatterplot showing pathways that were enriched for lengthened 3' UTR (yellow) in JEG3 cells upon ZIKV infection. Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Pathways enriched for 3' UTR lengthening (p-value <0.05) are labelled as space allows.

Table 3.3: Motifs enriched with 'X' nucleotides of up- or down-regulated poly-A clusters in human placental (JEG3) cells upon ZIKV infection.

Condition/Nts to PAS	100	250	500	1000		2000		KEY
UP	CATYTTC	RAAGRAAA	CCTCCCR	CTTGAACC	CRGCCTCC	CAGGCTGS	AACATGGY	R=purine (AG)
	GGHAGAA	CAGSMTG	AMAYACA	CRCCACCA	CCCRGCTA	GARGCTGA	TCACTGCA	Y= pyrimidine (CT)
	AAKGTA	GAAGAWGN	CAGGCTGG	ATAAAWAT	CCAGCCYG	CCCRGCTA	CACACAYA	W = weak (AT)
		ACAGAAAB	CACTGCAS	AGARGAAA	CCCASCAC	CTGYAATC	CTRCCTGC	S = strong (CG)
			ARAAGAAA	AGARAACA	AAAATAM	GTGAGCCR	AAWAATAA	K = keto (GT)
			GGAYTACA	AGCRAGAC	GTGAGCCR	CTCCTGAC	ARAAGAAA	M = amino (AC)
			CTCMTGCC	GAGAVAGA	CACTGCAC	GTGGTGGY	AAATRTGT	U/T interchangeable
			CTCYGTC	GAGAYGGA	AAAAAAWA	AYTACAGG	AGATCGC	D= not C
			CRCCACCA	CAGATCAB	AYTACAGG	GAGAYGGR	GAKCCACC	V = not T
			CTCCTGGR	ACATTTTY	GGTCAGGA	TATTTWTA	CTCYTCCC	H = not G
			AAATTAGC	AGGCCARG	GAGGHGGA	CSGCCTCC	CTRCCAC	B = not A
			GGAAAARA	ACAGARAT	GTAGAGAY	GGKGAAAC	AGCRAGAC	N = any base
			MAAATA	AAAABAAA	CTGYCWCC	GAGWGAGA	GCKTGAAC	X = ACGT = .
			TGCAGTGW	GAAACCCY	CCATGTTR	CCWCTGCA	CATATMC	
			GTTTTMCA	GGCCGGGM	GATCCDCC	CKCCTGCC	ACYCAGA	
			GAAGATGM	AGGYTGCA	AWATATG	CCTGTCTS	CCWCCACC	
			ACATGGTV	CCCYACCC	CRGTGGC	AAARTGCT	TAHTAAAA	
			ATRTATA	GAAGAKGA	CACACRCA	CTGTCWC	ATAACATY	
			AGTAGCWG	GCATGATS	GWTCGAGA	AMATACA	ACAGTATS	
			BGAGGCAG	TAATTTTW	CTCCTSCC	GCCCAGS		
				GAAATAC				
DOWN	AATAAA	AATAAA	RTATAAA	CAGGAGAA	CARCCTCC	CMGCCTCC	GGYCAGGA	
				CTSCAGCC	GCCACCRC	CTGTADTC	CCAGCTAC	
				RCTGAGA	GGAYTACA	ACTCCAKC	TGRTCTCR	
				CTCTACYA	AWATACA	GCCACCRC	CCGCCTYG	
						AGGCMGGA	GCAGTGGY	
						CAGTGAGC	GGTTTYAC	
						AGAGAYGG	GHGACAGA	
						CCCRGCTA	ATCACHTG	
						AGTGCTGG	GTRTGTA	
						AAWAAATA	AATAYAAA	
						CCTCCCAS	AAATATTT	
						GCRTGAGC	ARCCCAGG	

Figure 3.17: The ZIKV down-regulated poly-A clusters (PACs) appear to use canonical poly-A signals, whereas the ZIKV up-regulated PACs do not. *Centrimo* (Bailey and Machanick, *NAR*, 2012) shows enrichment of AAUAAA within 100 nucleotides of the the down-regulated PACs; about 20 nucleotides upstream of the poly-A cluster (PAC, position 0). Additionally, there is slight enrichment of KKKKKK (a GU rich region) downstream of the PAC. Neither of these sequences are enriched near the up-regulated PACs. These results suggest that ZIKV changes the pattern of poly-adenylation on a global scale.

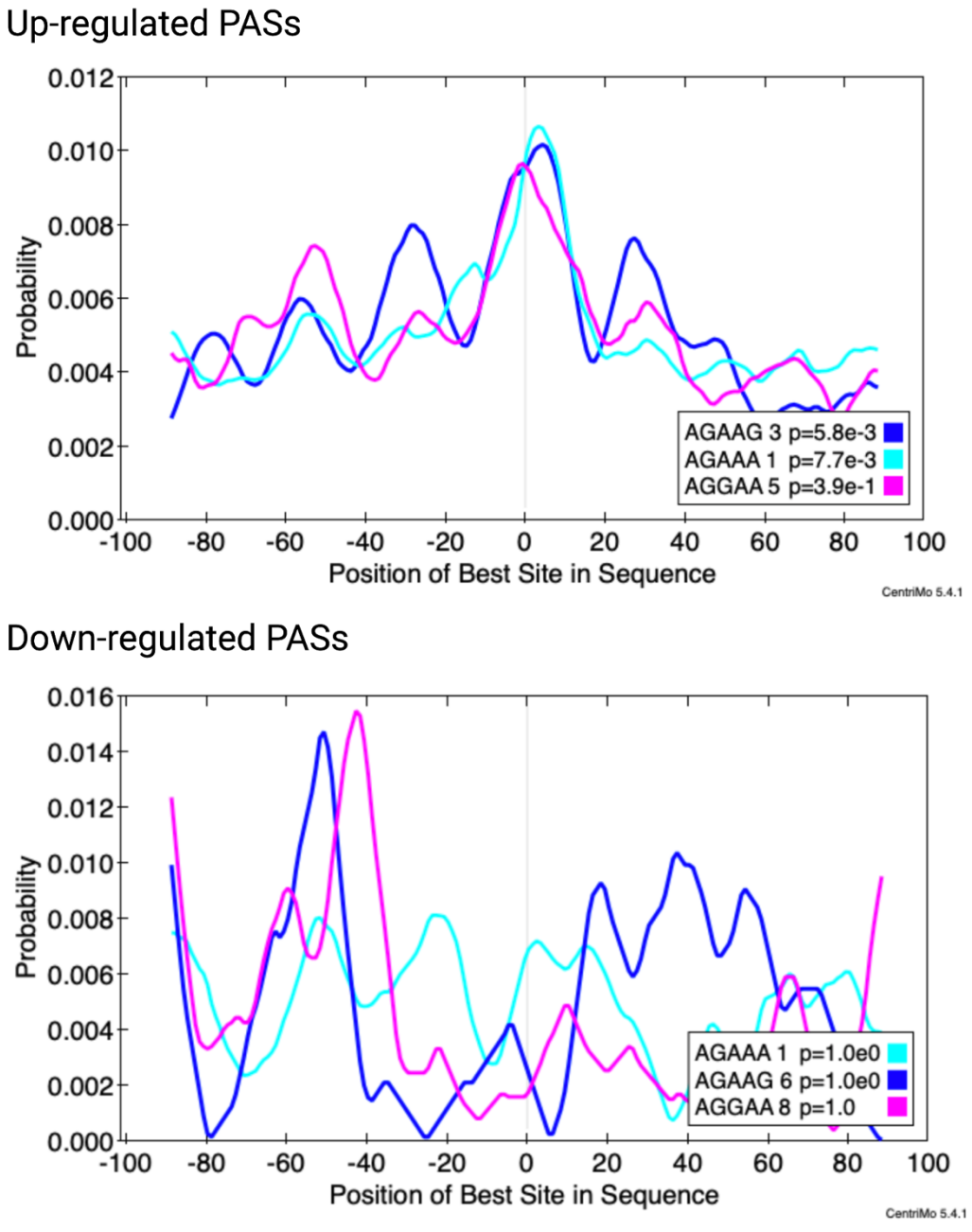


adenylation.

We did, however, manage to find a motif for PABPC1 (RAAGRAAA) enriched when we extended the distance to 250 nt of the up-regulated PACs. As we extended the sequences from 500 to 2000 nt away from the up-regulated PACs we retained enrichment of this motif. We also found additional motifs that serve as binding sites for HNRNPL and SRSF3 (CACCACCA) and SRSF1/7 (AGCRAGAC). We found the motif GCCACCRC enriched at both 1000 and 2000 nts from the down-regulated PACs. This serves as a binding site for U2AF2 (GCCACCAC) as well as PABPC1 (GCCCACCGC). These results suggest that the regions near up-regulated PACs are more likely to interact with splicing factors than those that are down-regulated.

Interestingly, we found enrichment of various motifs including AGRRR in the up-regulated PACs and not the down-regulated PACs. This is the 5-mer that ENCODE (*171*) shows SRSF11 (also annotated by p54 or NET2) to bind to. Our DGE study indicated that SRSF11 is up-regulated in ZIKV infection. Previous studies have shown that SRSF11 (*172, 173*) and other splicing factors (*174*) have played roles in APA. In the case of APA in COX2 SRSF11 binds just upstream of the poly-A signal, stabilizing interactions of core poly-adenylation factors at proximal PAC causing shortening of the 3' UTR of COX2. (*172*). We set to determine if these motifs are indeed upstream of the PACs that were differentially used in response to ZIKV infection by inputing the dys-regulated PACs we used above into another part of the MEME suite called *CentriMo* (*170*). This allows visualization of where in the sequence enriched motifs appear. We searched for each variation of AGRRR in the sequences within 100 and 500 bases of the PAC. The most significantly enriched within 100 nts of the up-regulated PAC regions are plotted in **Figure 3.18**. Interestingly, AGAAA and AGAAG are found enriched just downstream of the up-regulated PACs, but were

Figure 3.18: *CentriMo* (*170*) results searching for the 5-mer AGRRR that was reported to serve as a binding site for SRSF11 within 100 nts of our up- and down-regulated poly-A clusters (PACs). Here the center (position 0) indicates the poly-A cluster (PAC). Results indicate that sequences AGAAA and AGAAG are both enriched just downstream of the PAC. These sequences are not enriched in the down-regulated PAC regions and appear to be located further from the PAC compared to the up-regulated sequences.



found upstream of the down-regulated PACs (**Figure 3.19**). Our results suggest that the mechanism set forth by Hall-Pogar et al. was not accurate. This is primarily because they used an oligo with the sequence of upstream elements in COX-2 that SRSF11 (p54 in the article) bound *in vitro (172)*. This sequence (5'-UUGUUUGAUUUCUUAAAGU-3') is similar to the poly-pyrimidine track found at 3' splice sites, with the 3' end resembling the motif found by RNA Bind-N-Seq (AGRRR) in 2016(*171*). Therefore this binding may be artifactual. Additionally, others have suggested that SRSF11 interacts with nascent RNAs upstream of the PAC via interactions with SF-A (*175*). Instead, we propose the mechanism in **Figure 3.20** where ZIKV infection results in up-regulation of SRSF11 which leads to increased use of specific PAC by binding just slightly downstream of the poly-A signal.

We used *Flair* to investigate alternative splicing (AS) as a result of ZIKV infection in placental cells with Nanopore (long-read) libraries. Across the 6 samples there were ~4M reads ranging in length from 200 bases to 3kb (*147*). *Flair* allowed us to differentiate AS events that resulted in a larger than 5% change, for example – if a particular event occurs in the mock samples 100% of the time and in the infected samples 95% of the time. Using this cutoff we were able to generate lists of all of the AS events in each time point. We found 21 AS events: 12 exon skipping (ES), 4alternative 5' (ALT5), 3 alternative 3' (ALT3) and 2 intron retention (IR). (**Fig. 3.21**) We put the list of all alternatively spliced genes into *Enrichr* to identify any enriched pathways or processes (*167-169*). The results suggested that genes involved in glucose metabolism and anabolism, RNA processing and degradation, or viral translation and transcription were enriched in the alternatively spliced gene set, as highlighted in **Figure 3.22**.

In addition to looking at enrichment of particular processes or pathways in this gene set we used the bed files generated by *Flair* to validate these findings by uploading them into the Genome

Figure 3.19: *CentriMo* (*170*) results comparing enrichment of AGAAA (turquoise), AGAAG (blue), and AGGAA (magenta) within 500 nts of up-regulated (solid line) poly-A clusters (PACs) or down-regulated (dashed line) PACs. We see enrichment of these motifs near the poly-A site (position 0) in the sequences surrounding up-regulated PACs and enrichment upstream for down-regulated PACs.

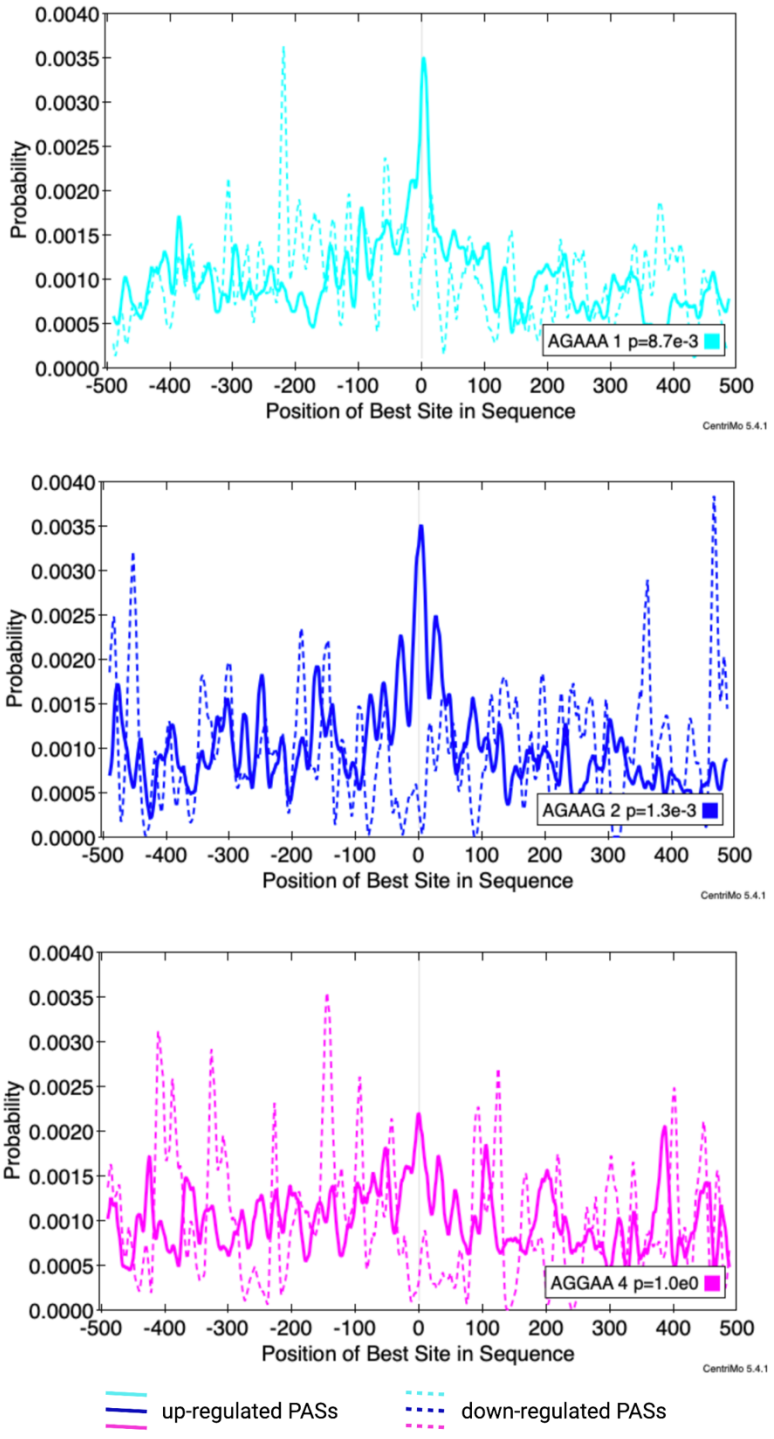


Figure 3.20: Potential mechanism of alternative poly-adenylation (APA) in response to ZIKV infection in JEG3 cells. ZIKV causes up-regulation of SRSF11 (p54/NET2) which is now more available to bind nascent RNA with AGRRR motifs This interaction just downstream of a poly-A signal aids in recruitment or stabilization of core poly-adenylation factors causing preferential poly-adenylation of transcripts with AGRRR motifs near the poly-A signal. EXO trims new transcripts and XRN2 removes the RNA, causing dissociation of RNAPII when it catches up to the polymerase. Image created by S. Sotcheff using BioRender under license.

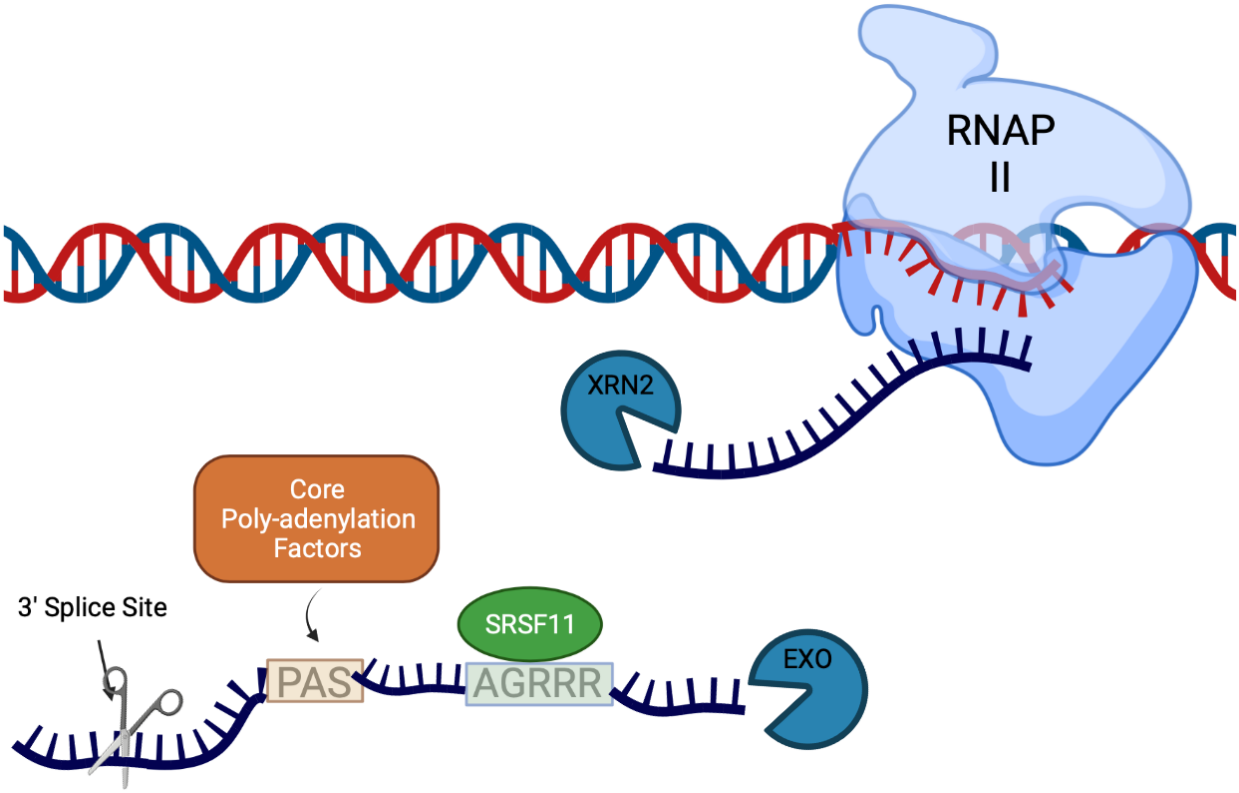


Figure 3.21: Pie chart depicting the relative percentage (and total counts) of different types of splicing events detected by *Flair* (*147*) analysis in ZIKV infected JEG3 cells. ES: exon skipping, IR: intron retention, Alt5: alternative 5' splice site, Alt3: alternative 3' splice site. Chart created by S. Sotcheff in Microsoft Excel under license.

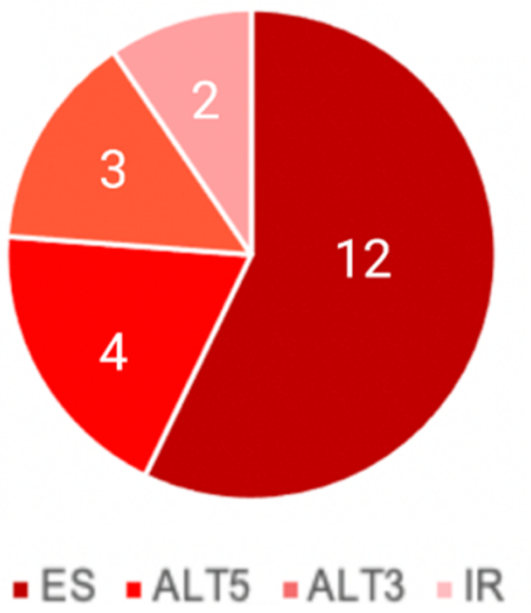
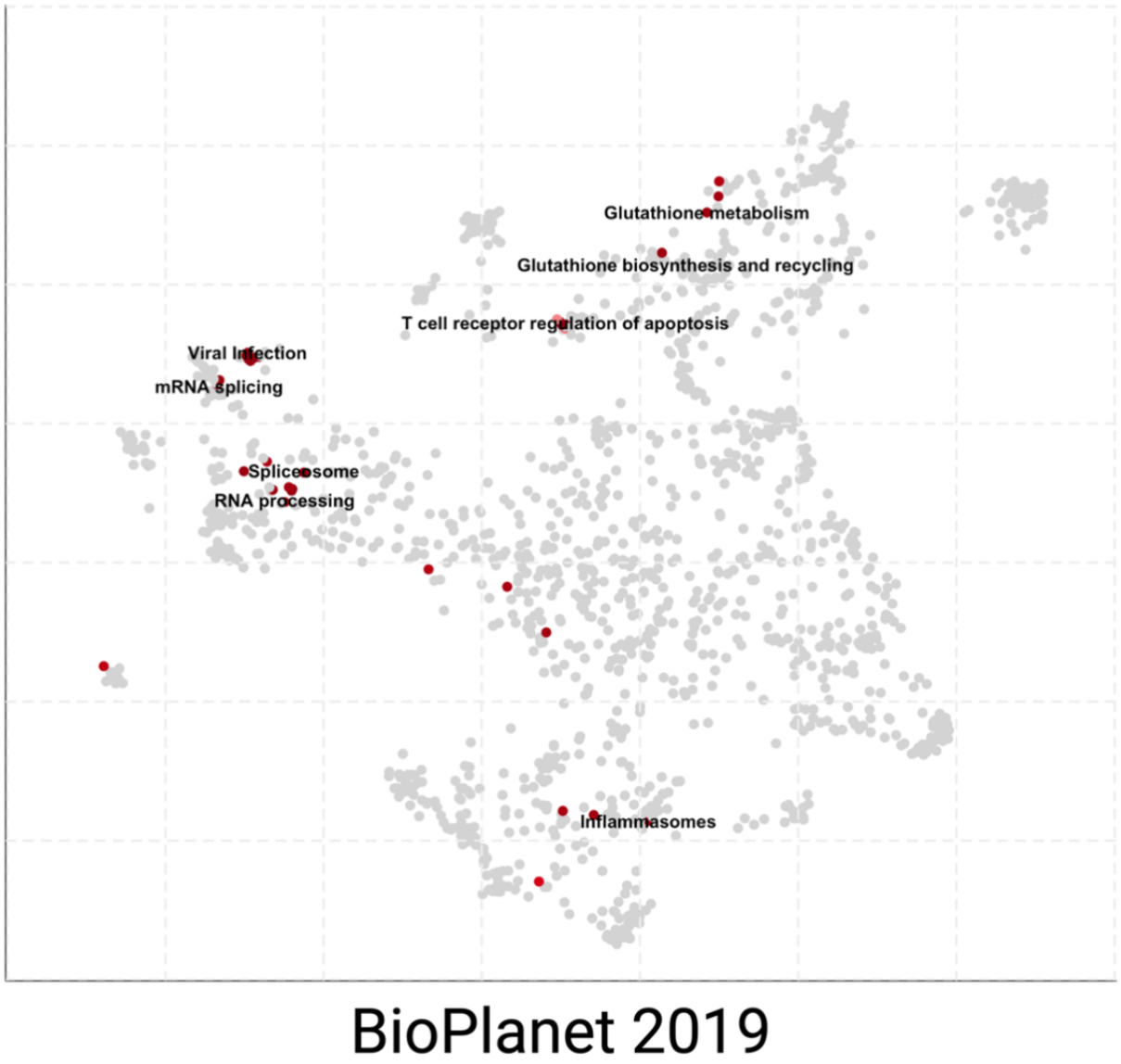


Figure 3.22: *Enrichr* (*167-169*) scatterplot showing pathways that were enriched for alternative splicing (AS) (red) in JEG3 cells upon ZIKV infection. Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Pathways enriched for AS (p-value <0.05) are labelled as space allows.



Browser as custom tracks – note that triplicates were uploaded as a single bed file for ease of viewing.[147, 161] We wanted to validate that the Flair results matched up with the actual reads in the bed files as displayed in the genome browser by looking at a few genes individually. We were able to determine that the Flair results were accurate based on the reads displayed. In **Figures 3.23 and 3.24** we show the primary splice variants in our dataset for the gene PEBP1, which encodes phosphatidylethanolamine binding protein 1 – a protein that modulates the MAP kinase pathway. In fact, this protein can be further processed to produce hippocampus cholinergic neurostimulating peptide (HCNP) which may be involved in neural development. In **Fig. 3.24** the sections are color coded for the variant they represent in **Figure 3.23**. From here you can clearly see the differential use of cassette exons 2 and 3 and an alternative 3' end of the terminal (4^{th}) exon.

DENV INFECTION OF JEG3 CELLS

PAC-seq data after *DESeq2* analysis yielded 10,724 genes to compare mock to DENV infected JEG3 cells. Of those 115 genes were differentially expressed (73 down- and 42 up-regulated) with a p-adj <0.1 and |log2FC| greater than 0.58 and the distributions of genes based on their p-adjusted and log2FC are shown in the volcano plots in **Figure 3.25**. We confirmed the data quality was sufficient by generating and loading bedgraph files into the UCSC Genome Browser[31] as shown in **Figure 3.26** for genes encoding DLG5 and DDX5, a scaffolding protein for cell-cell contact and a DEAD-box helicase, respectively. *Enrichr* (*32-34*) identified enrichment of different pathways, ontologies, or phenotypes in the 115 genes differentially expressed. The results suggest that DENV infection in JEG3 cells hinders AKT signaling and loading of class I MHC peptides and promotes Toll-like receptor signaling and TGF-beta signaling (**Figures 3.27 and 3.28**).

Figure 3.23: Primary splice isoforms of PEBP1 in ZIKV infected JEG3 cells detected using *Flair* (*147*) analysis. Isoforms are color coded with Figure 34 to show relative abundance in each sample. Figure created by S. Sotcheff using *Flair*.

Figure 3.24: Relative abundance of splice isoforms of PEBP1 upon ZIKV infection in JEG3 cells detected using *Flair* (*147*) analysis. Isoforms are color-coded with Figure 3.23. Plot created by S. Sotcheff using *Flair*.

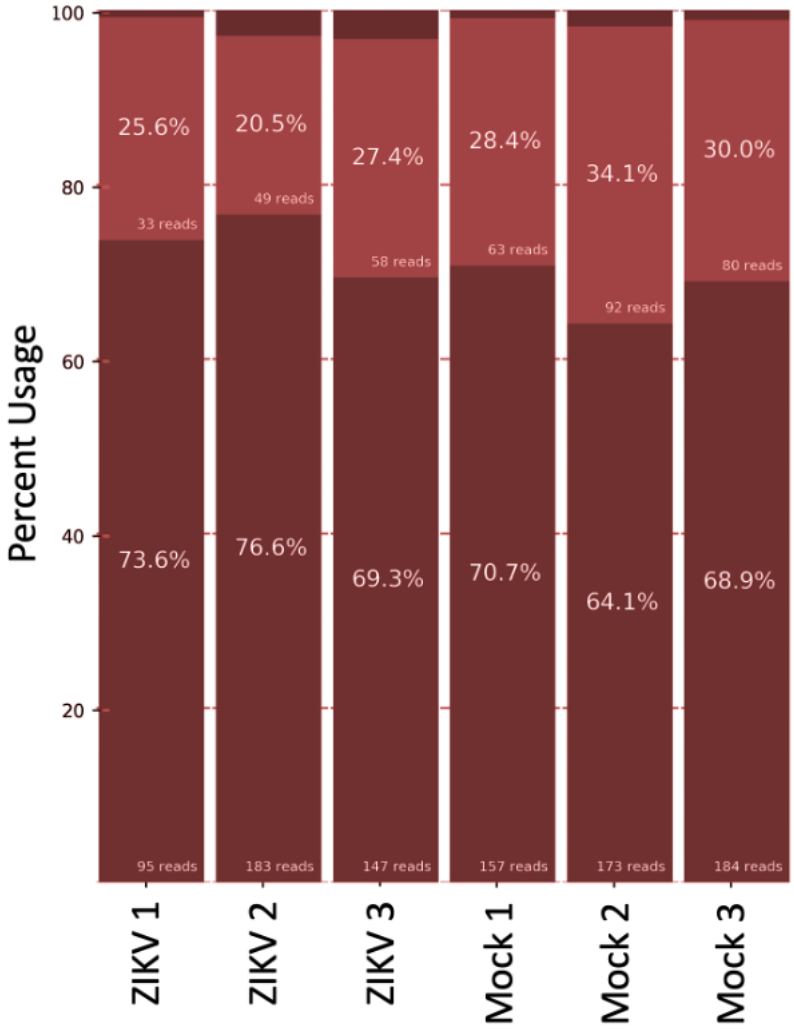


Figure 3.25: Volcano plot highlighting differentially expressed genes (DEGs) in human placental (JEG3) cells in response to DENV infection. Each point is a gene. Up- (orange) and down-regulated (purple) genes have a p-adjusted value <0.1 and log2FoldChange of >0.585 or <-0.585 respectively. DEGs are labelled as space allows. Plot created by S. Sotcheff using R under license.

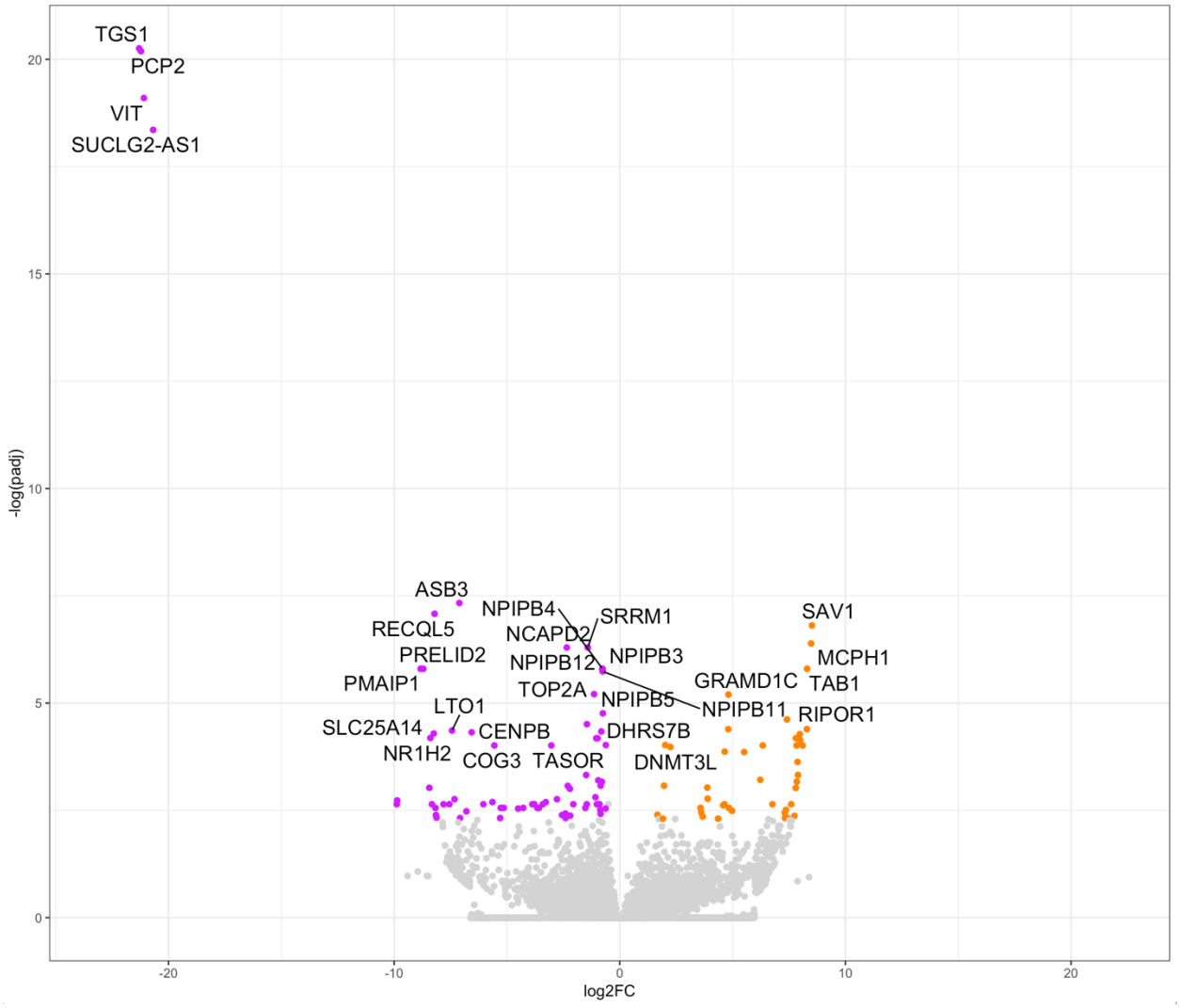


Figure 3.26: Alignment of bedgraph files for representative samples the ZIKV JEG3 data at one up-regulated (DLG5) and one down-regulated (DDX5) gene. As PAC-seq primes from the poly-A tail and cDNA was stochastically terminated in the reverse transcription step, we expect coverage to be highest at poly-A cluster (PAC) and to taper off as we extend into the gene. This is indeed the case.

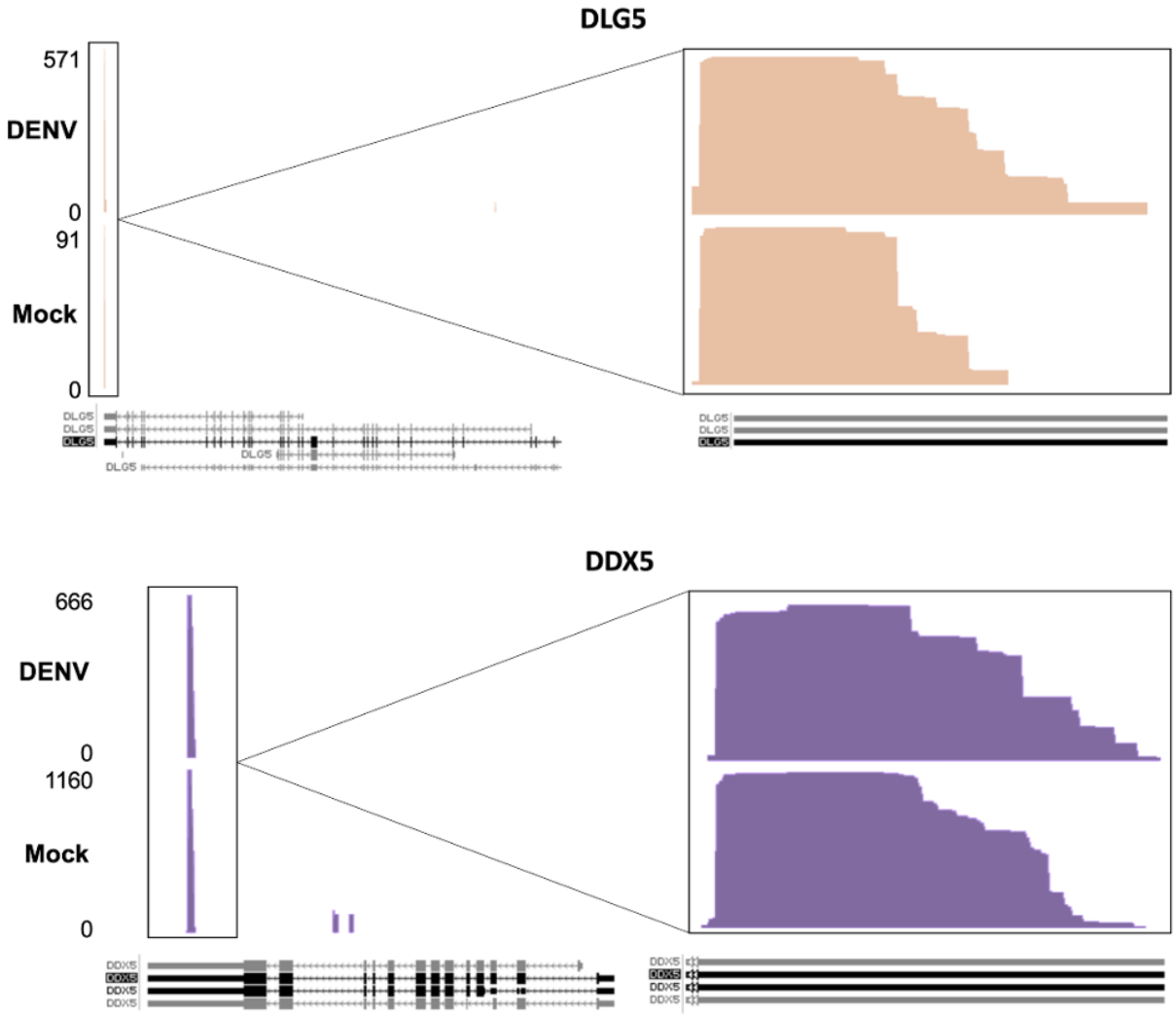


Figure 3.27: *Enrichr* (*167-169*) scatterplot showing up-regulated pathways (orange) in JEG3 cells in response to ZIKV infection. Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Up-regulated pathways (p-value <0.05) are labelled.

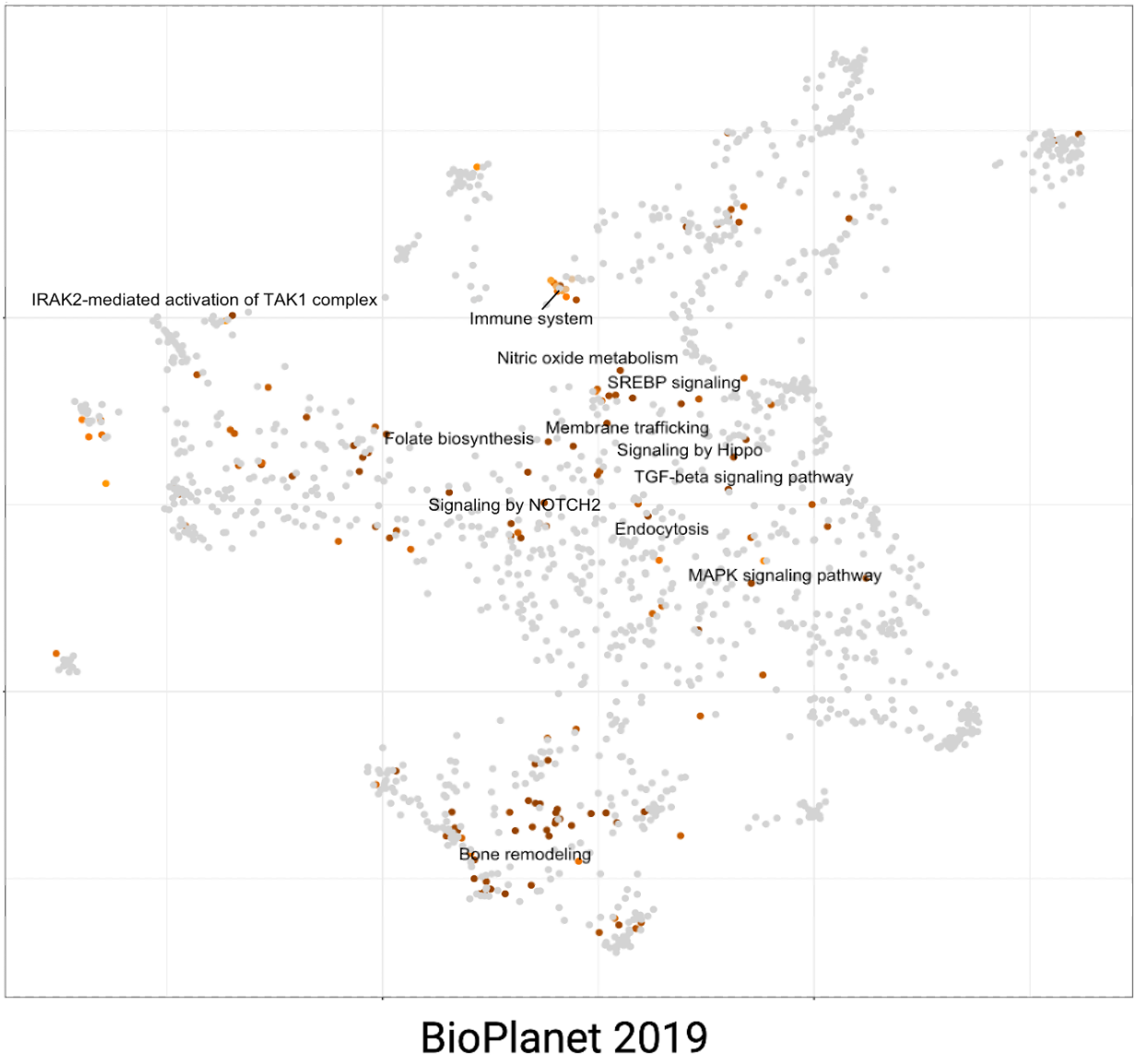
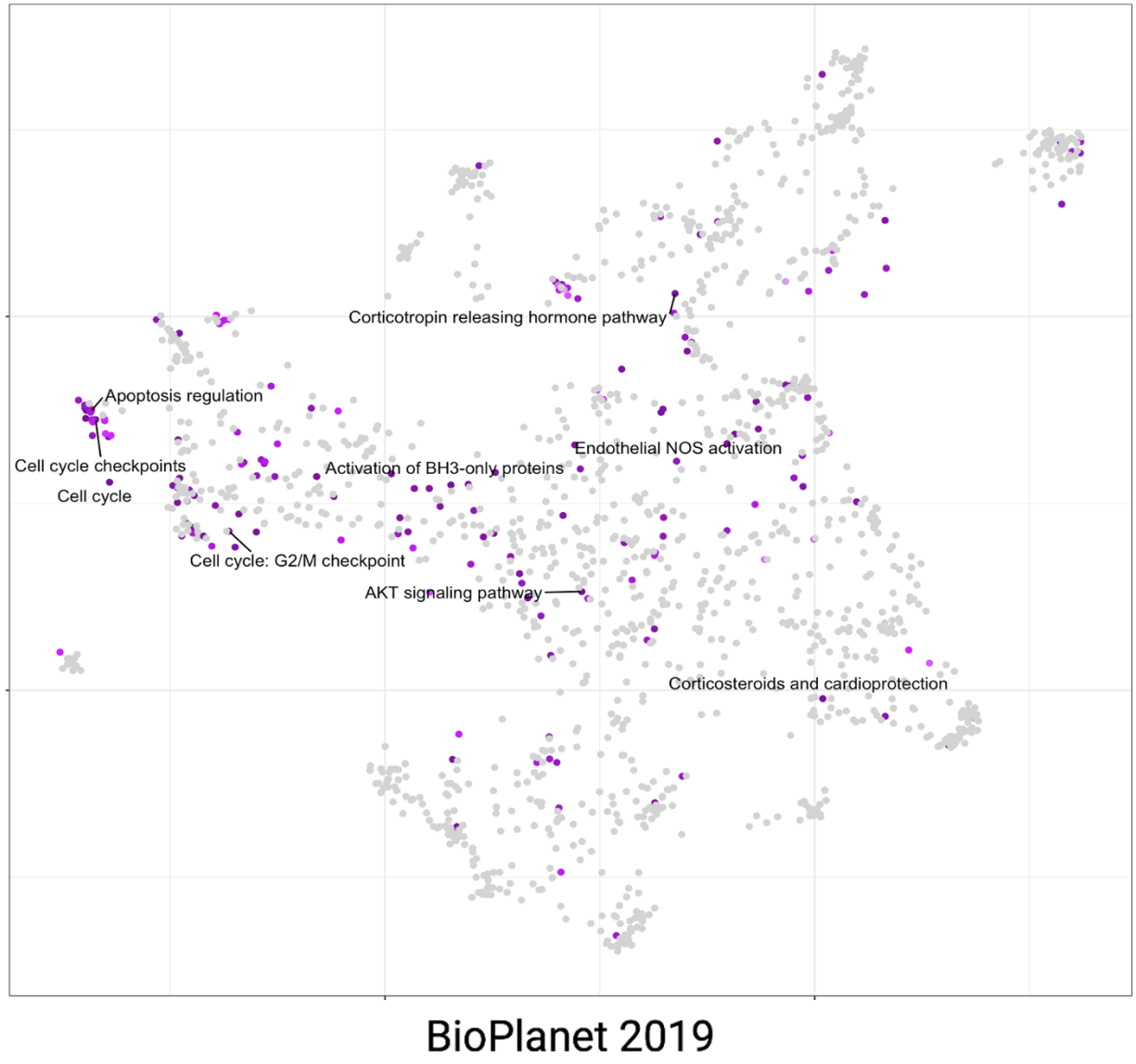


Figure 3.28: *Enrichr* (*167-169*) scatterplot showing down-regulated pathways (purple) in JEG3 cells in response to DENV infection. Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Down-regulated pathways (p-value <0.05) are labelled as space allows.



DPAC identified 2,013 genes with multiple PACs in the DENV infected JEG3 cells. **(Fig. 3.29)** For our analysis we focused on changes of at least 20% in PDU. To confirm, we uploaded the generated bedgraph files into the genome browser and viewed individual genes that were identified has having APA [161]. Bedgraph files confirmed mapping to the human genome at the 3' end of genes and indicated that NBR1 (an autophagy cargo receptor) and RBMS1 (an RNA binding protein) undergo shortening and lengthening of their 3' UTRs respectively **(Fig. 3.30).** Our analysis found 132 3' UTRs shortened and 126 lengthened 3' UTRs. As evidenced by the *Enrichr* Scatterplot plots in **Figure 3.31** (blue) there is shortening of transcripts involved in NOTCH signaling and actin cytoskeleton regulation.[167-169] In addition, these scatterplots show lengthening (yellow, **Figure 3.32**) of transcripts involved in protein processing in the endoplasmic reticulum (ER) and regulation of Toll-like receptor signaling.

We extracted sequences of varying lengths at up- and down-regulated PACs as described in Chapter 2. DREME, an online tool for identifying enriched motifs, was used to determine if any short sequences were enriched at these PACs. These enriched motifs are shown in **Table 3.4.** Browser[31] Further, these motifs were cross-referenced with RBPDB (*160*) to determine RNA binding proteins that may interact near these up- or down-regulated PACs. We found the canonical motif for poly-A binding protein 1 (PABP1, motif: AAUAAA) within 100 nts and 500 nts of the PAC for up-regulated PACs. While this motif was also found within the same distances for the down-regulated PACs we also found enrichment of AGAAGAA, a motif recognized by YBX1, TRA2, HRNRPA1, and SRSF proteins 1 and 9, and ACACACA which is recognized by cytoplasmic PABP1 (PABPC1). When we extended the included sequence to 1000 and 2000 nts we found enrichment of AGGCCGAG in regions near both up- and down-regulated PACs. These results indicate that splicing factors are more likely to bind the transcripts with the down-regulated

Figure 3.29: Plot depicting changes in percent distal usage (PDU) of poly-A clusters (PACs) between the mock-infected and DENV-infected human placental (JEG3) cells. Each point is a gene, those colored in yellow have an increase in PDU of 20% or more and those in blue have a decrease in PDU of 20% or more in response to DENV infection. These colored points (genes) have alternative poly-adenylation (APA) and are labelled as space allows. Plot created by S. Sotcheff using R under license.

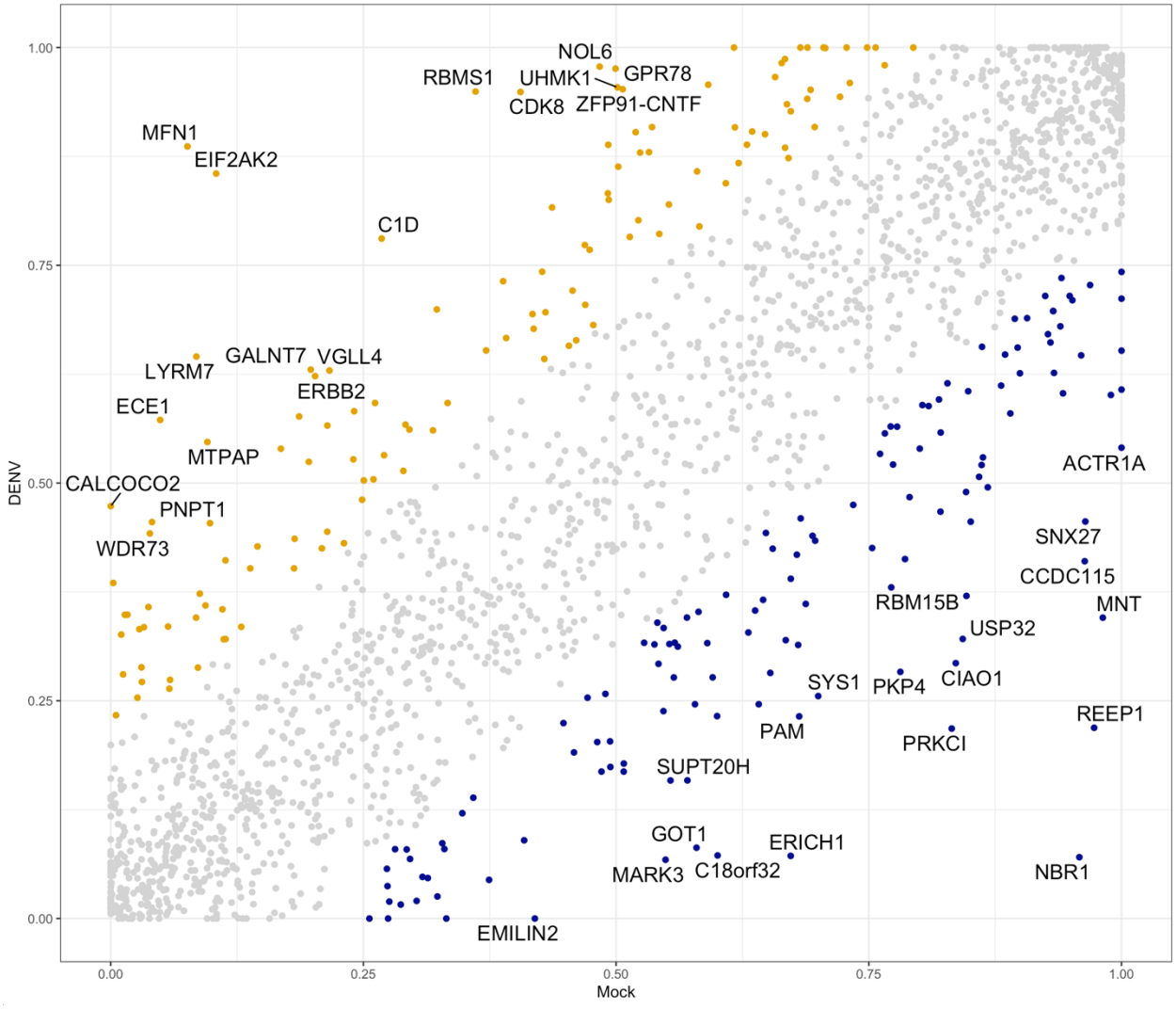


Figure 3.30: Alignment of bedgraph files for representative samples the DENV JEG3 data at one gene with a lengthened 3' UTR (RBMS1) and one with a shortened 3' UTR (NBR1).

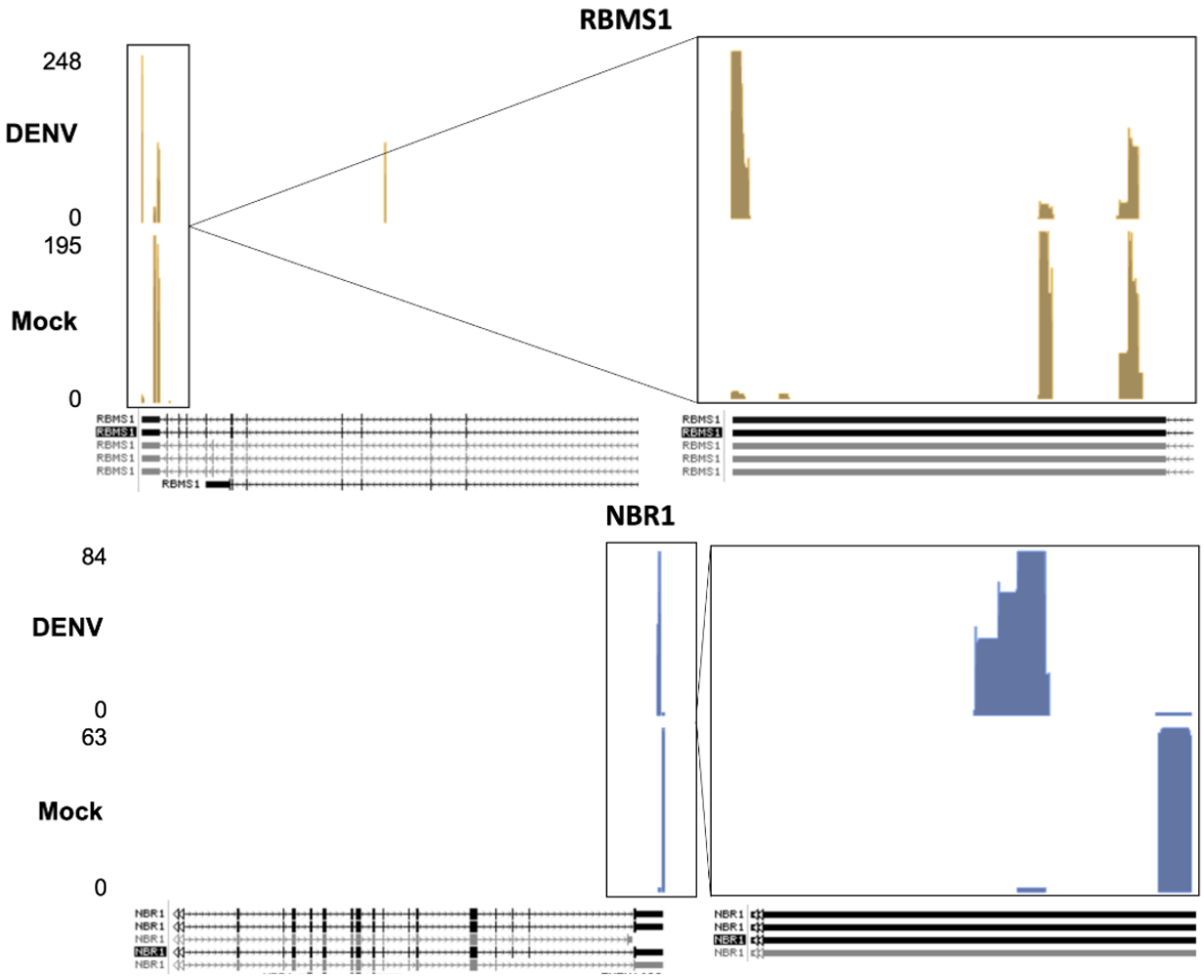


Figure 3.31: *Enrichr* (*167-169*) scatterplot showing pathways that were enriched for shortened 3' UTR (blue) in JEG3 cells upon DENV infection. Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Pathways enriched for 3' UTR shortening (p-value <0.05) are labelled as space allows.

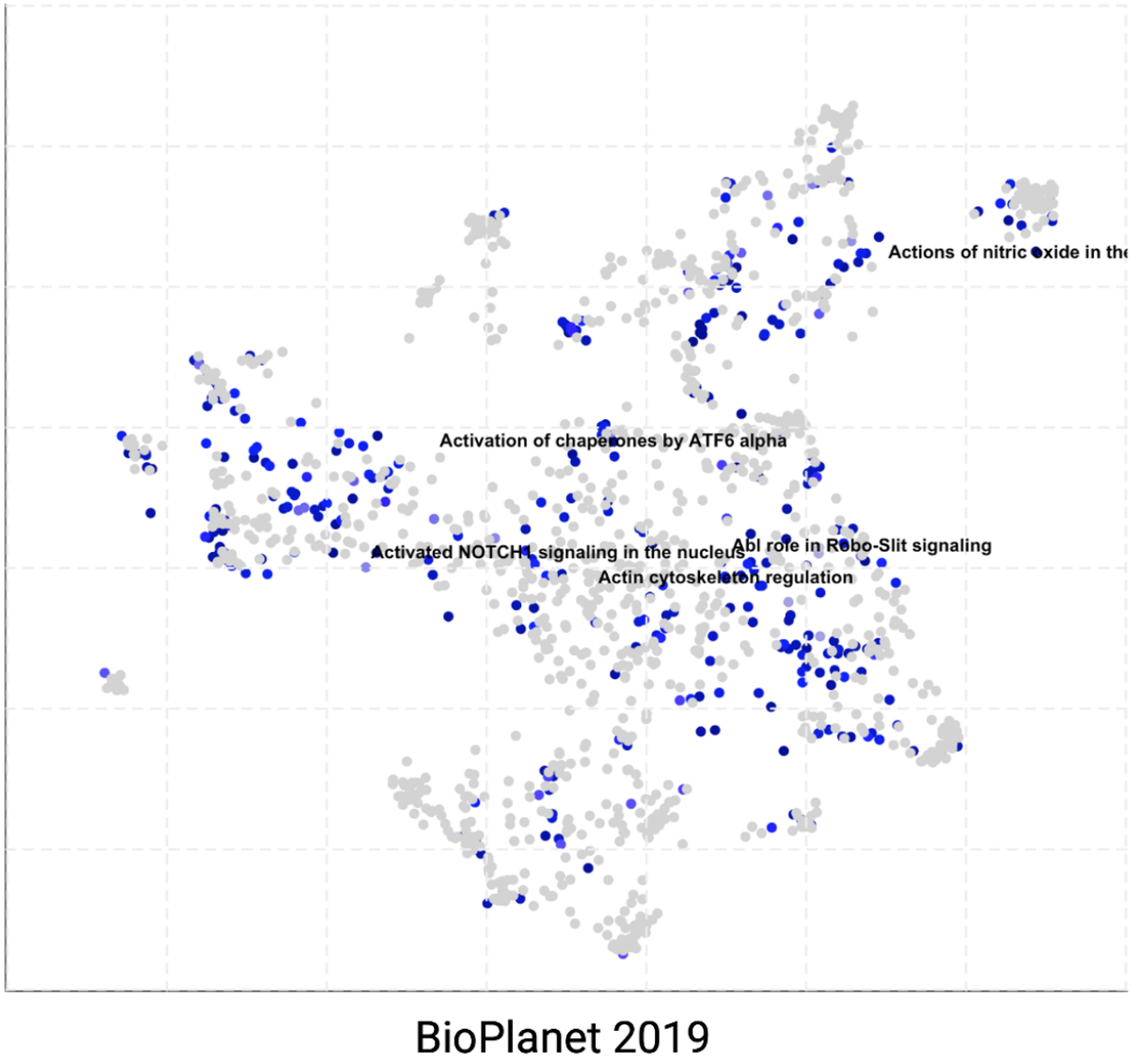


Figure 3.32: *Enrichr* (*167-169*) scatterplot showing pathways that were enriched for lengthened 3' UTR (yellow) in JEG3 cells upon DENV infection. Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Pathways enriched for 3' UTR lengthening (p-value <0.05) are labelled as space allows.

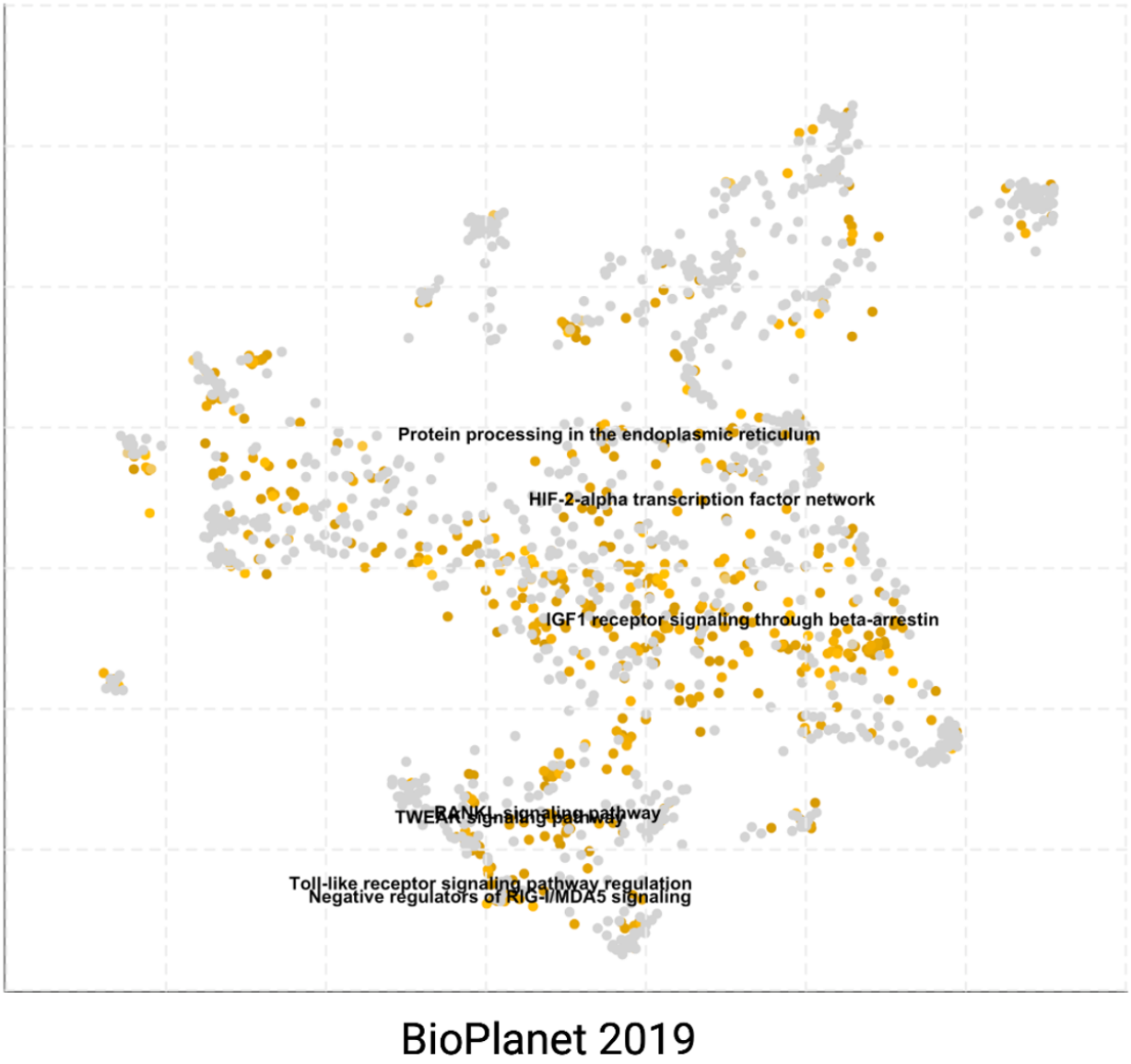


Table 3.4: Motifs enriched with 'X' nucleotides of up- or down-regulated poly-A clusters in human placental (JEG3) cells upon DENV infection.

Condition/Nts from PAS	100	250	500	1000	2000		KEY
UP	AATAAA	None	AAATAHAA	CTSCAGCC	CTGTAATC	ACAGAGHG	R=purine (AG)
				AGGCBGAG	GAGGCVGA	TCACTGCA	Y= pyrimidine (CT)
				AAATASAA	ACTMCAGC	TATATATA	W = weak (AT)
				GCCACYGC	KCCTCCCA	GCCTGTR	S = strong (CG)
				CTRTAATC	CYGCCTCA	GTAGAGAY	K = keto (GT)
				CMCATCTC	GACCAKCC	GTTRGCCA	M = amino (AC)
				CAGYTAC	AAAAATAC	GCCBGGC	U/T interchangeable
				TGAATGWA	CARTGGCA	ACCWCGTG	D= not C
				ACAGRAAA	CCGTSTC	CCGGGTTC	V = not T
				RAGAGAAA	GTGGCKCA	RAAGAAAA	H = not G
				ATGRAAAC	CATACATA	AAAATAWA	B = not A
				AAAAAMAA	CGAWCTCC	GWGACAGA	N = any base
					GGTTTCAC	SAGATGGA	X = ACGT = .
					CCCAGCTA	GCCACCAY	
DOWN	HAAAAATA	AAAABAAA	CTGTRRTC	TATTTTTW	KAAAAATA	AAAAMAAA	
	AGMAGAA	AGCAGWA	CCCAGGCY	CKCACTGC	CTCACTGC	AAAATAAA	
		CACWGCAC	CARCCTCC	AGGCWGGA	GTGGTGGB	GAGHGAGA	
		AYACACA	GAATAWA	CAGCCDCC	CTGAGGCA	AARAGAAA	
		CCAGCCTS	RCAGTGR	ATCYCAGC	GGGATTAC	GGTTTCAC	
		MATATATA	AAAWATA	GTGGCTCR	CAGCCTGG	CGATTCTC	
		CTGCARCC	AAAAVAAA	GACCAKCC	CACTGHAC	CTGTAGTC	
			AYACACA	AAATAWAT	ATCTCGG	ATRTATAY	
			CCAGCYAC	AAAASAAA	AGGCRTGA	CKTGAACC	
			GAGRCAGA	AGRCRGAG	GTAAAATD	AACKCCTG	
			AGSCAGGA	CRCCTGTA	CASCTTCC	CCRTCTCW	
			GCCCKGC	CACYGCAC	CCACCYGC	AAAATYAG	
			CYACCA	AAATAMAA	TTTAWAAA	AAAGTGCT	
				GAGRTGGR	CTGWGATC	AGCTACTC	
				CCACCACS	AAATACAR	AGGCBGAG	
				AACTCCTG	ATAWTAAA	GTGGCTCA	
				CCCRGCTA	CCTGGGCW	GCCCGGCY	
				AAAAAWAA	AAACATTT	AAATAHAT	
				TACAGRCA	GAGGCAGA	ATCWCCTG	
				GGCCAGAS	CCWCCTCC	GGCCAGRC	
				CCTCCCAM	CATGKTGG	CWGTCTC	
				TATATATR	CCRAGTAG	CGGHTCAC	
				CAGAGCGA	ACACACAC		
				ACASATAC			
				MAGGAAGA			
				GTATAY			
				CKTGAACC			
				GSCGGGCA			
				CCACCMGC			

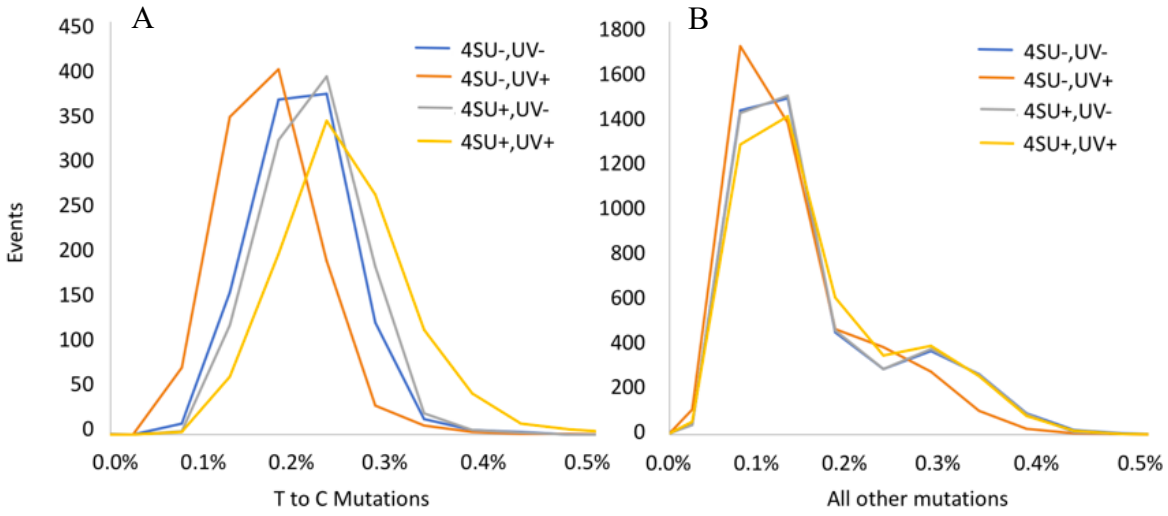
PACs.

v-PARCL

It is plausible to consider that transcriptomic changes, including alternative splicing, differential gene expression, and alternative poly-adenylation may be cause by interactions of soluble flavivirus proteins (C, NS5 and NS3) with host proteins or nucleic acids. Flavivirus C has been found on lipid droplets and in the nucleolus and has been shown to perturb ribosome biogenesis. However, this would be considered an extra-mural role of capsid which is necessary for the packaging of the viral genome. Interestingly, capsid appears to interact with nucleic acids non-specifically (no specificity for single or double strandedness, RNA or DNA, or sequence). Therefore, we were interested in how flavivirus capsid interacts with the viral genome in virions considering the low specificity for any particular nucleic acids.

We have previously validated our PAR-CL/NGS technique by investigating capsid-RNA interactions in FHV (an alphanodavirus originally isolated from grass grubs in New Zealand)(*107, 176*). The T to C mutations (in cDNA libraries) as a result of 4SU crosslinking to FHV capsid protein were validated by comparing this mutation rate to that of controls (4SU+/UV-, 4SU-/UV+ and 4SU-/UV-) as noted in **Figure 3.33**. The rate of all other mutations were unaffected by treatment with 4SU or UV. T to C mutations were found throughout the FHV RNA1 genome (*107*). Previously the PAR-CLIP method has been shown to look at mutation rate, but we opt to use fold change as a better comparison of T to C mutations in this technique. The benefit of using this method to detect the capsid-RNA interactions is that the output provides both positive and negative signal. If there are discrete binding sites one could expect to see peaks in the PAR-CL signal (a higher ratio of T to C mutations in the experimental vs. control), and if there were no

Figure 3.33: Validation of PAR-CL with NGS technique in determining RNA sequences involved in capsid-RNA interactions of FHV. A) T to C mutation rates are increased when cells were incubated with 4SU and particles were subject to UV 365 nm. B) Rates of all other mutations, indicating that PAR-CL treatment did not result in changes to mutation rates other than T to C mutations. Plots created by S. Sotcheff using R under license.



specific binding sites we would find to peaks, but rather a ratio of T to C that looks like noise throughout the genome. This also minimizes error introduced by PCR or the sequencing platform when adequate coverage (1000+) is accomplished at each nucleotide position. Therefore, there is a high confidence in the PAR-CL signals found in FHV. Subsequently these data were compared to output from RNAstructure (*177*) using soft constraints from DMS MaP-Seq data previously collected for FHV. The significant PAR-CL sites were complementary to the ssRNA regions noted by our DMS MaP-Seq experiments, suggesting that FHV capsid binds the FHV genome in double stranded regions. Interestingly, hindering these base pairings by introducing mutations reduced viral fitness. In addition, these results can be validated using psoralen crosslinking and NGS in future studies (*178, 179*).

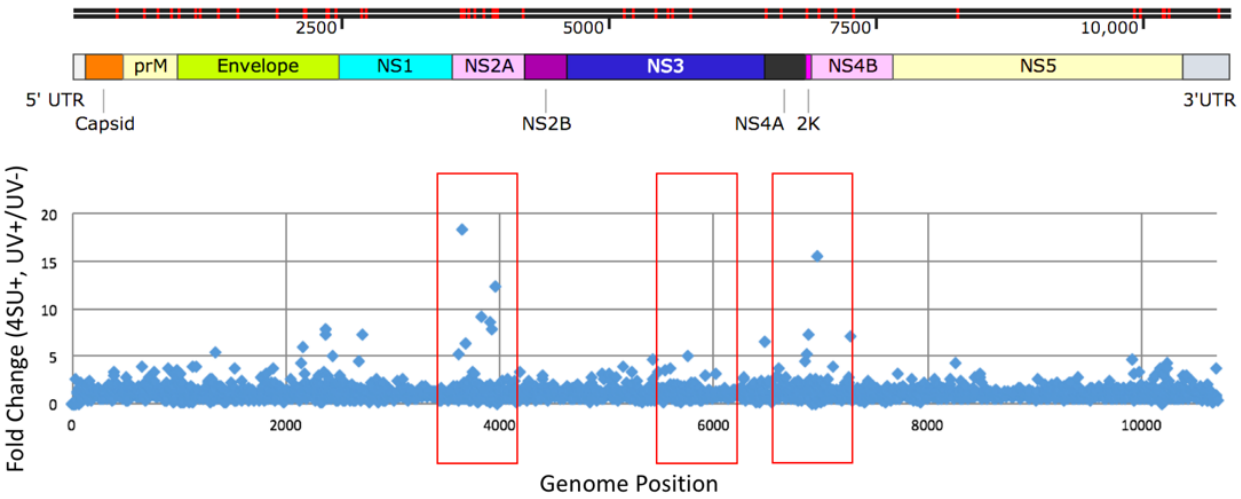
Since capsid is the least conserved flaviviral protein, it appears to have a broad range of binding partners within the host cell and roles in pathogenesis aside from packaging. Greater understanding of capsid's various roles in pathogenesis can ultimately aid in designing attenuated virus or identify potential antiviral targets (*180-188*). Although flavivirus capsid (C) protein is known to interact with various nucleic acids, both single stranded (ss-) and double stranded (ds-), RNA and DNA, we began with the hypothesis that there may be specific sequences or motifs in the genome that interact with ZIKV C in virus particles. Our rationale was that within the context of the replication factories, where packaging occurs at the ER membrane, the conditions may elicit more specific binding of capsid to the viral genome (vRNA). ZIKV RNA sequences that interact with ZIKV C inside released virus particles can be identified using v-PAR-CL (*107*) coupled with NGS.

We recently applied this novel method to identify binding positions for capsid protein in the Dakar strain ZIKV (named for the capital of Senegal in West Africa isolated 1954) (*189*). ZIKV was

amplified in Vero cells in the presence of 4SU and the PAR-CL libraries were made following our validated protocols. We have found various strong PAR-CL signals **(Fig. 3.34)**. Additionally, we found that the PAR-CL cluster regions are mostly conserved between both Dakar and two contemporary ZIKV strains: PRV59ABC (an Asian-lineage derived American strain isolated from Puerto Rico in 2015) and FSS13025 (an Asian strain isolated from a patient in Cambodia in 2010). We used Multiple Em for Motif Elicitation (*MEME*) (*190*) and Discriminative Relative Expression Motif Elicitation (*DREME*) (*159*), to identify sequence motifs throughout the Dakar genome or within regions of high PAR-CL signal respectively. The results were not conclusive across both software packages. Although MEME identified potential motifs near the major PAR-CL clusters, DREME did not find any short motifs present in the genome sequences within 50 nucleotides of the top 50 PAR-CL signals. This indicates that there may be a motif that is smaller or larger than the size searched for by the software or that the ZIKV C binds secondary structures within the viral genome. It may be possible that there is not a specific sequence responsible for the interaction between ZIKV C and specific RNAs, suggesting that the environment within the replication factories has no impact on the binding specificity of ZIKV C.

Due to incongruent results from the sequence analyses (*MEME/DREME*) we pondered if capsid binding within ZIKV particles was due to RNA structure. Although flavivirus capsids do not appear to specifically bind single stranded or double stranded nucleic acids, we compared our PAR-CL signals to a previously published structure of the ZIKV genome within particles (*191*). The previous study was done using ZIKV MR766 (a different African strain than Dakar), but the sequence identity is high. After aligning these two genomes we found that indeed, the ZIKV capsid appears to be interacting with both ds- and ss-RNA within Dakar particles.

Figure 3.34: PAR-CL/NGS data from particles produced by Dakar infected Vero cells, incubated with 4SU (100 uM). (Top) Map of ZIKV genome and red marks indicate the top 50 PAR-CL signals. (Bottom) Comparing 4SU+ samples that were or were not subject to UV radiation. (UV+/UV-) Consensus motifs identified by MEME software, corresponding to regions of high PAR-CL fold change are boxed in red. Figure created by S. Sotcheff using Microsoft Excel and Powerpoint under license.



DISCUSSION

Although ZIKV was discovered in rhesus monkeys in the Zika forest of Uganda in the 1940s, it was not necessarily on the map in terms of human infectious disease, so to speak, until it's emergence in Brazil during the 2015-16 outbreak.[2, 9, 10] This outbreak was associated with the development in Guillain-Barre syndrome in adults and microcephaly in infants born to women infected during the first trimester of their pregnancy. Since then we have found that ZIKV can persistently infect the testes. This is especially interesting as flaviviruses are typically transmitted via an arthropod vector – in this case a mosquito bite. Now that we are aware that ZIKV can be transmitted both sexually and from mother to child during pregnancy it is important to understand how these alternative transmission routes are made possible. We have focused on the incidence of microcephaly, caused by infection of neural progenitor cells in the developing fetus. In order for the virus to reach this destination the virus must first cross the placental barrier, therefore we have decided to investigate transcriptomic changes in ZIKV infected human placental cells. It is also important to note that microcephaly does not occur upon infection of expectant mothers with other flaviviruses, therefore we investigated transcriptomic changes in DENV infected JEG3 as well. As an additional control we also infected a human liver cell line (Huh7) with ZIKV.

Our study used PAC-seq to generate libraries for Illumina short-read sequencing and DPAC analysis to identify 83 DEGs and 380 APA genes at 16 hours post infection ZIKV infected Huh7 cells. To determine if there was enrichment for any pathways or processes in these changes in the transcriptome we used *Enrichr* to compare our results to the BioPlanet 2019 database (*167-169*). We found that the down-regulated genes were enriched in pathways regarding the electron transport chain, MAP kinase pathways, and one carbon metabolism. The up-regulated genes were enriched in pathways regarding interferon alpha and internal ribosomal entry. The 3' UTRs of

genes involved in pre-mRNA processing, cell cycle regulation and activation of NFkB in B cells were shortened, suggesting these genes are less-regulated than in mock samples. Additionally, genes involved in the minor splicing pathway and developmental biology had lengthened 3' UTRs in response to ZIKV infection. These results suggest altered splice patterns, metabolism, and immune response upon viral infection of the liver.

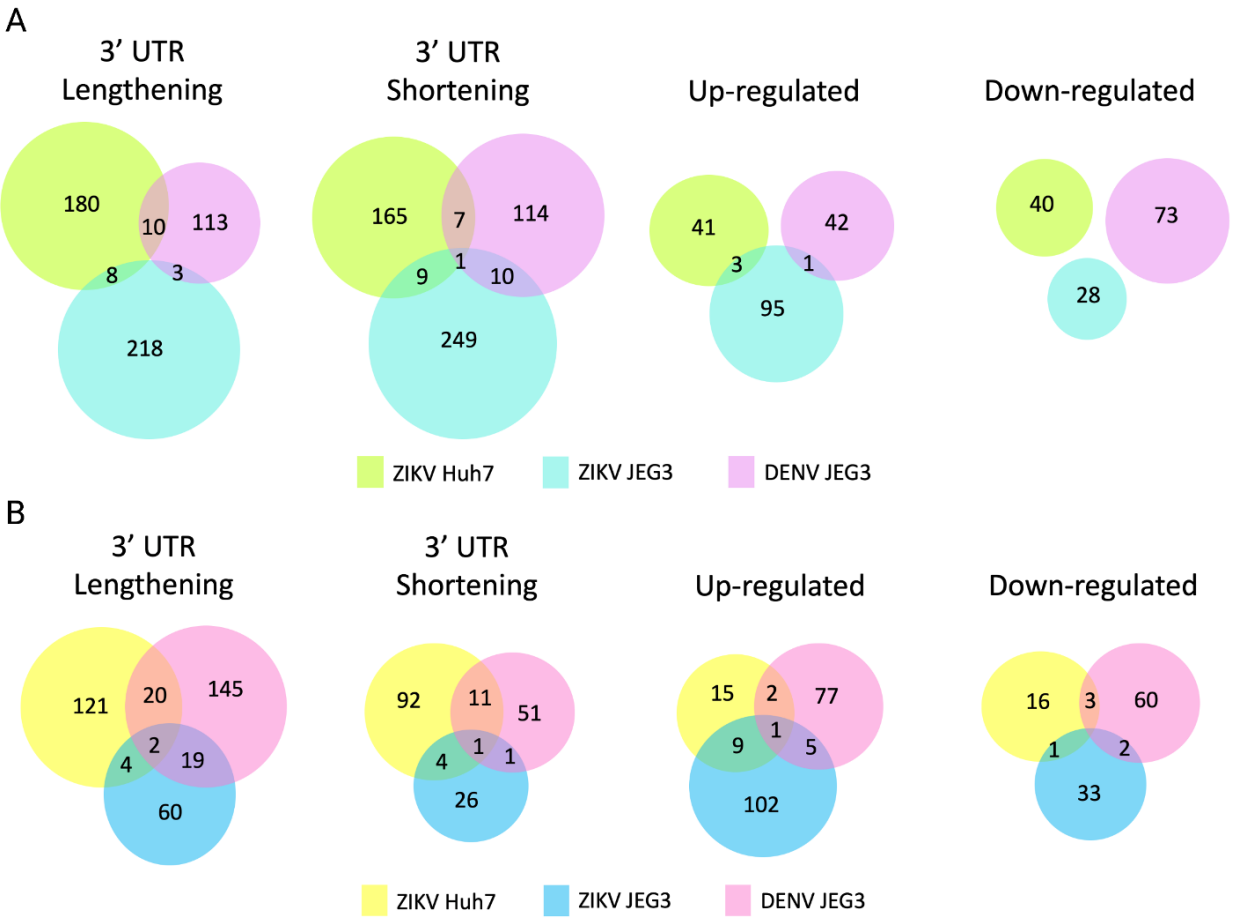
We used PAC-seq to generate libraries for Illumina short-read sequencing and *DPAC* analysis to identify 126 DEGs and 498 APA genes at 16 hours post infection for ZIKV infected JEG3 cells. We also used long-read sequencing followed by *Flair* (*147*) to identify 21 AS events in ZIKV infected JEG3 cells. To determine if there was enrichment for any pathways or processes in these changes in the transcriptome we used *Enrichr* (*167-169*). This identified that there was an enrichment in ZIKV-induced up-regulation of genes associated with interleukin and interferon signaling, as well metabolism and regulatory RNA pathways (including splicing) in the JEG3 cells. In addition, there was enrichment of genes associated with cell cycle regulation and DNA replication and repair. Interestingly, we found that the 3' UTRs of transcripts associated with splicing and mRNA processing as well as metabolism were shortened, whereas the 3' UTRs of transcripts associated with various aspects of development and T cell receptor signaling were lengthened. When plugged into *Enrichr*, the AS events identified in ZIKV infected JEG3 cells were enriched in genes associated with mRNA processing and splicing, translation, transcription, and glucose metabolism. These results suggested alternative splicing, impaired cell cycle regulation, and inflammation in human placental cells upon ZIKV infection.

Our study used PAC-seq to generate libraries for Illumina short-read sequencing and *DPAC* analysis to identify 115 DEGs and 258 APA genes at 16 hours post infection for DENV infected JEG3 cells. To determine if there was enrichment for any pathways or processes in these

changes in the transcriptome we used *Enrichr* (*167-169*). In response to DENV infection we found down-regulation of loading MHC I peptides, cell cycle control, and AKT signaling and up-regulation of toll-like receptor signaling and TGF beta signaling. The 3' UTRs for genes involved in protein processing at the ER membrane and toll-like receptor signaling regulation were lengthened while the 3' UTRs of genes involved in NOTCH signaling were shortened. These results suggest that DENV infected placental cells are potentially unable to identify DENV peptides using MHC I but like ZIKV infection in both other cell types also experience issues with regulating the cell cycle.

The results from our DEG and AS analyses appear to be in line with what has been previously noted by others. As noted above, there are a plethora of studies that have investigated changes in the host transcriptome upon ZIKV and DENV infection, focused on changes in gene expression and splice isoforms. These studies have spanned across various cell types and have quite a few results in common. First, the up-regulation of immune response and inflammation as well as lipid metabolism and the down-regulation of genes involved in fetal development in the case of ZIKV infection and the dys-regulation of metabolism in response to DENV infection. Second, AS events seemed to be centered on genes involved in cell death, RNA processing, and development (*76, 77, 79-82*). Although it is promising that these results seem to be rather conserved across various cell types, it also does not necessarily explain how infection of the placenta may be different from neural tissue, or even tissues less concerning to human health with regards to ZIKV infection. On this note we can focus on the differences between ZIKV infection of the placental cells compared to infection of the Huh7 cells. Interestingly, there is actually little over-lap in DEGs and APA genes across the three data sets (**Figure 3.35A and B** respectively). One notable difference is the down-regulation (3' UTR lengthening) of complement and

Figure 3.35: Comparison of each sample. A) Venn diagram comparing ZIKV JEG3, DENV JEG3 and ZIKV Huh7 enrichment of genes with regards to either 3' UTR lengthening, 3' UTR shortening, up-regulation, or down-regulation. B) Venn diagram comparing ZIKV JEG3, DENV JEG3 and ZIKV Huh7 enrichment of pathways with regards to either 3' UTR lengthening, 3' UTR shortening, up-regulation, or down-regulation.



coagulation when infected with ZIKV, seen in our PAC-seq results. Although interesting, as the development of microcephaly in infants born to ZIKV-infected mothers has been associated with dys-regulation of inflammation, this suggests that the inability to form thrombi may have some impact on ZIKV infection. It is also possible that this occurs to reduce the amount of host-damage done by the virus.

In addition, the result of PAC-seq analysis using the *DPAC* pipeline provides information on the use of alternative terminal exons. We compared these results to the alternatively 3' spliced findings from our long read sequencing with *Flair* analysis. What we found was that for genes already annotated as having multiple exons with a poly-A cluster (different terminal exons) we were able to find the differential use of those terminal exons in our PAC-seq data and they corresponded with the changes found in our *Flair* results. For genes with a single terminal exon with a previously annotated poly-A cluster we saw consistent 100% usage of this terminal exon across the samples in the PAC-seq data. This was also found in the Flair results. However, if what was detected by Flair was changes in the length of that terminal exon, typically shown including part of the 5' intron or including what would otherwise be the 3' UTR of that gene, this could not be identified as any type of change in terminal exon in *DPAC*. This suggests that were these altered (longer or shorter) terminal exons annotated as different exons then *DPAC* would be able to detect the difference.

Where this study sets itself apart is the investigation of APA. The enrichment of genes associated with splicing and mRNA processing in transcripts with shortened 3' UTRs, suggests that these mRNAs are likely to be dys-regulated, in particular de-regulated as there is less space for regulatory proteins to bind in the UTR. It had been previously noted that sfRNA or even the NS5 protein interacting with the spliceosome that caused changes in splice patterns in response to

flavivirus infection (*84, 88, 89*). Our findings suggest that this might not be the whole picture. In these data sets RNA degradation and miRNA biosynthesis also appear to be dys-regulated – suggesting that ZIKV may be finding a way to either a) evade these processes or b) cause them to act on host transcripts and allowing the translation of more viral proteins. Lengthened 3' UTRs of transcripts associated with host immunity also suggest that ZIKV infection alters these UTRs to increase their potential for regulation – serving as an additional means of evading this response. Finally, lengthening of transcripts important for fetal development and cell differentiation suggest that although these pathways may not be implicated in our DEG results, their altered 3' UTRs potentially suggests their down-regulation via increased regulation. Similar to our comparisons of specific DEGs, there are few conserved APA events across the three data sets.

Across all datasets we extracted the nucleotide sequences around each annotated PAC and put these sequences into DREME (*159*) to determine if any motifs for RNA binding proteins were enriched (results in Tables 3-5). As expected, we found enrichment for motifs recognized by PABP1 or PABPC1 in both down- and up-regulated PACs in each data set. Additionally, we were able to find binding sites for splicing factors in the regions flanking dys-regulated PACs across all data sets. Although the pattern is not consistent across each set (splicing factor binding sites found in both up- and down-regulated PACs for ZIKV infected Huh7, only in up-regulated in ZIKV infected JEG3, and only in down-regulated in DENV infected JEG3), these results, in combination with the finding of spliceosomal transcripts undergoing APA, suggest that poly-adenylation patterns may impact splicing patterns. Interestingly, this appears to be a two-way street as we found evidence that suggests that increased expression of SRSF11 may be the cause of many up-regulated PACs.

Our findings suggest that changes in poly-adenylation patterns may prove as an important means of transcriptomic change in response to ZIKV infection, as this provides supplement to the AS (*Flair*) or DGE (*DESeq2*) studies. In this study in particular, our APA results suggest that ZIKV may be able to evade the immune response because of dys-regulation of siRNA and miRNA synthesis and surveillance – which was not evident from the *DESeq2* or *Flair* results. This may also suggest, in combination with the AS results which implicated changes to proteins involved in transcription and translation, that this may serve as a mechanism to ensure translation of viral proteins. In the future, it may be interesting to identify how ZIKV induces these changes in APA, potentially via some viral protein interacting with poly-A polymerase or other factor involved in the poly-adenylation process. Presumably, this would be one of the three soluble viral proteins, either NS5, NS3, or capsid. It may also be interesting to see how any of these soluble proteins may result in other changes by determining what nucleic acids they bind, and any host-proteins they interact with, which has been started to some extent, but a the full extent to which each of these viral proteins promote pathogenesis is not entirely known.

In addition to investigating changes to the host transcriptome in human placental cells upon ZIKV and DENV infection we also looked into the capsid-RNA interactions within ZIKV particles using NGS and a novel method called v-PAR-CL. Our original hypothesis was that although capsid appears to have the propensity to bind with various nucleic acids (both ss- and ds-, RNA and DNA) that the environment within the replication factories at the ER membrane may increase specificity of ZIKV C binding. In line with this idea, we did actually find increased PAR-CL signal in various regions of the ZIKV Dakar genome, as well as little to no signal in other regions. These peaks in signal indicate specificity binding sites of ZIKV C to the vRNA in virions. Interestingly, these peak regions are fairly well conserved across ZIKV genomes. However, motif enrichment analysis

over the entire Dakar genome and for the discrete regions around these peaks was inconclusive. Additionally, we compared the PAR-CL signals to previous annotations of the genome structure within MR766 particles (after alignment) and found that ZIKV appears to bind both ds- and ss-RNA in this context.

Although both of these findings suggest no binding specificity for ZIKV C, and are congruous with previous flavivirus capsid binding studies, this fails to explain why we see peaks in PAR-CL signal in discrete genome regions. It has previously been noted that a non-structural protein (NS2A) initially binds the vRNA and chaperones it to capsid prior to packaging. Additionally, we may consider that folding of the vRNA occurs as the RNA is being synthesized, folding locally first, then forming tertiary base-pairing as extension continues. As others have been able to investigate the overall structure of ZIKV RNA in virions, this suggests that base-pairing and folding patterns are consistent from particle to particle. This is the case for FHV for example, with a dodecahedral cage of RNA easily discernible from EM images. We might expect that the globular nature of capsid proteins prevent the resolution of distinct RNA structures. This has led me to a new hypothesis, that capsid proteins are simply interacting with the regions of the genome that are on the exterior of the final vRNA structure, similar to holding a ball of yarn in your hand where you can only make contact with regions of the yard on the outside. In this case the ball of vRNA with capsid around it is not covalently attached to the membrane, similar to a ball compass.

Considering that these PAR-CL regions are highly conserved across ZIKV strains, it may be interesting to see if these interactions also occur in these regions in other Zika strains. That being said, future studies could elucidate if the sequences identified using v-PAR-CL are important for viral packaging by introduction synonymous mutations at those sites, potentially hindering these interactions. Additional rounds of v-PAR-CL can be done to see if binding still occurs in this

region, but if packaging is impeded it will be worth investigating the impact on viral fitness as well, via plaque assays and growth curves. Further, as the exact residues of capsid that interact with vRNA *in virio* has been hypothesized as the positively charged N-terminus of capsid, but various basic residues on the surface of flavivirus capsid have been implicated in interactions with nucleic acids – studies mutating these positively charged regions of capsid and investigating how this may perturb viral fitness are of interest as well. These studies are critical as there is currently no treatment or vaccine available for ZIKV.

Chapter 4 Opioid use disorder (OUD) as a co-morbidity for viral infections

Although none of the data in this chapter has been previously published, the manuscript below for opioid studies is in collaboration with Dr. Cunningham's lab in the UTMB Center for Addiction Research (CAR). This is in preparation for submission to Nature Communications.

Sotcheff, S.; Zheng, J.; Stafford, S.; Merritt, C.; Smith, A.; Routh, A.; Anastasio, N.; Mendoza, I.; Cunningham, K. Acute withdrawal from fentanyl use may exacerbate severity of COVID-19. (in Preparation)

There have been recent studies that have documented the impact of SARS-CoV-2 upon the central nervous system leading to neurological symptoms including dysphoria, confusion and delirium (*192-196*). Additionally, this pandemic has coincided with a drug overdose crisis (*197-199*). Between June 2020 and June 2021 there was a reported 18.2 percent increase in drug overdose deaths in the United States (*200*). In fact, since December 2019 the number of drug overdose deaths in a 12 month window has been increasing drastically. From 2016 to 2019 this number has remained around 70K deaths in the U.S. per year. In June 2021 this number was just under 100K for the previous year (*41*). Many of these deaths are occurring in individuals who survived previous overdoses. The incidence of repeated overdose events is a pattern consistent with opioid use disorder (OUD). Although opioids like fentanyl and oxycodone are regularly prescribed to treat pain, this does not mean that these patients are immune to OUD development, actually 15-25% of overdose survivors are pain patients (*201*). Both prescription and elicit (e.g. heroin) opioids act in deep brain regions and cause respiratory depression contributing to

overdose deaths. That being said, patients using opioids may be especially vulnerable to COVID-19 impacts on the respiratory, cardiovascular and central nervous system.

Opioids have been shown to modulate the immune system in complex ways, consistent with the expression of opioid receptors in immune cells. For example, clearance of influenza virus from rat lungs is impeded by morphine use. It has also been shown that medications that aid in the treatment of OUD, such as methadone, increase viral replication of influenza *in vivo* and *in vitro (202)*. Interleukin 6 (IL-6) has been shown to be up-regulated in response to SARS-CoV-2 (*203*) and interestingly high expression of IL-6 has a negative impact on individuals undergoing medication-assisted treatment for OUD (*201*). Additionally, fentanyl self-administration in male rats was shown to increase expression of various cyto- and chemokines, altering the immune response in the nucleus accumbens (NAc) but not the hippocampus (*44*). Other studies of peripheral blood suggest that opioid usage hinders the antiviral response (*204*). Finally, one report has shown that COVID-19 mortality is higher in opium-addicted populations compared to non-users (*205*). Considering the host response to various viral infections is similar and that flaviviruses can be neurotropic, we posit that opioid use may also exacerbate neurological disease in the case of flavivirus infections.

INVESTIGATING OPIOID USE AS AN INDICATOR OF COVID-19 OR ZIKV DISEASE SEVERITY

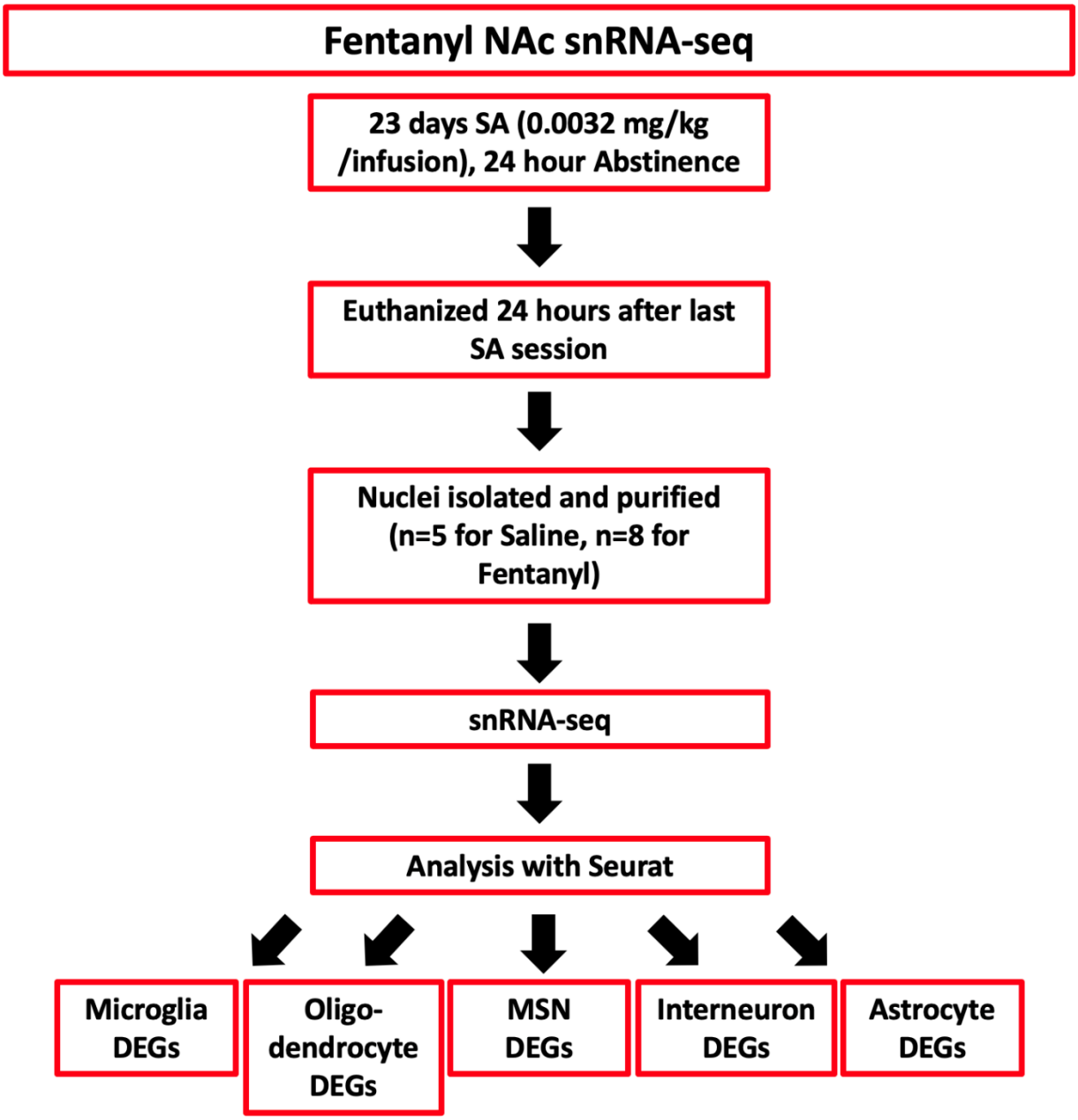
The ventral tegmental area (VTA) and nucleus accumbens (NAc) are the primary brain regions where addictive drugs regulate the dopamine pathway (*38*). The NAc is one of two parts of the ventral striatum. The VTA is located in the midbrain and consists of primarily glutamatergic, GABA-ergic and dopaminergic neurons (*148, 149*), whereas the NAc is primarily medium spiny neurons. Reward-seeking behavior elicited by use of opioids and other addictive substances or predictive uses (ex. seeing paraphernalia) is caused by release of dopamine in the

ventral striatum from neurons originating in the VTA (*149-153*). These neurons are referred to as the mesolimbic pathway, which is involved in salience, motivation, and reinforcement learning (*38, 154*). Therefore, to investigate how opioid use may serve as an indication of severity of COVID-19, we conducted transcriptomic studies on nuclei from the NAc and total cellular RNA from both the NAc and the VTA.

Expression of genes involved in SARS-CoV-2 or ZIKV infection and altered make-up of the NAc upon fentanyl self-administration

In our work with Dr. Kathryn Cunningham's group in the Center for Addiction Research at UTMB we have generated single nuclei RNA sequencing (snRNAseq) libraries from rat NAc. The experiment was designed to determine if fentanyl withdrawal has negative impacts on the immune system and to determine if genes necessary for SARS-CoV-2 cell entry are expressed in specific cell types within the NAc. We hypothesized that OUD is a co-morbidity for CoVID-19. There was a total of 13 rats, 5 that were self-administering (SA) saline, and 8 SA fentanyl. Rats SA for 23 days, following which they were forced to be abstinent from their treatment for 24 hours prior to euthanasia (**Figure 4.1**). Brain was dissected by Sonja Stutz in Dr. Cunningham's lab and samples were frozen until nuclei could be extracted for sequencing. Nuclei extraction was completed by Susan Stafford and Dr. Junying Zheng in Dr. Cunningham's group and filtered nuclei were counted and provided for 10X library preparation and subsequent data analysis. Libraries were sequenced on the NovaSeq (an Illumina platform) at Baylor College of Medicine in Houston, TX yielding ~400M reads/sample.

Figure 4.1: Experimental design for the acute withdrawal study, investigating single nuclei expression in the nucleus accumbens (NAc) of rats self-administering fentanyl or saline (control). Schematic created by S. Sotcheff in Microsoft Powerpoint.



Of the thirteen rat NAcs, I generated high-quality 10X libraries for all but one (a fentanyl SA rat) as determined by BioAnalyzer trace. The number of cells, reads, and genes per cell are noted in **Table 4.1**. There was data for ~100K cells, an average of about 8.2K per sample. These were loaded into R using a package from the Satija lab called Seurat (*162*). After filtering for number of genes per cell, percentage of mitochondrial genes expressed and doublet score, we retained ~52K cells for subsequent analysis. These were separated into 26 clusters by Seurat (**Figure 4.2,** breakdown of number of cells per sample/cluster in Appendix H). The following genes were used to generate feature plots and aid in determining cell types of the clusters: Snhg11 and Syt6 (neurons); Ache and Chat (cholinergic); Grm8, Drd1, Drd2, and Drd3 (medium spiny neurons); Pdgfra (Ng2, an oligodendrocyte precursor); Mog and Sox10 (oligodendrocytes); Gfap and Gja1 (astrocytes); Arghap15 and Cx3cr1 (microglia); Sst, Elavl2 and Kit (interneurons); and Slc17a7 (glutamatergic) (**Figure 4.3**). One cluster expressed a single neuronal marker but no others, further investigation of that cluster revealed expression of Tpx2 and Cenpf – indicating that cells in this cluster were undergoing mitosis (**Figure 4.4**). Clusters were re-labelled by cell-type to generate **Figure 4.5** (Appendix I). Note that three clusters appeared to express markers for two cell types (and indeed the UMAP plot shows these clusters located in two places) shared between the cell types. These clusters were excluded from differential gene analysis.

Before differential gene expression analysis, we noticed that the relative number of cells from fentanyl SA rats compared to saline SA rats in the glutamatergic neuron and mitosis clusters was skewed. The cells in these clusters are primarily from fentanyl SA rats. On average there are 10 of glutamatergic neurons in the NAc datasets of the saline rats, and 261 in the fentanyl SA rats. Similarly, there is an average of 13 cells undergoing mitosis in the NAc of the

Table 4.1: Table summarizing the total number of cells, genes and reads per cell, as well as total number of reads per sample in our single nuclei RNAseq (snRNAseq) experiment.

Sample	# cells	Avg reads/cell	Median genes/cell	Total reads
1_BK4_S_introns	5,539	48,099	1,832	266,419,976
2_BL5_S_introns	9,212	31,927	1,645	294,110,163
3_BK7_F_introns	7,247	38,545	1632	279,334,573
4_BK8_F_introns	6,305	49,872	1,346	314,441,865
5_BK1_S_introns	10,857	31,826	1,131	345,535,655
6_BK6_F_introns	10,721	38,303	1,193	410,647,745
7_BL9_F_introns	8.154	44,041	1,296	359,111,078
8_BL12_F_introns	8,700	34,441	1,155	299,634,647
9_BK2_F_introns	6,310	97,469	1,723	615,029,408
10_BL2_S_introns	10,374	33,326	904	345,727,716
11_BK5_F_introns	6,758	45,287	1,487	306,047,889
12_BK11_F_introns	8,776	36,957	1,400	324,334,777

Figure 4.2: UMAP plot of snRNAseq data from all twelve rat nucleus accumbens (NAc) samples where the cells (each individual point is a single nuclei) are clustered into groups based on similarities or differences in expression profiles. Seurat (*162*) provided us with 26 clusters based on the parameters we set. Plot created by S. Sotcheff using Seurat in R.

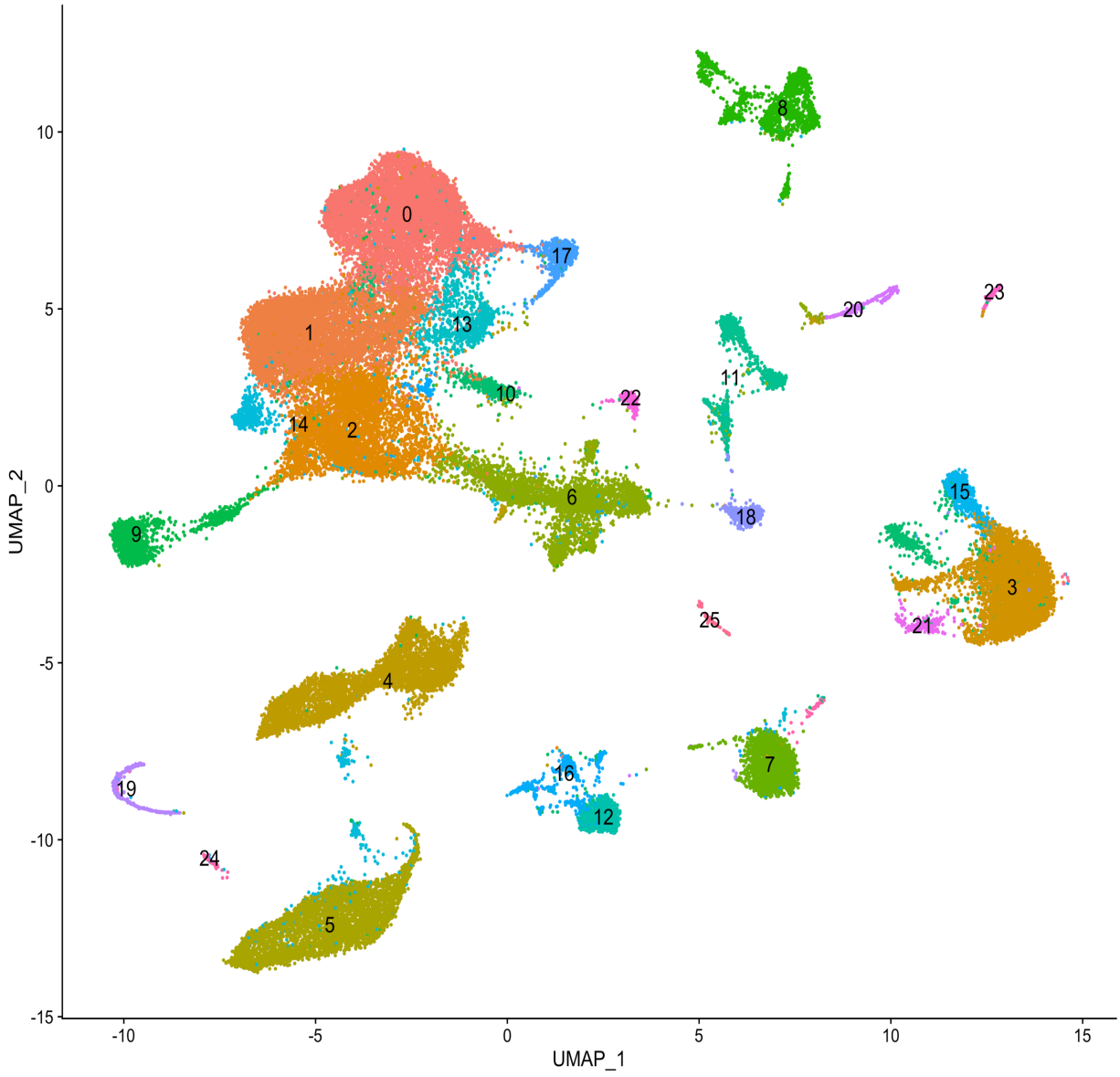


Figure 4.3: Feature plots showing expression in all cells of various expression markers as highlighted in the text. Snhg11 and Syt6 (neurons); Ache and Chat (cholinergic); Grm8, Drd1, Drd2, and Drd3 (medium spiny neurons); Pdgfra (Ng2, an oligodendrocyte precursor); Mog and Sox10 (oligodendrocytes); Gfap and Gja1 (astrocytes); Arghap15 and Cx3cr1 (microglia); Sst, Elavl2 and Kit (interneurons); and Slc17a7 (glutamatergic). Feature plots created by S. Sotcheff using Seurat in R.

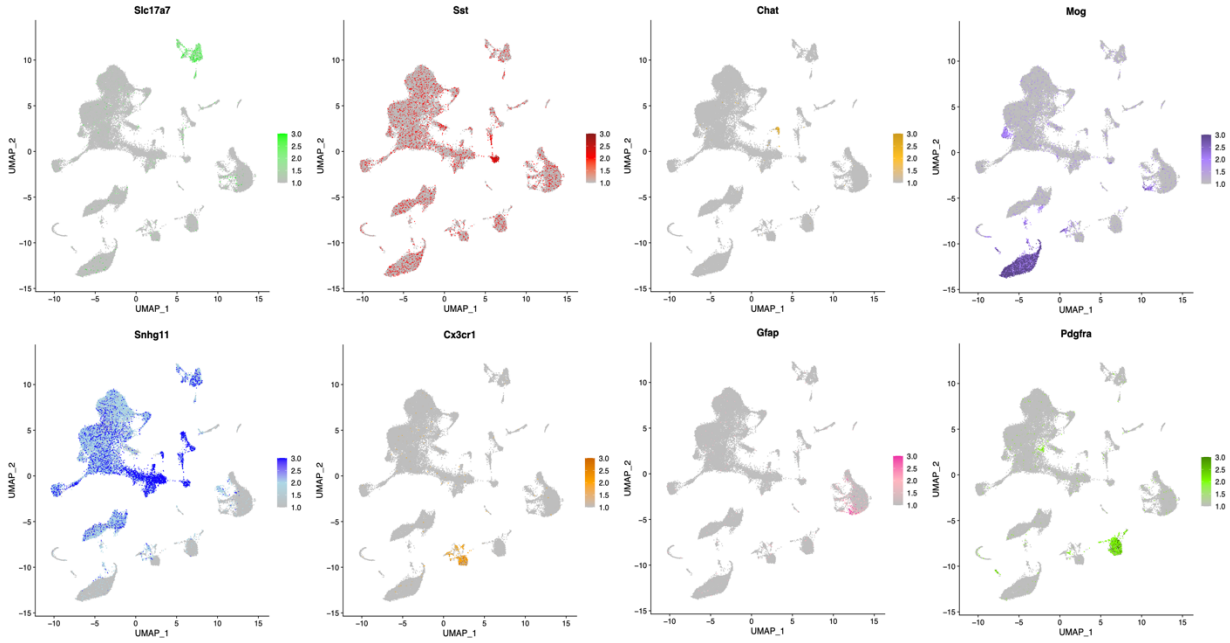


Figure 4.4: Dotplot displaying expression of various marker genes across all clusters – aiding in cell type determination. Plot created by S. Sotcheff using Seurat in R.

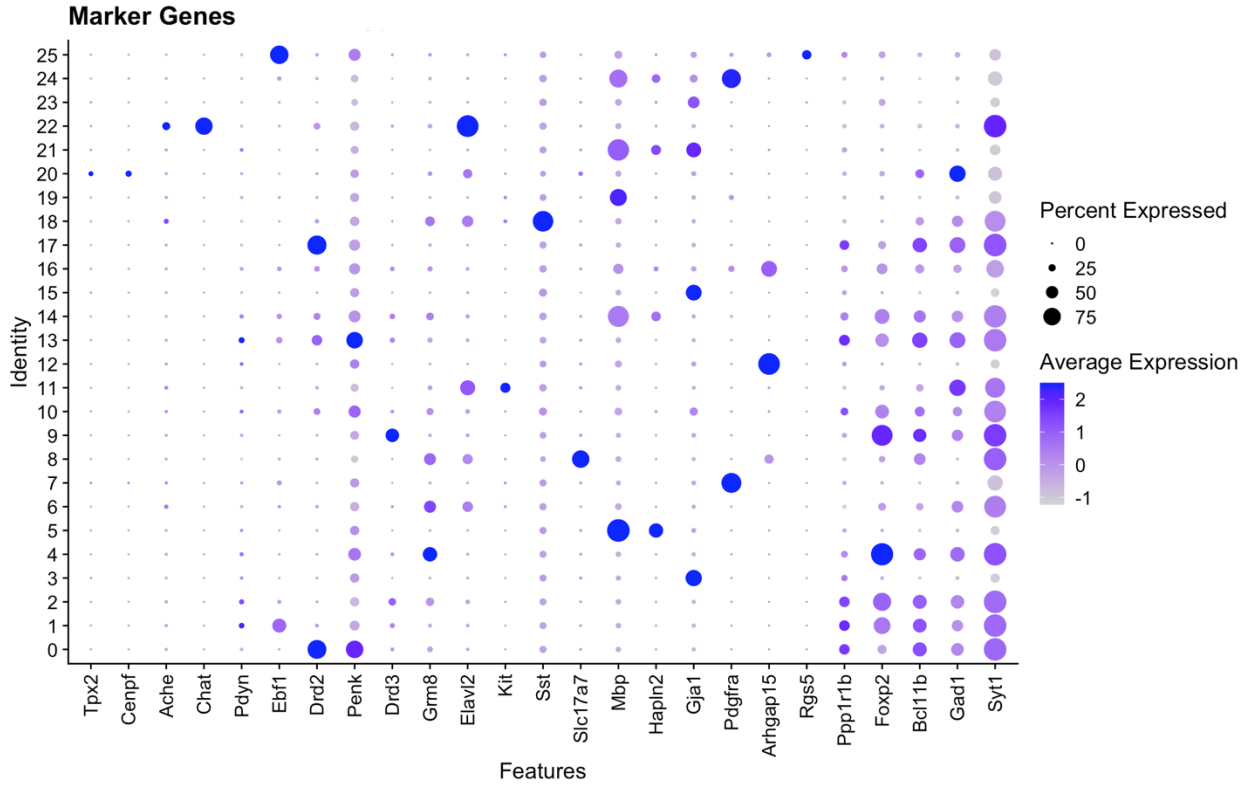
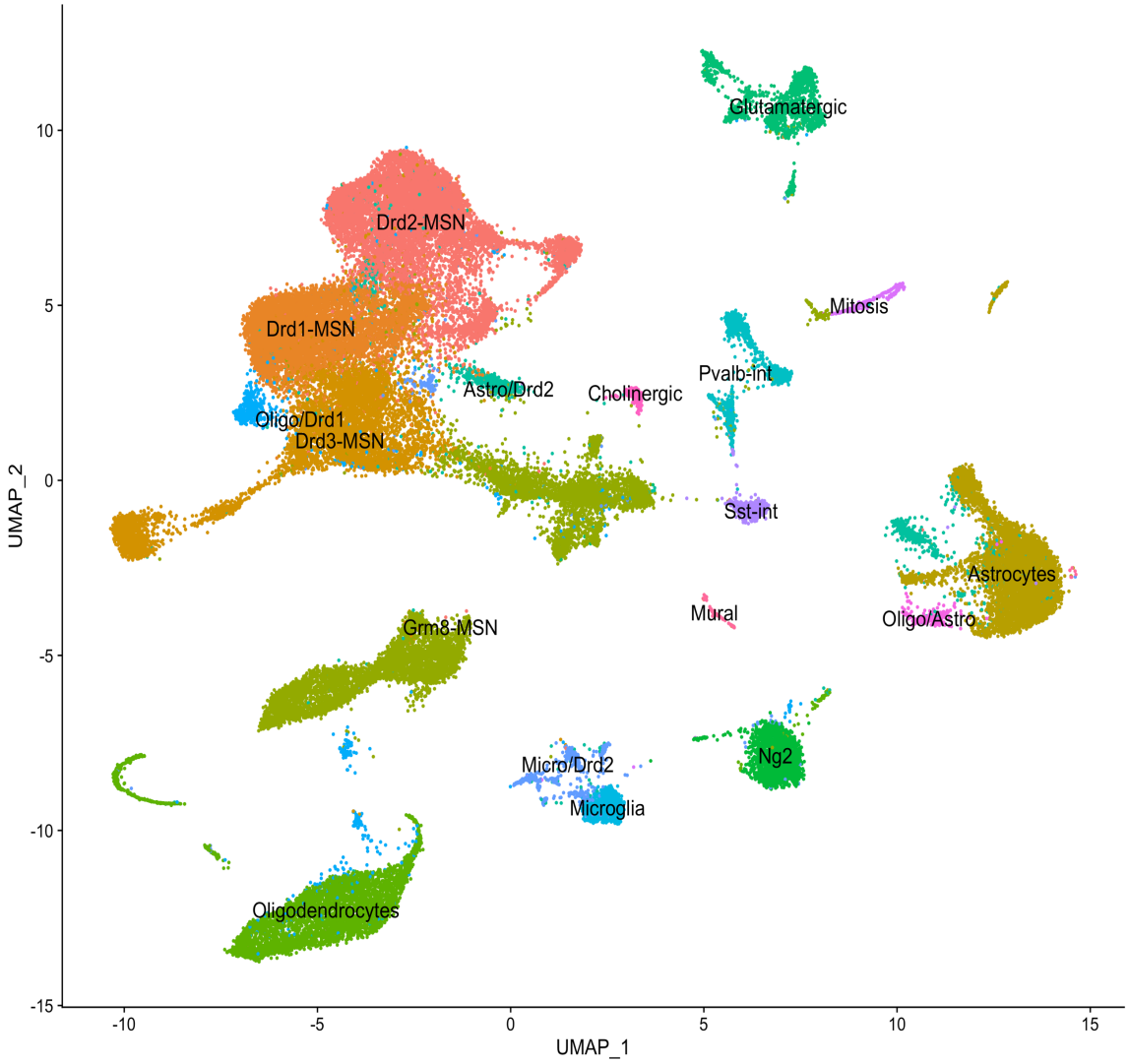


Figure 4.5: UMAP plot of snRNAseq data from all twelve rat nucleus accumbens (NAc) samples where the cells (each individual point is a single nuclei) are clustered into groups based on similarities or differences in expression profiles. Here we have re-labelled each cluster to its corresponding cell type based on marker gene expression. Plot created by S. Sotcheff using Seurat in R.

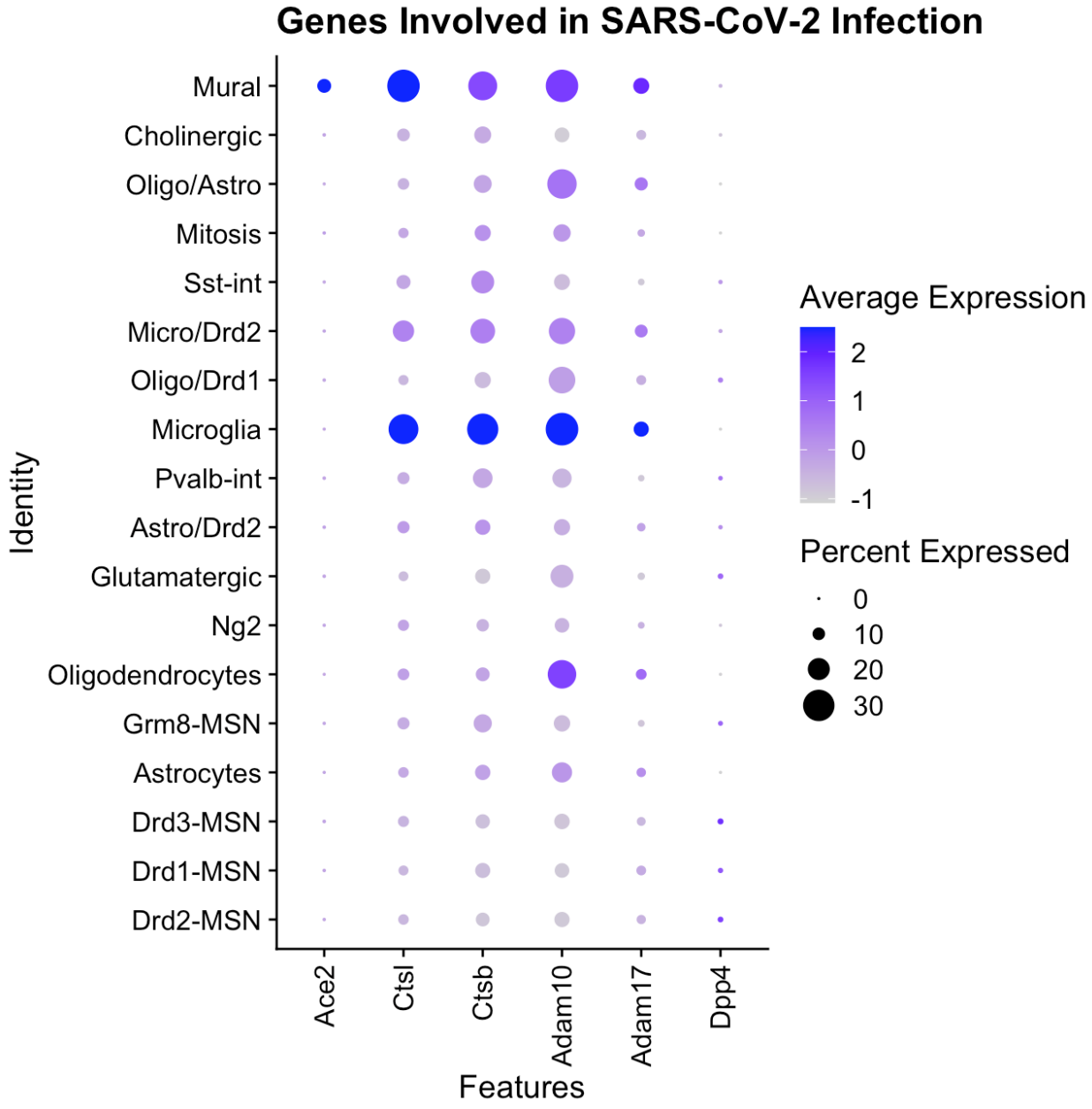


saline SA rats and 46 in the fentanyl SA rats. This suggests a shift in the rate of cell proliferation and distribution of cell types in the NAc in response to either fentanyl or the forced abstinence from fentanyl. This in and of itself is an interesting find considering adult rats should have very few neurons undergoing mitosis and the Cunningham group is in the process of validating these results.

Differential gene expression analysis was conducted as previously described: clustered cell types are subset from the original complete data set, treatments are labelled, and markers are found that distinguish cells from the treatments from each other. Although this was conducted for all cell types, we will focus on the cell types that express genes that have previously been associated with SARS-CoV-2 infection (ACE2, TMPRSS2, Cathepsins L and B, ADAMs 17 and 10, and Dpp4) (*206-211*) or ZIKV permissibility (*32, 212*) and modulation of inflammation (*32, 213*) (AXL and TLR3 respectively).

While most of the proteins listed above are enzymes capable of cleaving proteins or peptide bonds in either host proteins or the coronavirus spike protein – enhancing SARS-CoV-2 infection; angiotensin converting enzyme 2 (ACE2) is highly expressed in the lungs and has been shown to act as a receptor for the virus (*209, 214, 215*). Previous studies have indicated that the central nervous system (CNS) can be infected (*192-195*), but the expression of ACE2 in the brain was not clear. Our data show that cathepsins L and B and ADAMs 17 and 10 are expressed in three cell types in the NAc: oligodendrocytes, microglia, and mural cells (**Figure 4.6**). Interestingly, only one of these cell types expresses ACE2: mural cells. Mural cells are epithelial cells associated with blood vessels such as vascular smooth muscle cells and pericytes. This is the smallest population in the data set, accounting for about 0.25% of all cells sequenced. However, their expression of ACE2 as well as various other genes associated with SARS-CoV-2

Figure 4.6: Dotplot displaying expression of genes necessary for SARS-CoV-2 entry in the various cell types in the nucleus accumbens. Plot created by S. Sotcheff using Seurat in R.



infection, indicates that these cells have the potential to be infected. When we investigated the expression of receptors for ZIKV in our single cell data, we found that TLR3 was expressed in microglia and AXL was expressed in both astrocytes and mural cells **(Figures 4.7 and 4.8).**

Only a handful cell types express these genes associated with SARS-CoV-2 infection, so we investigated how these cells in particular were impacted by cessation of fentanyl self-administration. Within mural cells, genes that were dys-regulated appear to mostly be involved in cytoskeleton organization or re-arrangement (**Figure 4.9**). However, when we investigated pathways that may alter response to a viral infection, we saw that MAVS, IRF3/7, IFIT2, and ISG20 were up-regulated in response to abstinence from regular fentanyl use. This suggests fentanyl or acute withdrawal stimulates type I interferon pathways and the antiviral response. Additionally, the expression of ACE2 was up-regulated in the fentanyl SA rats, suggesting that fentanyl or withdrawal may make individuals more susceptible to SARS-CoV-2 infection of the CNS. Microglia expressed cathepsins B and L as well as ADAMs 17 and 10. However, none of these were differentially expressed between the fentanyl abstinence and saline rats. We did find that IL-6 was up-regulated in a sub-set of microglia, as was μOR and STING. Up-regulation of STING indicates that other pro-inflammatory cytokines are likely to be up-regulated in NAc microglia upon fentanyl use or acute withdrawal.

Oxycodone triggers differential gene expression in rat NAc and VTA

Also in collaboration with the Cunningham group in the CAR at UTMB, we investigated differential expression in both the NAc and VTA after SA of oxycodone or saline with and without a period of forced abstinence. Experimental design for these studies is highlighted in **Figure 4.10**. Rat brains were dissected after sacrifice and RNA was extracted by Christina

Figure 4.7: Feature plots showing cells expressing TLR3 and AXL, genes involved in ZIKV infection in the brain. Plots created by S. Sotcheff using Seurat in R.

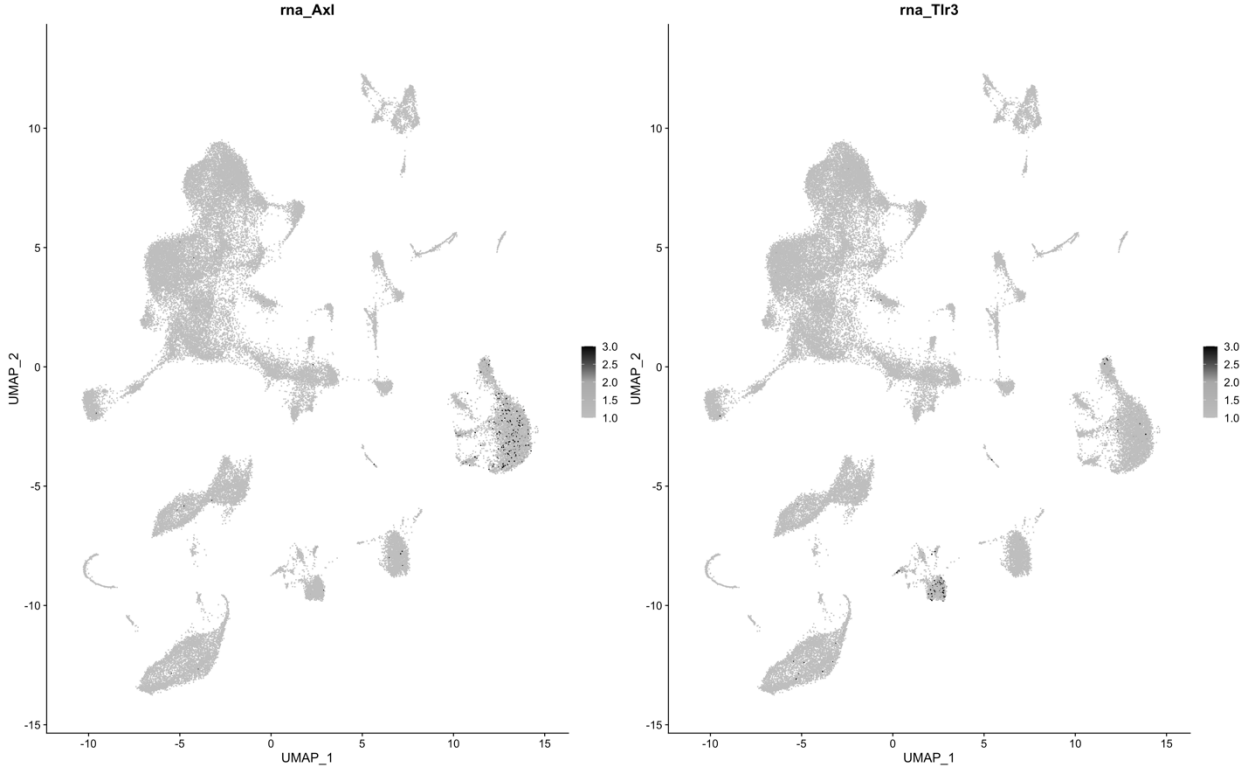


Figure 4.8: Dotplot showing that AXL is expressed by mural cells and astrocytes and TLR3 is expressed by microglia in the NAc. Plot created by S. Sotcheff using Seurat in R.

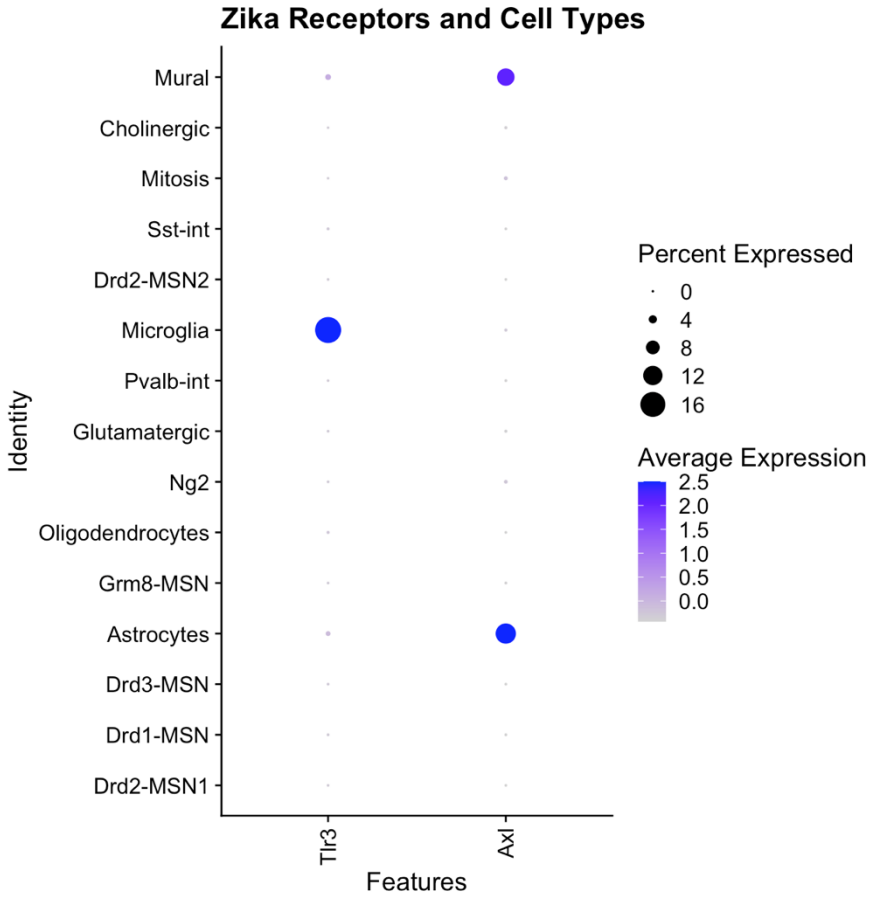


Figure 4.9: Heatmap of differentially expressed genes in mural cells upon acute withdrawal from fentanyl self-administration. A majority of these genes are involved in cytoskeleton organization or re-arrangement. Plot created by S. Sotcheff using Seurat in R.

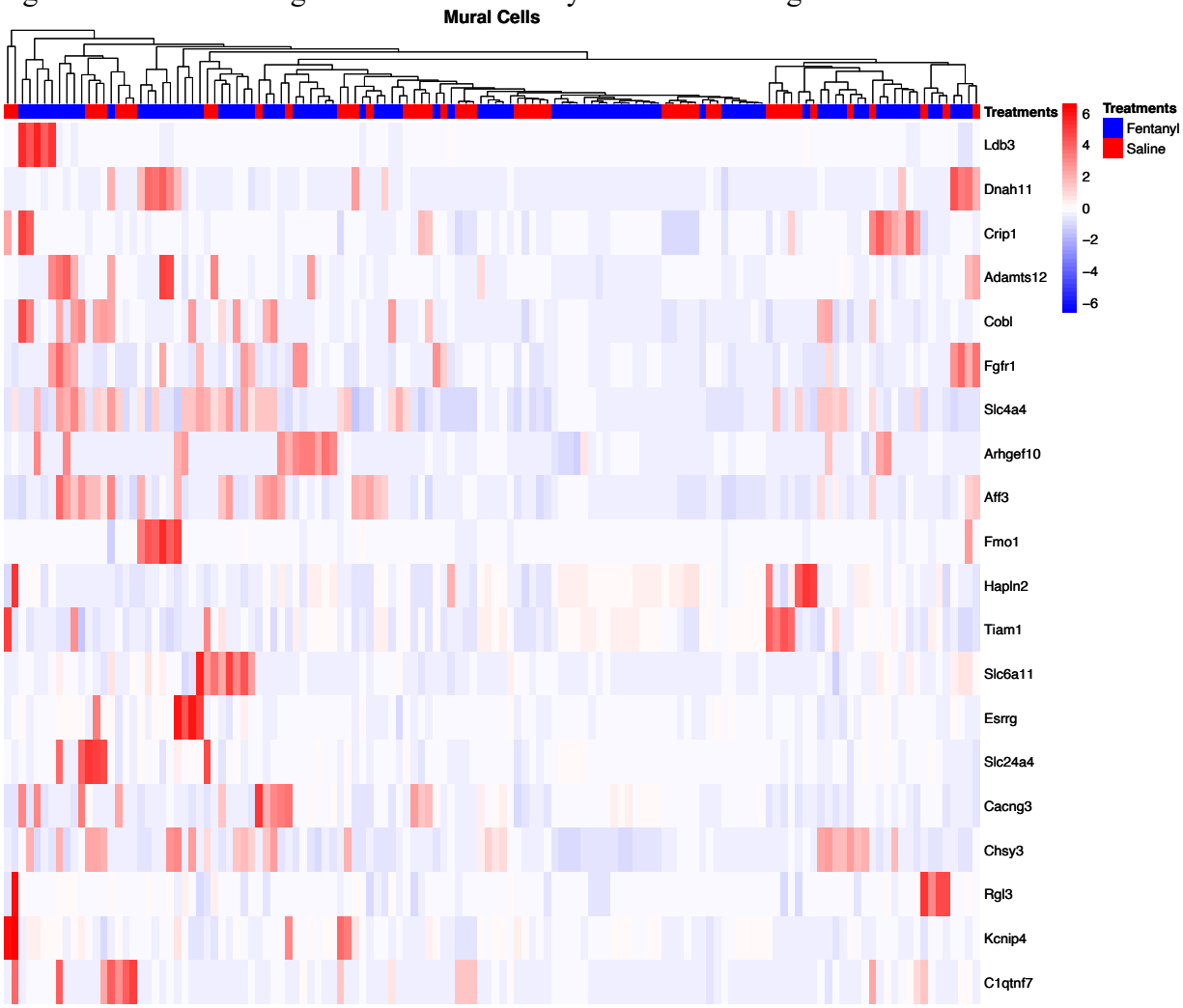
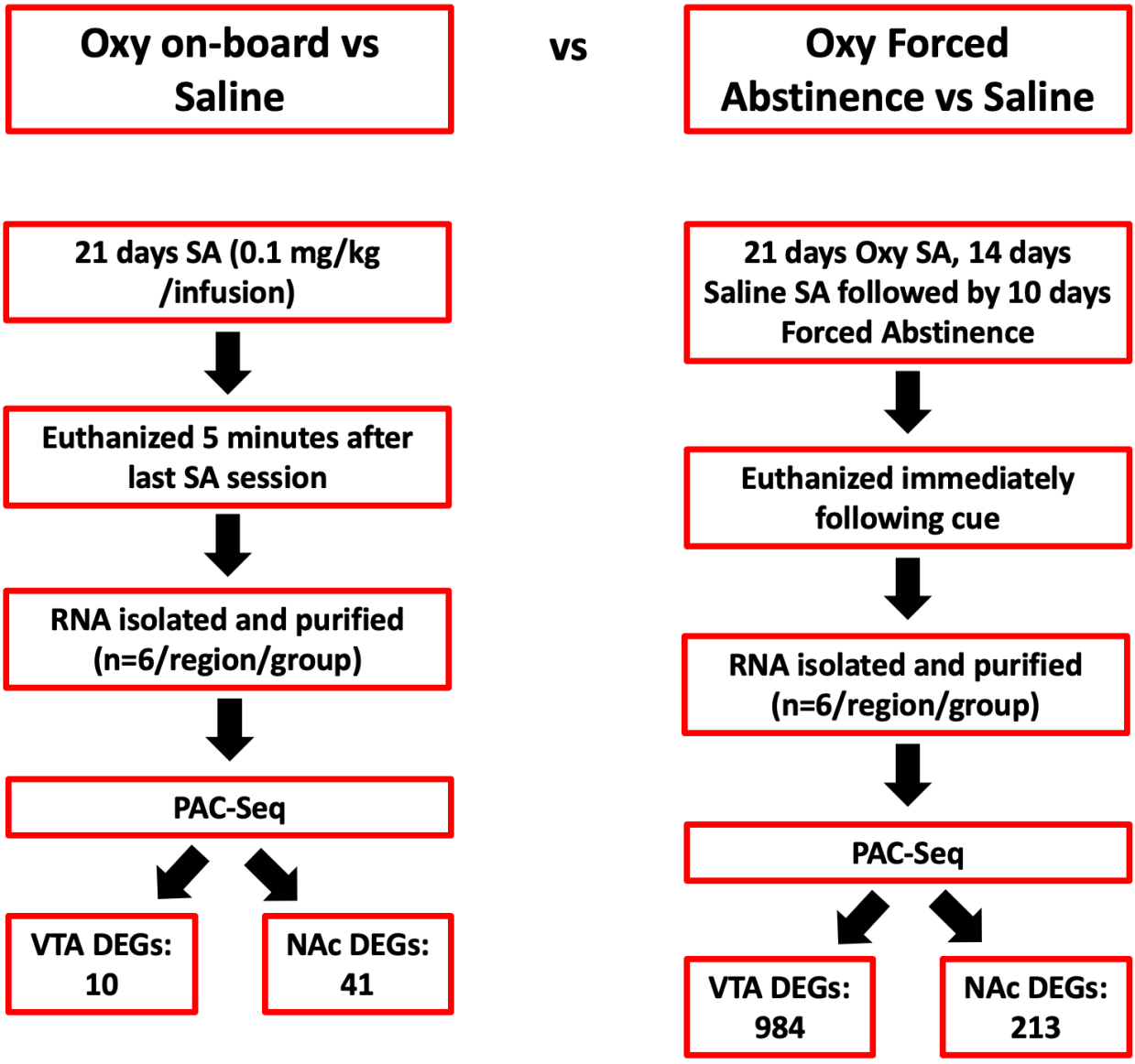


Figure 4.10: Experimental design for Oxy-On Board and Oxy Forced Abstinence studies to investigate differential gene expression in the nucleus accumbens (NAc) and ventral tegmental area (VTA) after oxycodone use. Schematic created by S. Sotcheff in Microsoft Excel.



Merritt from Dr. Cunningham's lab. We determined differential gene expression using *DPAC* analysis after PAC-seq library construction and sequencing. Quality of these data was assessed by reviewing the alignment of the bedgraph files in the UCSC Genome Browser (**Figure 4.11**). Principle component analysis showed distinct clustering for the saline rats, away from the more variable oxycodone rats, in each experiment. A representative plot, for the NAc forced abstinence experiment is shown in **Figure 4.12**.

Our analysis detected between 9,500 and 10,200 genes per experiment per brain region. Differentially expressed genes (DEGs) were identified by filtering the resultant DESeq2 results for a p-adjusted value of less than 0.1 and a log2FoldChange greater than 0.58 (up-regulation) or less than -0.58 (down-regulation). We found 10 DEGs in the VTA and 41 in the NAc when rats were sacrificed 5 minutes after their last SA session and 984 DEGs and 213 DEGs in the VTA and NAc respectively when rats were sacrificed following an extended period of forced abstinence (**Figure 4.10**). In the NAc we found 7 up-regulated and 3 down-regulated genes in response to oxycodone on-board (**Figure 4.13**) and 96 up-regulated and 117 down-regulated genes in response to oxycodone after forced abstinence (**Figure 4.14**). The up-regulated genes in the NAc oxy on-board study include multiple heat shock proteins and calreticulin, a chaperone protein involved in protein folding and calcium homeostasis. The down-regulated genes in the NAc oxy on-board study are Gnl3, Net1, and Id3: a protein in the nucleolus that interacts with p53 and involved in stem cell proliferation, neuroepithelial cell transforming 1 (a Rho guanine nucleotide exchange factor), and inhibitor of DNA binding 3 respectively. In the VTA we found 25 up- and 16 down-regulated genes in response to oxycodone on-board (**Figure 4.15**) and 42 up- and 942 down-regulated in response to oxycodone forced abstinence (**Figure 4.16**). From this we see the trend that on-board oxycodone resulted in more up-regulated genes compared to

Figure 4.11: Bedgraph alignment for gene EGR4 in the PAC-seq data generated for these studies. We see increased coverage at the poly-A site and coverage tapers off as we move into the 3' region of the gene. As the reverse transcription reaction for PAC-seq library preparation includes priming from the poly-A tail and stochastic termination of cDNA synthesis this validates that the quality of the libraries is sufficient.

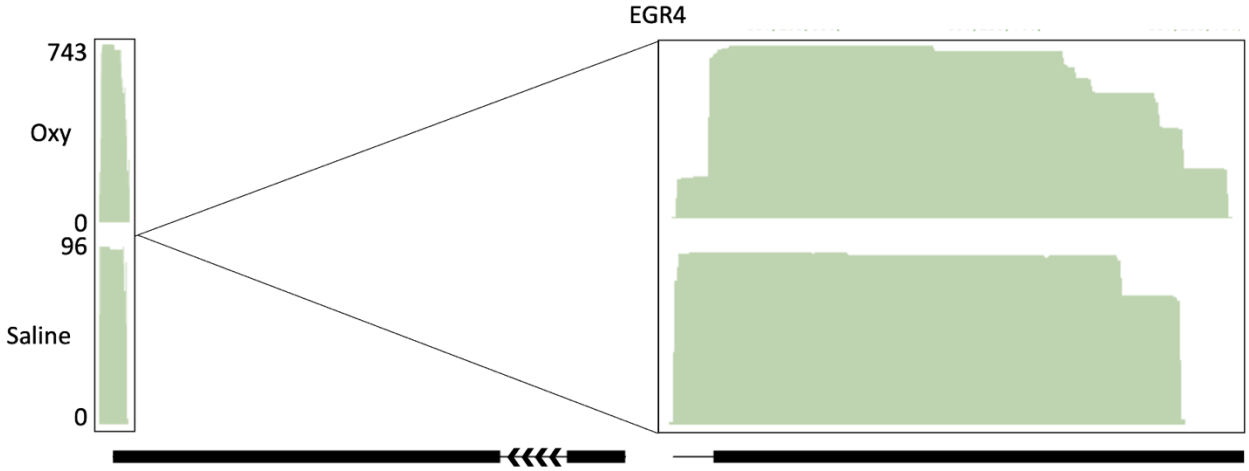


Figure 4.12: Principal component analysis (PCA) plot for the PAC-seq data generated from rat nucleus accumbens (NAc) mRNAs after saline or oxycodone forced abstinence. The PCA plot shows clear clustering of the saline self-administering rats and larger variations in rats that were self-administering oxycodone. Plot created by S. Sotcheff using DESeq2 in R.

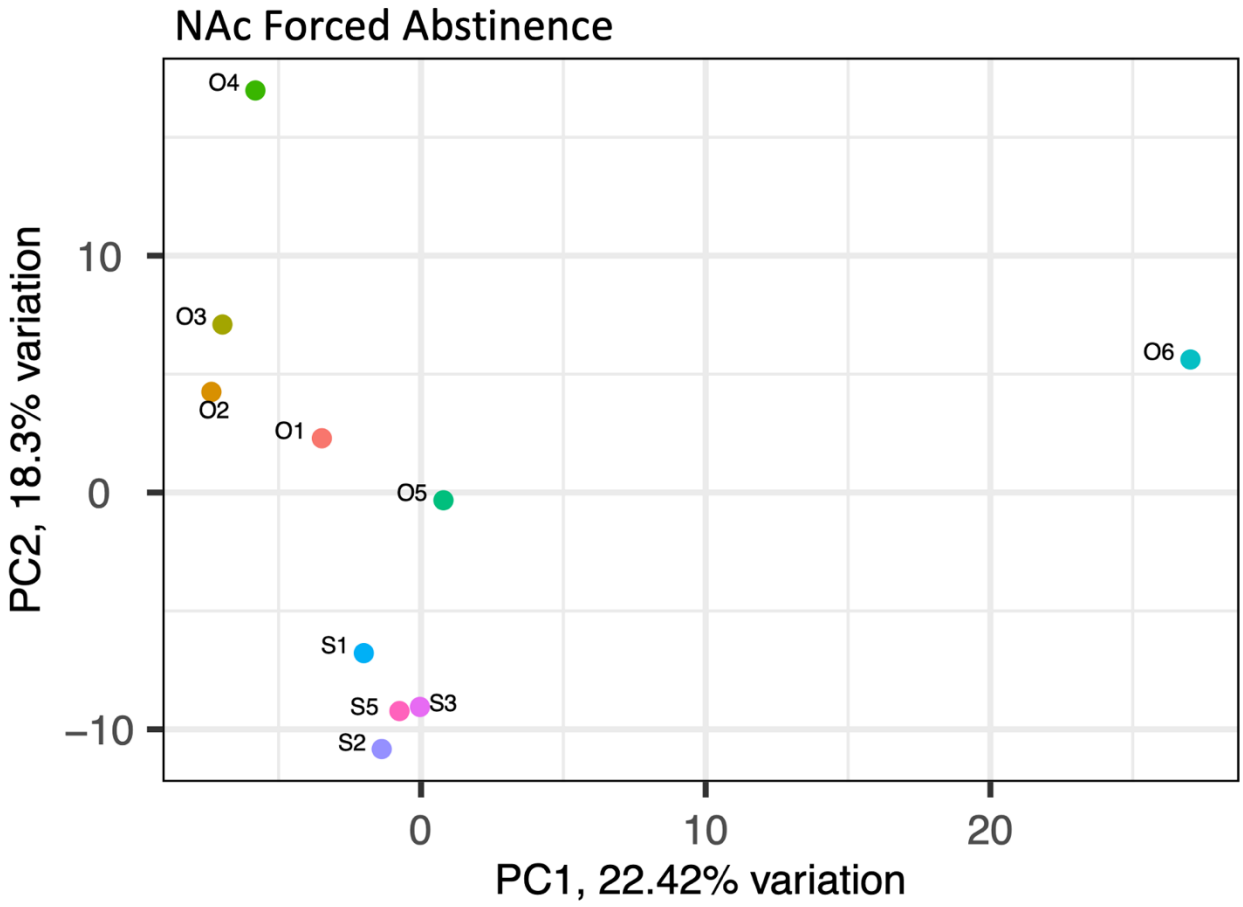


Figure 4.13: Volcano plot to display differentially expressed genes in the nucleus accumbens (NAc) between saline and fentanyl self-administering rats in the Oxy-On Board (sacrificed 5 minutes after last oxycodone session) study. Plot created by S. Sotcheff using R.

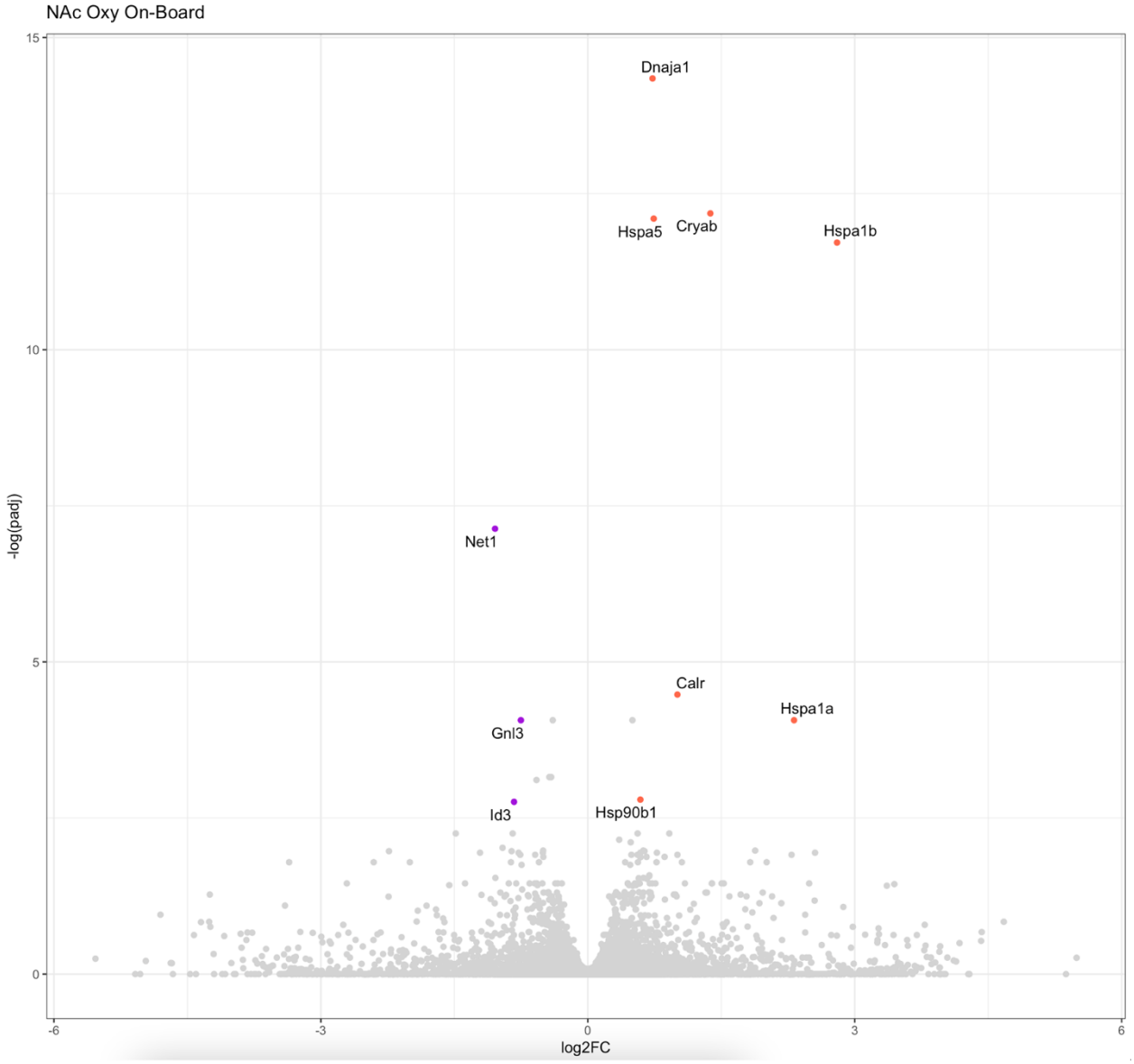


Figure 4.14: Volcano plot to display differentially expressed genes in the nucleus accumbens (NAc) between saline and fentanyl self-administering rats in the Oxy-Forced Abstinence (sacrificed 14 days after last oxycodone session) study. Plot created by S. Sotcheff using R.

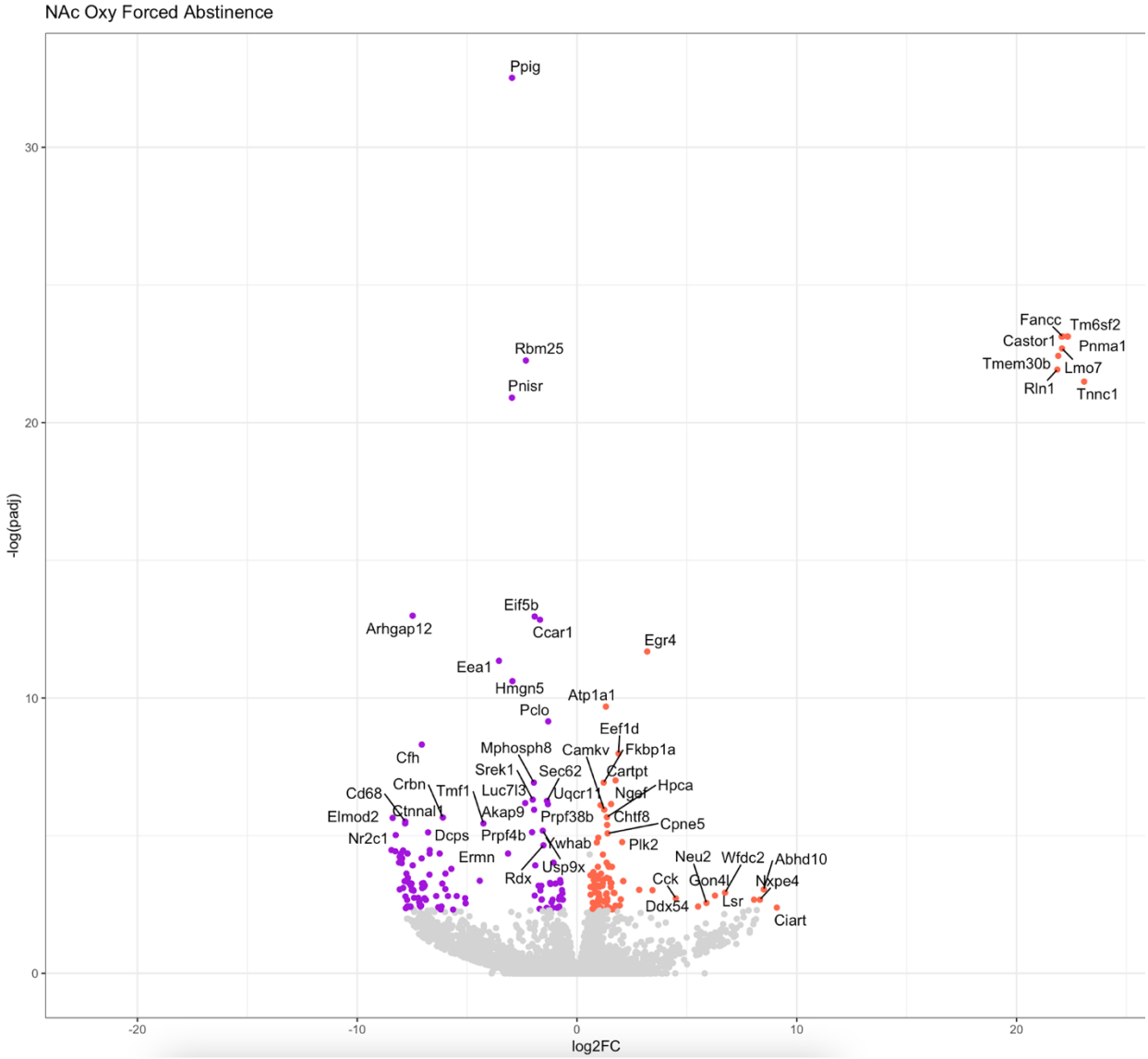


Figure 4.15: Volcano plot to display differentially expressed genes in the ventral tegmental area (VTA) between saline and fentanyl self-administering rats in the Oxy-On Board (sacrificed 5 minutes after last oxycodone session) study. Plot created by S. Sotcheff using R.

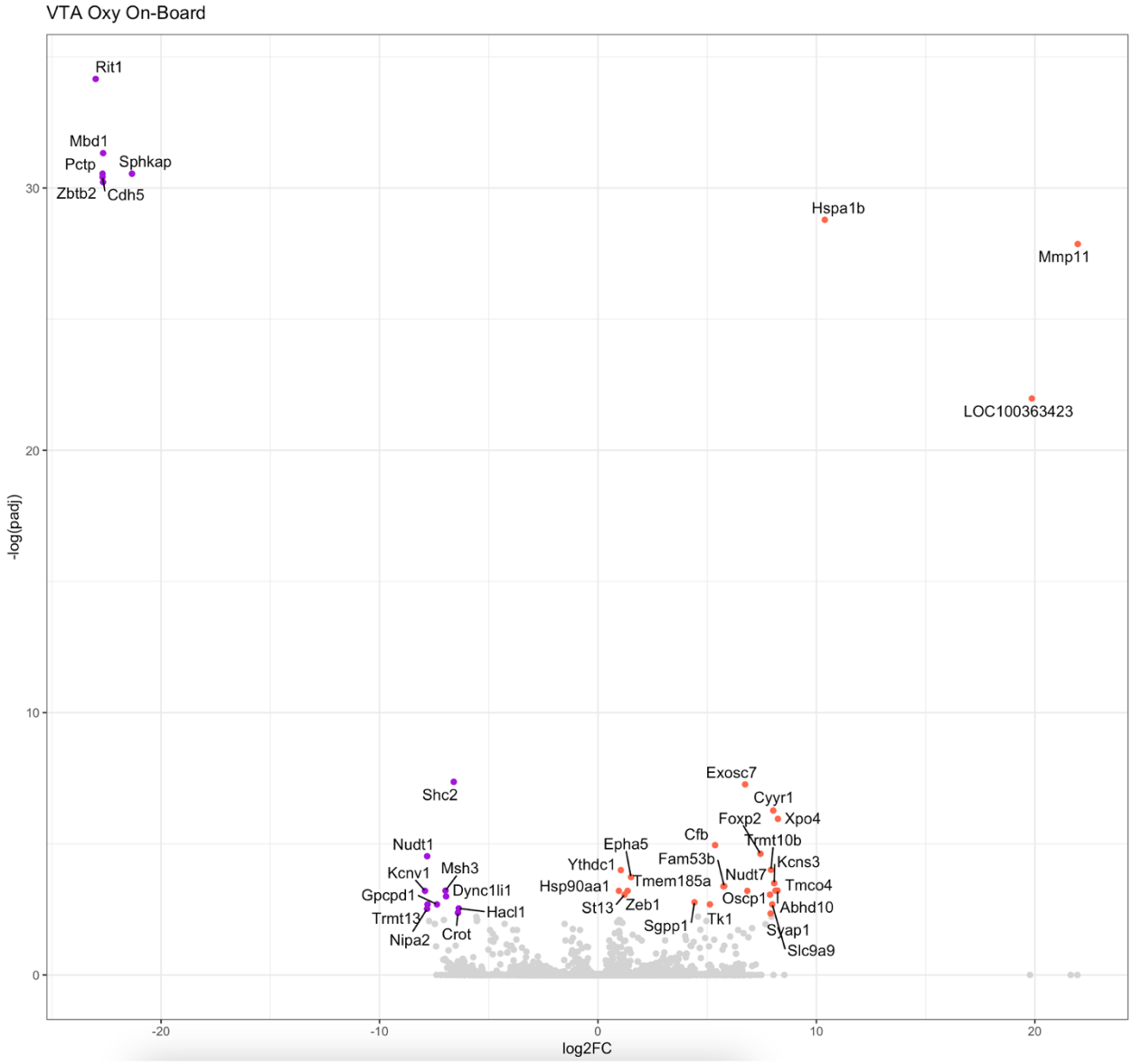
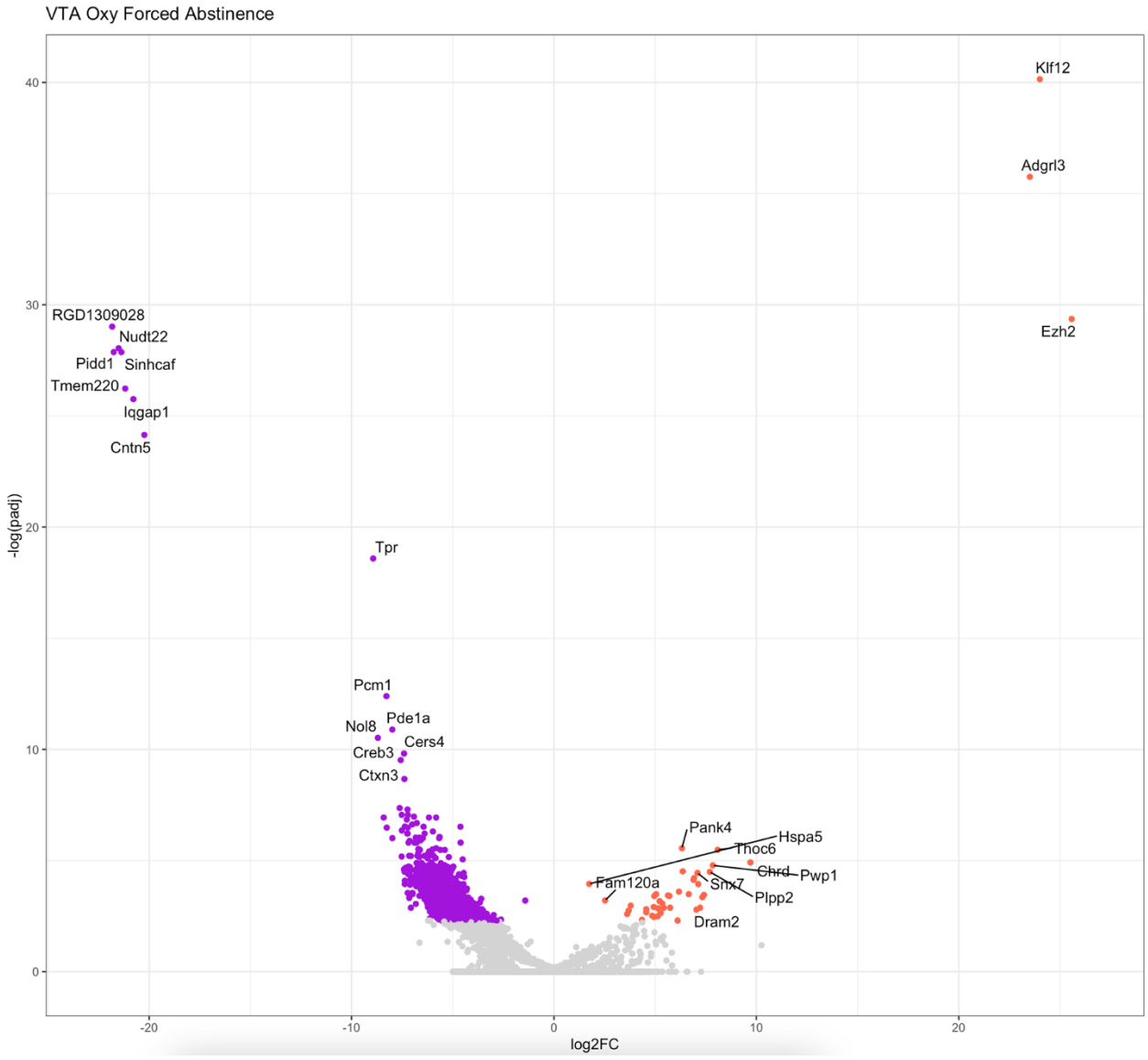


Figure 4.16: Volcano plot to display differentially expressed genes in the ventral tegmental area (VTA) between saline and fentanyl self-administering rats in the Oxy-Forced Abstinence (sacrificed 14 days after last oxycodone session) study. Plot created by S. Sotcheff using R.



down-regulated genes, and the opposite is true for the forced abstinence rats across both brain regions.

As there were a larger number of DEGs in the VTA studies as well as the NAc forced abstinence study, we input them into *Enrichr* (*167-169*) to determine enrichment of any biological pathways or processes (**Figures 4.17-4.22**). In the NAc forced abstinence study, we found up-regulation of genes in response to viral infection and dopamine receptors and down-regulation of clathrin-derived vesicle budding. In the VTA forced abstinence study, we found up-regulation of interferon signaling, calcium-activated potassium channels and the alternative complement pathway. In this data set we found down-regulation of various pathways including those involved in mRNA processing, leptin signaling, and caspase pathways of apoptosis. Lastly, in the VTA oxycodone on-board data we found up-regulation of complement, interleukin and interferon signaling and down-regulation of signal attenuation. Combined, these results support the findings in our snRNAseq data where the immune response is altered in the NAc upon fentanyl SA and 24 hours of abstinence.

Additionally, we wanted to compare DEGs across these datasets to determine if any were conserved in a given brain region after immediate euthanasia as well as forced abstinence (**Figure 4.23**). We found that there was overlap in 4 differentially expressed genes in the VTA that were up-regulated in the oxy on-board study but down-regulated in the forced abstinence study. These include Cfb, Tmem185a, Cyyr1, and Xpo4. These genes encode complement factor B, transmembrane protein 185a, cysteine and tyrosine rich protein 1, and exportin4 respectively. We also found that Gnl3 (noted above) is also down-regulated in the VTA after oxycodone forced abstinence. Hspa1b, a heat-shock protein, is up-regulated in the oxy on-board study in both the NAc and the VTA. We also found that there was overlap in 4 differentially expressed

Figure 4.17: *Enrichr* (*167-169*) scatterplot showing pathways that were enriched for up-regulated genes (orange) in rat nucleus accumbens (NAc) upon forced abstinence from oxycodone. Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Pathways enriched for up-regulated genes (p-value <0.05) are labelled as space allows.

Figure 4.18: *Enrichr* (*167-169*) scatterplot showing pathways that were enriched for down-regulated genes (purple) in rat nucleus accumbens (NAc) upon forced abstinence from oxycodone. Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Pathways enriched for down-regulated genes (p-value <0.05) are labelled as space allows.

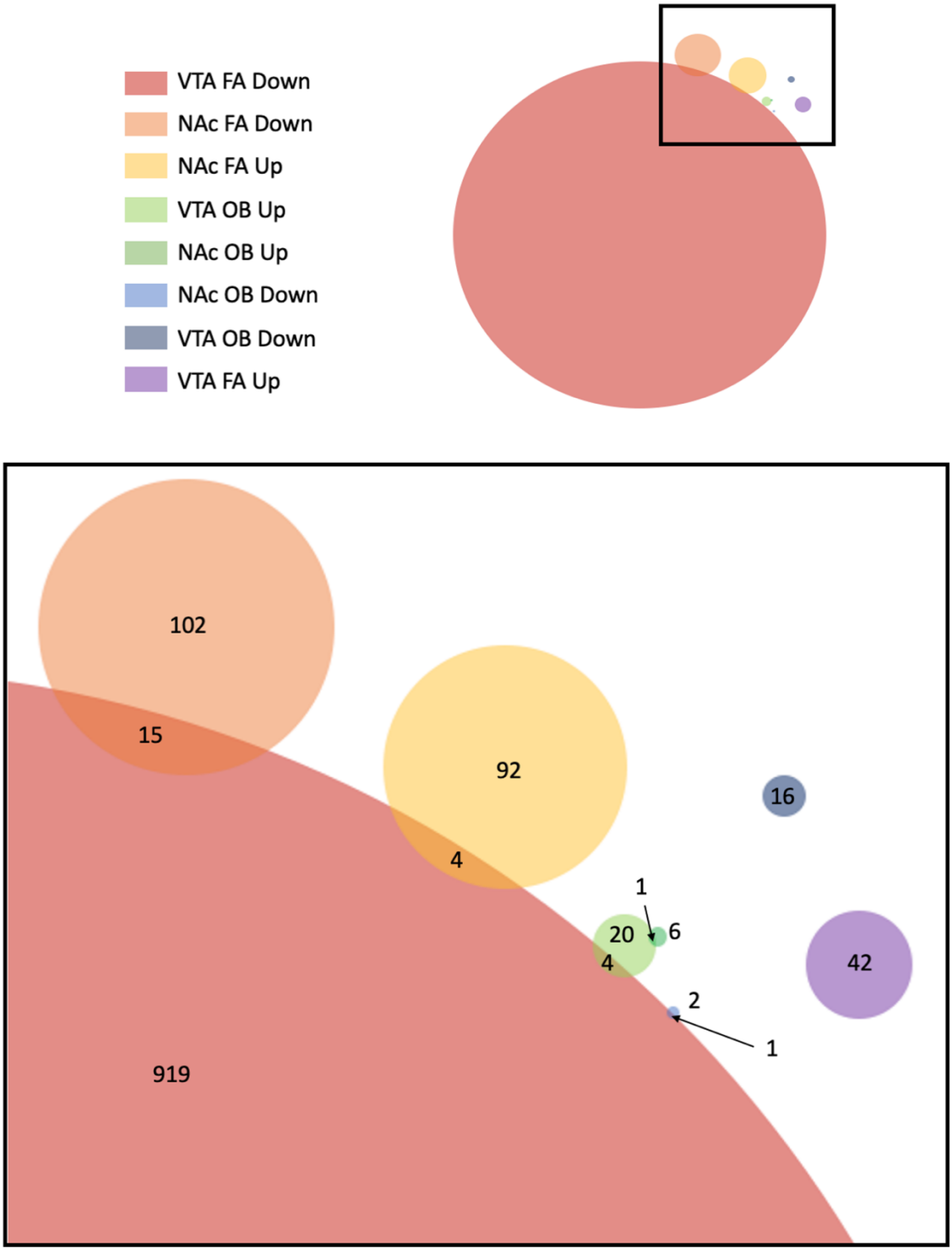
Figure 4.19: *Enrichr* (*167-169*) scatterplot showing pathways that were enriched for up-regulated genes (orange) in rat ventral tegmental area (VTA) with oxycodone on-board. Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Pathways enriched for up-regulated genes (p-value <0.05) are labelled as space allows.

Figure 4.20: *Enrichr* (*167-169*) scatterplot showing pathways that were enriched for down-regulated genes (purple) in rat ventral tegmental area (VTA) with oxycodone on-board. Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Pathways enriched for down-regulated genes (p-value <0.05) are labelled as space allows.

Figure 4.21: *Enrichr* (*167-169*) scatterplot showing pathways that were enriched for up-regulated genes (orange) in rat ventral tegmental area (VTA) upon forced abstinence from oxycodone. Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Pathways enriched for up-regulated genes (p-value <0.05) are labelled as space allows.

Figure 4.22: *Enrichr* (*167-169*) scatterplot showing pathways that were enriched for down-regulated genes (purple) in rat ventral tegmental area (VTA) upon forced abstinence from oxycodone. Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Pathways enriched for down-regulated genes (p-value <0.05) are labelled as space allows.

Figure 4.23: Complex venn diagram comparing genes that were differentially expressed in either the nucleus accumbens (NAc) or ventral tegmental area (VTA) in both studies (forced abstinence or oxy-on board). Figure created by S. Sotcheff using BioRender under license.



genes that were up-regulated in the NAc in the forced abstinence study but down-regulated in the forced abstinence study in the VTA. These genes were Ddx54, Exd2, Kctd1, and Camkv which encode for dead-box helicase 54, exonuclease 3'-5' domain containing 2, potassium channel tetramerization domain containing 1, and Cam kinase like vesicle associated protein. There were 15 genes that were down-regulated in both brain regions in the oxy forced abstinence study. These genes were input for *Enrichr* (**Figure 5.24**) which suggested that force abstinence from oxycodone results in the down-regulation of some facets of the immune response, adrenergic pathway, insulin signaling, and interestingly genes involved in Alzheimer's disease in both the VTA and the NAc.

DISCUSSION

As noted above, ACE2 and at least one of the protease/peptidases listed above is necessary for entry of SARS-CoV-2 into host cells and in order for a cell to be permissive to ZIKV infection it must express a receptor that the virus can bind to. Two receptors that Zika can bind to in the brain are toll-like receptor 3 (TLR3) and receptor tyrosine kinase AXL. As previous studies have indicated that SARS-CoV-2 and various flaviviruses have an impact on the CNS and that opioid use alters immune regulation in the brain, we wanted to determine if opioid use may exacerbate the effects of these viral infections. Opioids and other drugs of abuse studies in addiction research act in the mesolimbic pathway (involved in learning and reward seeking) and therefore we investigated DGE in the NAc and VTA. In response to oxycodone there were very few changes when rats were sacrificed five minutes after their last session. However, when animals were forced to be abstinent from the drug for an extended period there were much larger changes in gene expression patterns in both brain regions. Ultimately, the withdrawal they experienced seemed to hinder the immune response overall, but result in increased expression of

Figure 4.24: *Enrichr* (*167-169*) scatterplot showing pathways that were enriched for down-regulated genes (purple) in both rat nucleus accumbens (NAc) and ventral tegmental area (VTA) upon forced abstinence from oxycodone. Each point is a pathway listed in the BioPlanet 2019 database. Points that are near to each other are most similar or share genes. Pathways enriched for down-regulated genes (p-value <0.05) are labelled as space allows.

pro-inflammatory cytokines and genes in the type I interferon pathway. Ultimately this was consistent with the snRNAseq data we generated for rats self-administering fentanyl with one day of abstinence prior to sacking.

Importantly, we also found expression of ACE2 in mural cells and expression of the proteinase/peptidases involved in SARS-CoV-2 entry in mural cells, microglia, and oligodendrocytes. We also found expression of a receptor necessary for ZIKV infection in mural cells and astrocytes (AXL) and a marker for inflammation that is typically induced by ZIKV infection expressed in microglia (TLR3). In mural cells we also found increased expression of antiviral and type I interferon pathways in response to acute fentanyl withdrawal, but in microglia we found increased expression of IL-6 and STING, indicating release of proinflammatory cytokines.

Interestingly, hyper-inflammatory responses have been shown to serve as an indicator for severe disease in SARS-CoV-2 infected individuals. Specifically, elevated IL-6 has been proposed as a predictor for non-survivors of COVID-19 (*216, 217*). Additionally, flaviviruses have been shown to increase inflammasome production and stimulate inflammatory cell death, releasing many pro-inflammatory cytokines (*213*). Our results suggest that within the NAc only mural cells have the potential to be directly infected with SARS-CoV-2 but multiple cell types could be infected by ZIKV. There is also an increase in antiviral and type I interferon expression in mural cells coupled with the release of pro-inflammatory cytokines from nearby microglia, which suggests that regular fentanyl use could potentially exacerbate the effects of SARS-CoV-2 of ZIKV infection and increase severity of COVID-19 or flavivirus induced neurological disease.

Chapter 5 Collaborative projects investigating recombination in SARS-CoV-2

In this chapter some sections are directly lifted from one of my published works (citations below). As first author on the Virus Recombination Mapper manuscript, I conducted all computational analyses, wrote those sections, and aided in revision of the entire manuscript. Libraries were previously prepared and sequenced by or for collaborators and prepared all of the figures. For the additional manuscripts I prepared libraries and conducted data analysis. In the final paper I prepared a supplementary figure. These citations are also included in the references section.

Sotcheff, S.; Zhou, Y.; Sun, Y.; Johnson, J.E.; Torbett, B.E.; Routh, A. *ViReMa*: A Virus Recombination Mapper of Next-Generation Sequencing data characterizes diverse recombinant viral nucleic acids. *BioRxiv*. 2022 Mar. https://doi.org/10.1101/2022.03.12.484090

Jaworski, E.; Langsjoen, R.; Mitchell, B.; Barbara, J.; Newman, P.; Plante, J.; Plante, K.; Miller, A.; Zhou, Y.; Swetnam, D.; **Sotcheff, S**.; Morris, V.; Saada, N.; Muchado, R.; McConnell, A.; Widen, S.; Thompson, J.; Dong, J.; Ping, R.; Pyles, R.; Ksaizek, T.; Menachery, V.; Weaver, S.; Routh, A. Tiled-ClickSeq for targeted sequencing of complete coronavirus genomes with simultaneous capture of RNA recombination and minority variants. *eLife*. 2021 Sept. http://dx.doi.org/10.7554/eLife.68479

Vu, M.; Lokugamage, K.; Plante, J.; Scharton, D.; Johnson, B.; **Sotcheff, S**.; Swetnam, D.; Schindewolf, C.; Alvarado, R.; Crocquet-Valdes, P.; Debbink, K.; Weaver, S.; Walker, D.; Routh, A.; Plante, K.; Menachery, V. QTQTN motif upstream of the furin cleavage site plays key role in

SARS-CoV2 infection and pathogenesis. *BioRxiv*. 2021 Dec. https://doi.org/10.1101/2021.12.15.472450

Coronavirus disease 2019 (COVID-19) has emerged as a historical pandemic that may have an impact on our lives for years if not generations to come. To date there have been over 300 million cases and 5 million deaths world-wide since the causative agent, severe acute respiratory syndrome (SARS) coronavirus 2 (SARS-CoV-2), was identified. Of that, the United states has documented roughly 80 M cases and 980K deaths as of April 7th, 2022 (*218*). Additionally, this is the third coronavirus to spread from animal reservoirs and into humans in the last two decades (including SARS-CoV in 2003 (*219*) and Middle East respiratory syndrome [MERS] in 2012 (*220, 221*)). SARS-CoV-2 has had a more significant impact on the world at large. Although, chronic co-morbidities such as pulmonary disease, obesity, and diabetes are thought to be key contributors in severe COVID-19 cases (*222, 223*) we have aimed to identify key changes between this novel coronavirus and its predecessors that have resulted in this effect.

Recombination of viral genomes, both homologous and non-homologous, is essential for viral evolution and adaptation. Homologous recombination can occur within a single cell among a number of genomes including SNVs or even produce viral chimeras if two related viruses happen to co-infect a cell (*224*). Non-homologous recombination results in simple insertions or deletions or even more complex rearrangements such as large deletions, duplications, or insertions of host genetic material into the viral genome. This can produce structural variants as well as defective viral genomes (genomes that replicate when in a cell with functional "helper" virus but do not code for all necessary viral proteins themselves) (*225, 226*). These changes in viral genomes can result in the emergence of outbreak strains. One example includes the large

structural variants in ORF 8 upon adaptation of SARS-CoV-1 to human transmission (*227*). Small deletions near the furin cleavage site of SARS-CoV-2 have been found upon serial passaging that appear to attenuate the virus (*228, 229*) – suggesting that the presence of these sequences (not found in MERS or SARS-1) contributes to pathogenicity.

In collaboration with Dr. Vineet Menachery, and other collaborators at UTMB, we have set out to identify recombination events in SARS-CoV-2 from patient samples and understand how some of these events may alter pathogenicity and/or infectivity. We used the Tiled-ClickSeq protocol to generate libraries of SARS-CoV-2 from patient samples as well as ARTIC libraries from these same samples to serve as a control for the method. I have also aided in transcriptomics projects investigating differences in gene expression for various mutants (deletions near the furin cleavage site or mutations in nsp16). I will describe these studies in more detail below. We found various recombination events in the patient samples ranging from large deletions to micro insertion/deletions (indels) in SARS-CoV-2. We investigated how a small deletion near the furin cleavage site may impact disease outcomes by conducting transcriptomic studies in lung tissue from infected Syrian golden hamsters. Additionally, Michelle Vu in Dr. Menachery's lab monitored symptoms of disease and viral load in the infected hamsters.

VIRUS RECOMBINATION EVENTS

Viral genomes are constantly evolving. Though many consider this a bi-product of an error-prone viral polymerases, there is evidence to show that recombination in viral genomes aids in both the pathogenesis and evolution of these pathogens (*63, 224, 230*). In fact, recombination in viral genomes has been shown to generate outbreak strains of coronaviruses (*227, 230*). Our lab has produced two analysis pipelines to investigate and identify recombination in viral genomes. The first, Virus Recombination Mapper or *ViReMa* (*73*) looks at

recombination events on their own in short read data and identifies events from single nucleotide to large deletions, duplications, copy-backs, and host-virus recombination. The second, Co-variation Mapper or *CoVaMa* (*74, 75*) can be used to determine if specific recombination or single nucleotide poly-morphisms exist in a genome relative to one another (i.e. are they mutually inclusive or exclusive for example – see Appendix G). This section highlights my work in validating the updates to the *ViReMa* pipeline.

Virus Recombination Mapper (*ViReMa*)

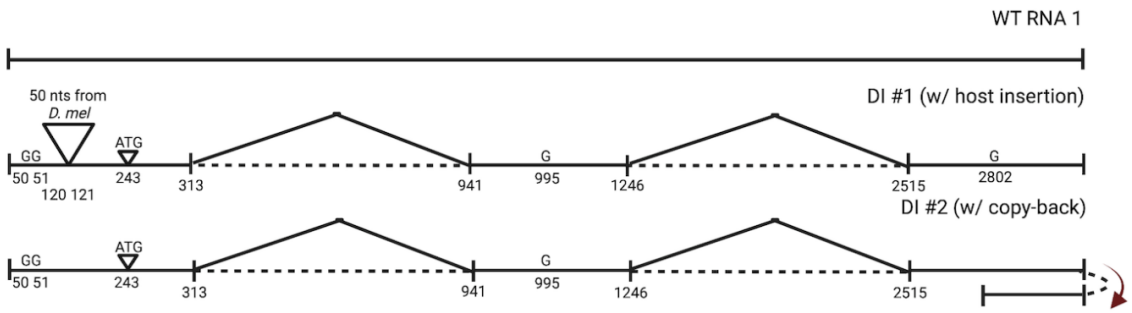
1. FHV Simulated and Previously Available Data

We generated three artificial RNA templates based upon recombination events previously seen in Flock House virus as well as hypothetical events designed to challenge and test our platform. These templates are illustrated in **Figure 5.1**, and contain a deletion event, simple insertions, duplications, points mutations, and multiple point mutations as well as a copy-back type event. We also included a D-RNA like sequence with an insertion of a 50 nt fragment of host mRNA (chr3L:23086340-23086389 from *dm6*). Using these input reference sequences, we generated simulated reads using the 'ART' tools for read simulation generated as part of the 1000 genomes project (*231*). With ART, we generated a dataset containing approximately 37'000 X coverage over each of the simulated DI-RNAs (74,000X total for both D-RNAs) and 300'000X over the wild-type virus. An error profile reflecting the HiSeq 2500 was imposed, yielding 10,935,310 reads. These simulated reads are packaged with the *ViReMa* software for user-testing/verification.

We ran the *ViReMa* pipelines using standard parameters: (--X 3 --Defuzz 0 --Host_Seed 25 - BED --MicroInDel_Length 5) mapping to both a padded FHV genome and dm6 genome. The resulting BED files produces are shown in **Figure 5.2.** As can be seen, each of the simulated

Figure 5.1: *ViReMa* analysis of simulated and previously reported Flock House virus (FHV) data. A) Defective RNA1 (DI) genomes we generated simulated reads for using ART Illumina. On top we have WT FHV RNA1. D-RNA1 #1 includes an insertion from *D. melanogaster* before the start codon as well as two deletions. D-RNA1 #2 also includes the same deletion events but also includes a copy-back event instead of the insertion. B) *ViReMa* results of our simulated data found in the Virus_Recombination_Results.bed file. Output indicates the reference (in this case FHV has two references, one for each strand of its bi-partite genome), where the first sequence aligned stops, where the alignment picks back up, the type of recombination event, how many times it occurred, which genome sense the reads map to, and how many total reads at the "stop" or 5' of the recombination event and the "start" or 3' of the recombination event. C) *ViReMa* results for our simulated data found in the MicroRecombinations.bed file. D) *ViReMa* results for our simulated data found in the Virus-to-Host-Recombination output file. E) *ViReMa* results for our simulated data found in the Virus_Fusions.BEDPE file. Similar information is depicted as in the Virus-to-Host-Recombination output. Here we show the detection of the copy-backs we included in our simulated dataset. Figure created by S. Sotcheff using BioRender under license, used here with permission from Sotcheff et al. *BioRxiv* 2022 (*232*).

A



B

Reference	Stop	Start	Event Type	Count	Sense	5' Reads	3' Reads
FHV RNA1	1245	2514	Deletion	17869	+	178827	160111
FHV RNA1	313	941	Deletion	17857	+	178247	179034
FHV RNA1	1243	2512	Deletion	55	+	180052	150621
FHV RNA1	311	939	Deletion	46	+	179551	151039
FHV RNA1	2512	1243	Deletion	17788	-	150663	179631
FHV RNA1	940	312	Deletion	17174	-	176001	178738
FHV RNA1	2516	1247	Deletion	53	-	179842	151208
FHV RNA1	942	314	Deletion	47	-	177183	150706

C

Reference	Stop	Start	Event Type	Count	Sense	5' Reads	3' Reads
FHV RNA1	243	244	Ins:ATG	16292	+	177962	178615
FHV RNA1	243	244	Ins:TTG	45	+	177962	178615
FHV RNA1	243	244	Ins:AAG	43	+	177962	178615
FHV RNA1	243	244	Ins:ACG	43	+	177962	178615
FHV RNA1	243	244	Ins:AGG	41	+	177962	178615
FHV RNA1	243	244	Ins:ATA	40	+	177962	178615
FHV RNA1	243	244	Ins:ATT	38	+	177962	178615
FHV RNA1	243	244	Ins:GTG	36	+	177962	178615
FHV RNA1	2293	2295	Deletion	3	+	149948	149934
FHV RNA1	1538	1540	Deletion	2	+	150632	150533

D

Reference 1	Stop 1	...	Reference 2	Start	...	Event Type	Count	Sense 1	Sense 2
chr3L	23093239	23093240	FHV RNA1	120	121	Virus-Host-Fusion	9062	-	-
chr3L	23093238	23093239	FHV RNA1	119	120	Virus-Host-Fusion	27	-	-
chr3L	23093240	23093241	FHV RNA1	121	122	Virus-Host-Fusion	18	-	-
chr3L	23093234	23093235	FHV RNA1	115	116	Virus-Host-Fusion	17	-	-
FHV RNA1	120	121	chr3L	23093289	23093290	Virus-Host-Fusion	8820	-	-
FHV RNA1	118	119	chr3L	23093287	23093288	Virus-Host-Fusion	33	-	-
FHV RNA1	121	122	chr3L	23093290	23093291	Virus-Host-Fusion	20	-	-
chr3L	15255214	15255215	FHV RNA1	1631	1632	Virus-Host-Fusion	1	+	-
chr3L	22512568	22512569	FHV RNA1	26	27	Virus-Host-Fusion	3	-	+
FHV RNA1	38	39	chr3L	16741900	16741901	Virus-Host-Fusion	1	-	+
chr3L	23093288	23093289	FHV RNA1	121	122	Virus-Host-Fusion	8909	+	+
chr3L	23093287	23093288	FHV RNA1	120	121	Virus-Host-Fusion	15	+	+
chr3L	23093289	23093290	FHV RNA1	122	123	Virus-Host-Fusion	14	+	+
FHV RNA1	119	120	chr3L	23093240	23093241	Virus-Host-Fusion	8628	+	+
FHV RNA1	120	121	chr3L	23093241	23093242	Virus-Host-Fusion	24	+	+
FHV RNA1	118	119	chr3L	23093239	23093240	Virus-Host-Fusion	16	+	+
FHV RNA1	503	504	chr3L	22311648	22311649	Virus-Host-Fusion	1	+	+

E

Genome 1	Stop	Start	Genome 2	Stop	Start	Event Type	Count	Sense 1	Sense 2
FHV RNA1	3106	3107	FHV RNA1	3105	3106	Copy/Snap-Back	5823	+	-
FHV RNA1	3104	3105	FHV RNA1	3109	3110	Copy/Snap-Back	17	+	-
FHV RNA1	3108	3109	FHV RNA1	3105	3106	Copy/Snap-Back	14	+	-

recombination events are successfully detected with no erroneous mappings. The correct deletion recombination events at 313^941and 1246^2515 are found, as is a small ATG insertion between nts 243 and 244. The correct host insertion (virus-host fusion type) event is detected, as reported using the BEDPE output as described in the methods section.

As we know the identity of the simulated reference genomes, we can determine how many reads should have been able to map across the region containing the hypothetical host insertion event with the appropriate sized mapping segments either side (25nt seed of the host, 20 nt seed for the virus). Using these parameters therefore, we calculated that of the 10,935,310 reads there should have been, which would be in both positive and negative sense, the host insertion event should be present in 9,250 reads in each sense. In the negative sense the event was identified 8,941 times (the average of both ends of the read: 9,062 and 8,820), however in the positive sense the event was detected 8,769 times (the average of 8,909 and 8628) **(Fig. 5.1D)**. Therefore the accuracy of detecting this event was 96.7% in the negative sense and 94.8% in the positive sense.

As seen in **Figure 5.1D**, *ViReMa* found 17,857 and 17,869 occurrences of the above mentioned deletion recombination events (518^1146 and 1450^2719) in the positive sense, yielding 92.3% and 84.5% mapping efficiencies respectively. The imperfect sensitivity is simply due to the presence of single nucleotide mismatches that are found near recombination junctions. While these reads are still mapped to the reference genome and can be found in the output SAM file, *ViReMa* does not annotate these reads as containing recombination events if mismatches are found near the putative recombination junction at a distance defined by the optional –X parameter (set to 3 above). This value can be adjusted to increase sensitivity, but at the cost of introducing false positive events (*233*). In addition, we used *ViReMa* on a set of three replicates to determine the precision of the pipeline in identifying particular recombination events. We investigated two known recombination

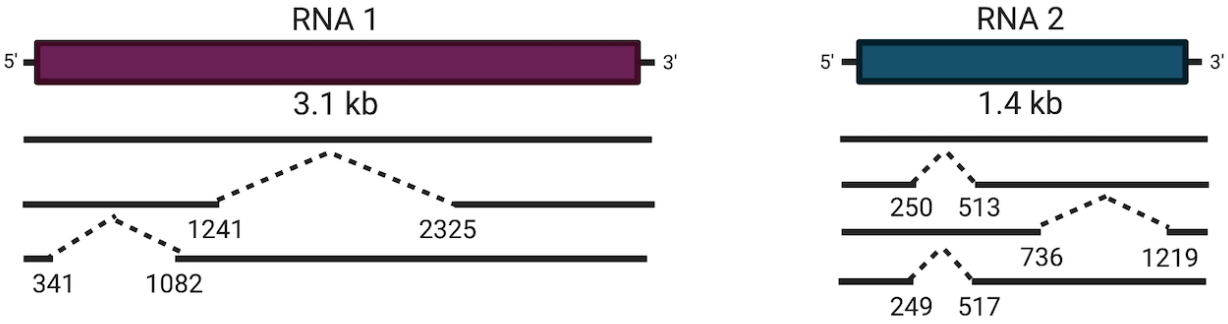
events in RNA1 and three in RNA2 from late Passage FHV. These events are illustrated in **Figure 5.2**. The percentage of reads that include each event per replicate are provided in the table in **Figure 5.2**, as well as the average and standard deviation for each recombination event. The standard deviation ranged from 0.31 to 2.35% indicating that *ViReMa*'s ability to identify recombination events is precise.

2. HIV back splicing

In a previous study, we reported the analysis of covariation of amino acid and nucleotide variants within a large cohort of HIV infected patients as part of the ARMY Longitudinal study. We used the *ViReMa* pipeline to detect insertions, deletion, or recombination events in the patient data. To begin, processed reads were mapped to the reference HIV genome (2B) using *ViReMa*. We invoked the -back-splicing option to report short insertions and duplications in the HIV genome. Strikingly, this revealed a large number of recombination events occurring at close to the proteolysis sites between capsid and matrix protein, as well as in the near the PTAP region of the p6 protein. These events are largely characterized by in-frame duplications ranging from 3 to 24 nucleotides in length.

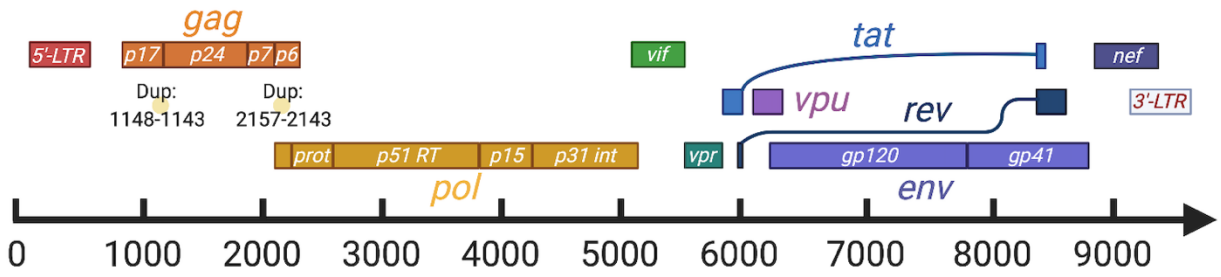
The most commonly seen events (2 and 3 illustrated in **Figure 4.3**) cluster into three regions and are most frequently: 1) 'AQQA' duplications 13 amino acids upstream of the p17/p24 proteolysis site; 2) 'QSRPE' duplications two amino acids downstream of the p1/p6 proteolysis site; and 3) 'PTAP' duplications 10 amino acids downstream of the p1/p6 proteolysis site. There is some variation in the exact nature of these duplications, varying in length from 3 to 24 nucleotides, and sometime containing inexact duplications that introduce variant amino acids at the duplication site. Nonetheless, the same or closely similar events were seen in multiple different patient samples, indicating that these duplications were a common response to a common

Figure 5.2: Depiction of five previously annotated FHV recombination events across both strands (top). Table indicating the precision of detecting those events with *ViReMa* across three replicates of late passage FHV. Figure created by S. Sotcheff using BioRender under license, used here with permission from Sotcheff et al. *BioRxiv* 2022 (*232*).



Strand	Stop	Start	Rep1 %	Rep2 %	Rep3 %	Average	St. Dev.
RNA1	1241	2325	8.83	8.9	8.33	8.69	0.31
RNA1	342	1083	8.37	12.55	11.22	10.71	2.14
RNA2	250	513	25.88	30.55	27.7	28.04	2.35
RNA2	249	517	11.84	12.61	11.47	11.97	0.58
RNA2	736	1219	9.2	5.94	9.66	8.27	2.03

Figure 5.3: Duplications in *gag* in HIV patient data. Map of the HIV genome with the identified duplication events noted at nucleotide positions 1148-1143 and 2157-2143. Figure created by S. Sotcheff using BioRender under license, used here with permission from Sotcheff et al. *BioRxiv* 2022 (*232*).



evolutionary selection pressure. Importantly, many of these duplication have been previously observed, and have been characterized as a response to antiviral drug treatment.

In **Figures 5.3 and 5.4** we illustrate where these duplications are occurring in the context of the HIV genome and what these duplications may look like at the nucleotide level for two examples: 2157-2143 and 1148-1143 that were found in a patient's sequencing data over the course of ~5 years. Note that in the third time point (Feb 2001) the 1148-1143 is replaced with 1150-1145, shown in the table in **Figure 5.5.** This is likely the same recombination event and therefore is used in the calculations for percentage of the 1148-1143 event in **Figure 5.6**. The gray dotted line in **Figure 4.6** illustrates when the patient was switched from indinavir (IDV) to another anti-retroviral – nelfinavir (NFV). In addition to showing changes in viral duplications over time, we used this data set to investigate how changing the error density parameter, noted above, may affect the permissibility of *ViReMa* and the percentage of reads that are found to contain these duplication events over time as well as one additional duplication that was found at a lower percentage in the Feb 2001 data (**Figure 5.7**). The largest deviations in duplication events per reads at these particular nucleotide positions was found when using the least permissive parameters of 3,100 and 1,100 (number of mismatches allowed per x nucleotides), which reduced the number of mapped reads and proportion of these reads identified as containing the above mentioned duplications as expected. The more permissive settings (1,10 and 1,15) did not vary much from the results seen in the Default, 2,25, or 1,20 runs in terms of proportion of reads containing the duplications, but did increase the number of mapped reads by ~3%.

This example demonstrates the ability of *ViReMa* to faithfully identify duplication events and how altering the permissibility of *ViReMa* by changing the error density parameter may affect the results. Note that it is important to keep the error density parameter fairly permissive to be able to

Figure 5.4: Depiction of the duplication events (1148^1143 and 2157^2143) at the sequence level. Figure created by S. Sotcheff using BioRender under license, used here with permission from Sotcheff et al. *BioRxiv* 2022 (*232*).

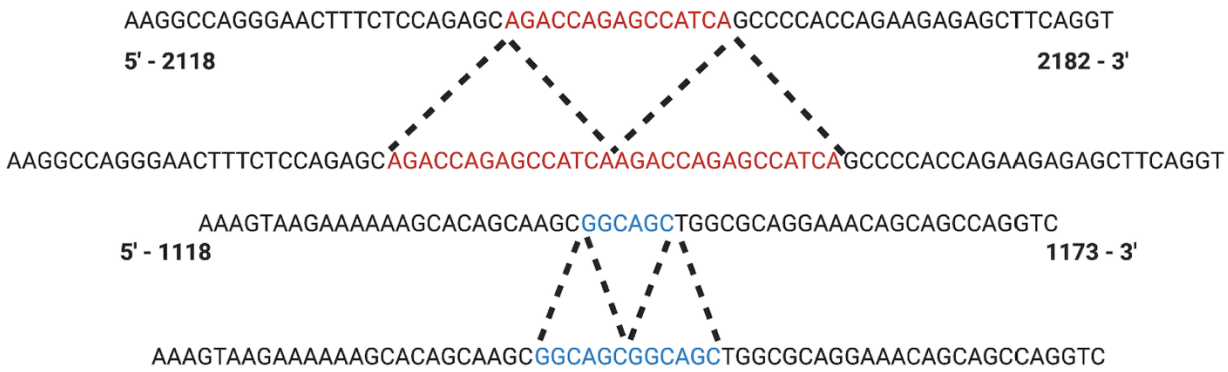


Figure 5.5: *ViReMa (73)* output in the Virus_Recombination_Results.bed with all time points compiled and focused on the primary duplication events, used here with permission from Sotcheff et al. *BioRxiv* 2022 (*232*).

Month-Year	5' Position	3' Position	Event Type	Count	Read Sense	Read Count 5'	Read Count 3'
Feb-97	2157	2143	Duplication	4511	+	24442	19408
	1148	1143	Duplication	5310	+	30052	30146
Aug-00	2157	2143	Duplication	634	+	23562	19515
	1148	1143	Duplication	1475	+	16678	17877
Feb-01	2157	2143	Duplication	1000	+	24519	16494
	1150	1145	Duplication	927	+	14149	15192
Oct-01	2157	2143	Duplication	1478	+	27527	18505
	1148	1143	Duplication	4966	+	22217	23104
Jun-02	2157	2143	Duplication	1127	+	17914	12967
	1148	1143	Duplication	2851	+	12385	12003

Figure 5.6: Plot of the percent of mapped reads at these nucleotide positions that identify the noted duplication events over time. Note that the patient's anti-retroviral treatment was changed after the third timepoint from indinavir (IDV) to nelfinavir (NFV). Figure created by S. Sotcheff using R and BioRender under license, used here with permission from Sotcheff et al. *BioRxiv* 2022 (*232*). Data also shown in Wang et al. NAR 2022 (*75*).

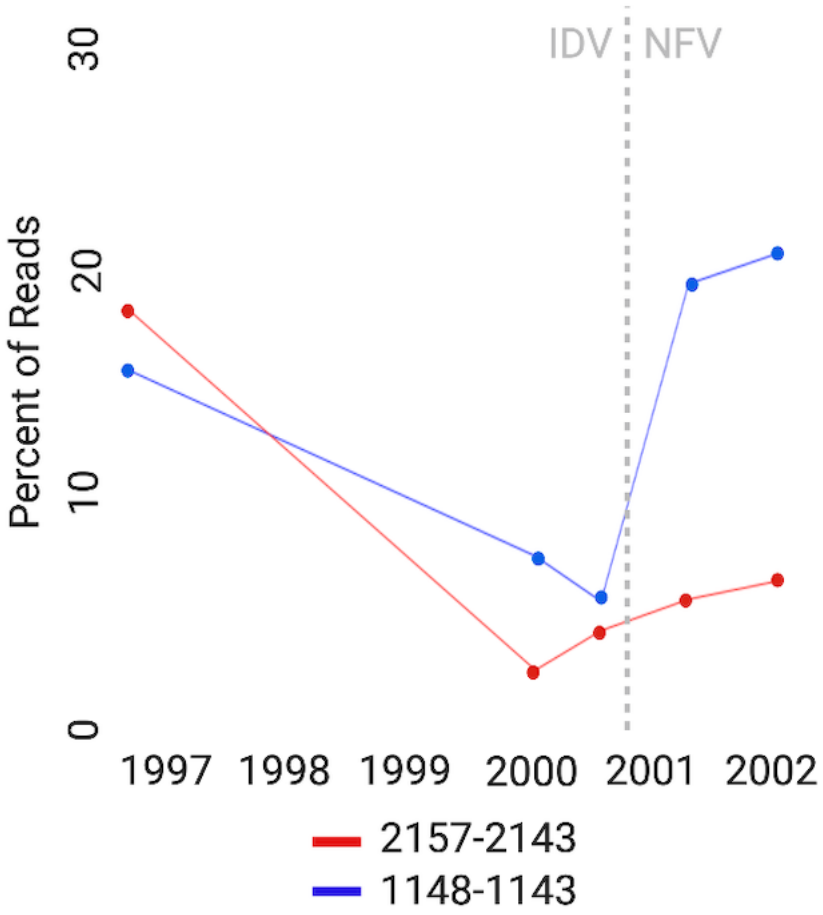


Figure 5.7: Tables depicting how the number of duplication events detected is dependent on the error density x,y where x is the number of allowed mismatches in y nucleotides. We included the two above mentioned events and an additional duplication event at position 1174, used here with permission from Sotcheff et al. *BioRxiv* 2022 (*232*).

Error Density	Reads at 1148	Reads at 1143	Duplication Events	Dup. Percentage
Default	14301	15005	1305	8.91
2,25	15849	16577	1474	11.70
3,100	12165	13032	788	6.25
1,100	11223	12085	487	4.18
1,20	15904	16719	1482	9.09
1,15	16279	16944	1612	9.70
1,10	16300	16978	1614	9.70

Error Density	Reads at 2157	Reads at 2143	Duplication Events	Dup. Percentage
Default	21510	17690	321	1.64
2,25	21580	17735	354	1.80
3,100	21217	17625	100	0.51
1,100	20868	17320	97	0.51
1,20	21599	17752	355	1.80
1,15	21605	17760	360	1.83
1,10	21617	17766	361	1.83

Error Density	Reads at 1174	Reads at 1160	Duplication Events	Dup. Percentage
Default	17464	13615	259	1.67
2,25	17566	13610	274	1.76
3,100	17144	13185	3	0.02
1,100	16609	10091	3	0.02
1,20	17565	13613	274	1.76
1,15	17574	13616	276	1.77
1,10	17575	13619	276	1.77

detect these recombination events.

3. Copy-Back DVGs - Sendai

Copy-back or snap-back RNAs are types of defective viral genomes (DVGs) commonly found in negative sense RNA viruses including measles (*234, 235*), mumps (*236*), Nipah (*237*) and Sendai (SeV) virus (*238, 239*). They are synthesized during synthesis of the anti-genome (positive-sense) if the viral polymerase stalls and detaches from the template strand and begins copying the nascent strand creating a hairpin loop with complementary sequences in the strand at both ends. These were identified in SeV, or murine parainfluenza virus 1 - a paramyxovirus closely related to human parainfluenza viruses 1 and 3, as early as 1983, where sequencing found defective interfering genomes that consisted of the 5' end of the genomic (negative sense) strand and the 3' end of the positive sense or coding strand (*238, 239*). Since then, these recombination events have been noted by various labs and have been shown to promote persistent paramyxovirus infections by preventing apoptosis in DVG enriched cells versus highly infected cells with full-length virus (*235, 240-242*). In addition, copy-backs in SeV have also been shown to increase the heterogeneity of virus particles produced by infected cells. Copy-back RNAs can occur at multiple different loci in the 5' ends of the -ve sense RNA genome. Copy-back species, and other recombination events, can also to appear in sequencing data as an artifact of cDNA library prep synthesis (*243, 244*). As part of a previous study, genomic SeV RNA is purified from cells in culture and used to synthesize cDNA libraries using an Illumina TruSeq Stranded Total RNA LT kit with Ribo-Zero Gold. These libraries were sequenced on a Illumina NextSeq 550, obtaining 21-53M 75 bp single-end reads for each of 6 samples with high DVGs (*241*). We mapped these reads using ViReMa to both the SeV genome (AB855654.1) and the human genome to confirm the presence of these known copyback RNAs previously reported (**Figure 5.8**).

Figure 5.8: Detection of copy-back RNAs in Sendai virus (SeV) using *ViReMa*. Table showing the number of copy-back events in the data set. The 'Percent' column is calculated from the count of reads with the recombination event (from Virus_Fusions.bedpe) compared to the total number of reads at those nucleotide positions from the 'cuttingsites' bedfiles, used here with permission from Sotcheff et al. *BioRxiv* 2022 (*232*).

Genome 1	Stop	...	Genome 2	Start	...	Event Type	Count	Orientation 1	Orientation 2	Genome 1 Reads	Genome 2 Reads	Percent
SeV	15290	15291	SeV	14933	14934	Copy/Snap-Back	50574	-	+	77288	51536	78.52
SeV	14931	14932	SeV	15292	15293	Copy/Snap-Back	3611	-	+	46943	9841	12.72

These copy-back DVGs are made by template switching during production of the negative sense genome as shown in **Figure 5.9**. These recombination events appear to primarily occur in sites with CTT motifs (**Fig. 5.10 and 5.11**). The data in the table in **Figure 5.8** derived from the *ViReMa* output file '*Virus_Fusions.BEDPE'*. As described in the 'Methods' section, this is a format used to report eukaryotic genomic fusion events and is a convenient standardized format to use here. The 'Reads' columns describe the number of reads for a particular nucleotide position of SeV and can be populated using the data in the 'cuttingsites' bedfiles and the 'Percent' column was populated by calculating the number of recombination event counts over the average of the 3' and 5' read counts. Interestingly, this table indicates we have replication of the original DVG, which we propose a model for in **Figure 5.12**.This example demonstrates the ability of *ViReMa* to faithfully identify copy-back recombination events that have been previously identified by other groups. In addition, this provides insight to other output files of *ViReMa* that can provide information on coverage as opposed to recombination events.

4. Host to virus recombination - STIV

Sulfolobus turreted icosahedral virus (STIV) infects the thermophile *Sulfolobus solfataricus*. These archaea, and the viruses that infect them, are found in hot springs. To avoid degradation at high temperatures and low pH, STIV contains an inner membrane within its icosahedral capsid. Virus particles are 75 MDa and enclose a 17.6 kb double-stranded circular DNA genome (*245*).
To characterize rates of recombination in a DNA sourced viruses, we obtained genomic DNA from samples of STIV purified in culture (from Mark Young at Montana State University). Genomic viral DNA was prepared for NGS using the ClickSeq method and sequenced on an Illumina HiSeq 1000 obtained ~38M single-end 150 bp reads in total across four samples. We used ViReMa to map these reads to the STIV genome (NC_005892.1) and the host genome (Sulfolobus:

Figure 5.9: Map of the negative sense (packaged) SeV genome with a box around the 5' end where the copy-backs occur. This box is blown up to show how these copy-back DVGs are made. Figure created by S. Sotcheff using BioRender under license, used here with permission from Sotcheff et al. *BioRxiv* 2022 (*232*).

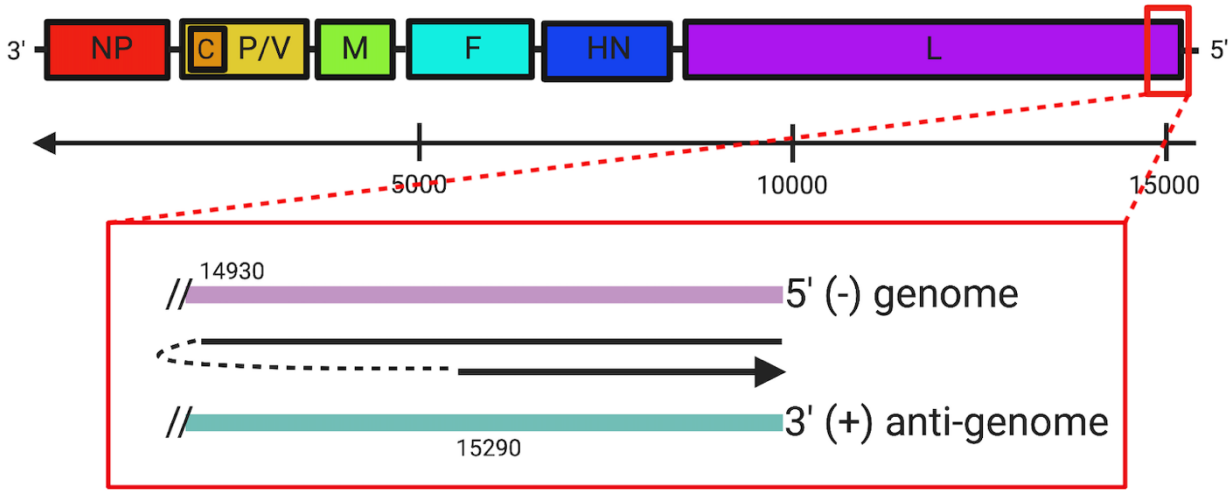


Figure 5.10: Sequences in the positive and negative sense strands of SeV that undergo recombination to produce copy-back RNAs, where light purple recombine and purple recombine. Figure created by S. Sotcheff using BioRender under license, used here with permission from Sotcheff et al. *BioRxiv* 2022 (*232*).

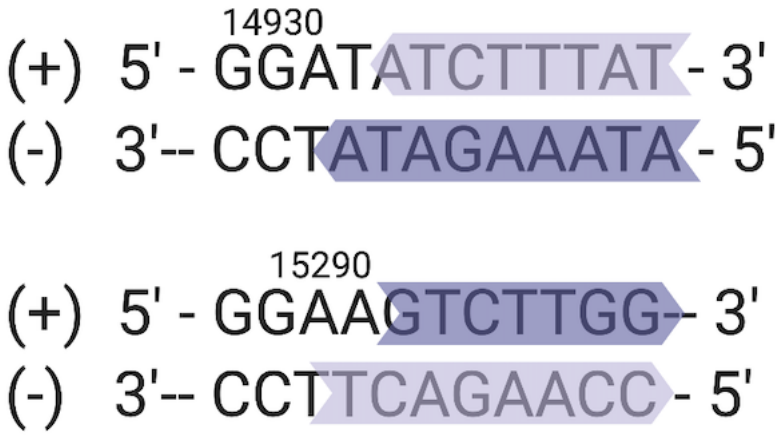


Figure 5.11: Depiction of a sample read identifying a copy-back and how that aligns to the 5' end of the negative sense strand and 3' end of the positive sense strand. Figure created by S. Sotcheff using BioRender under license, used here with permission from Sotcheff et al. *BioRxiv* 2022 (*232*).

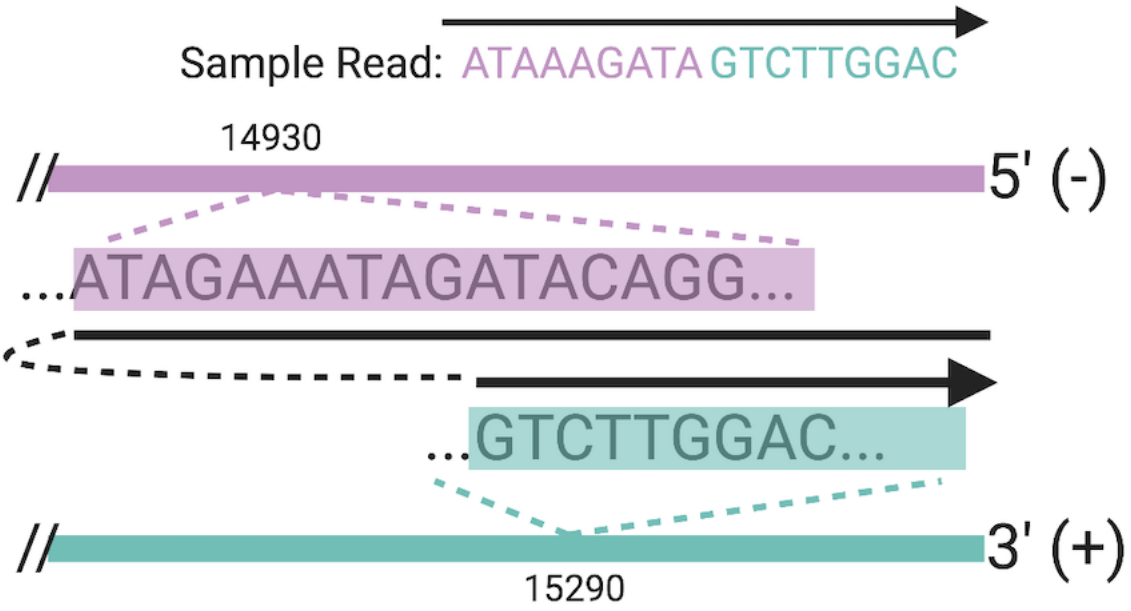
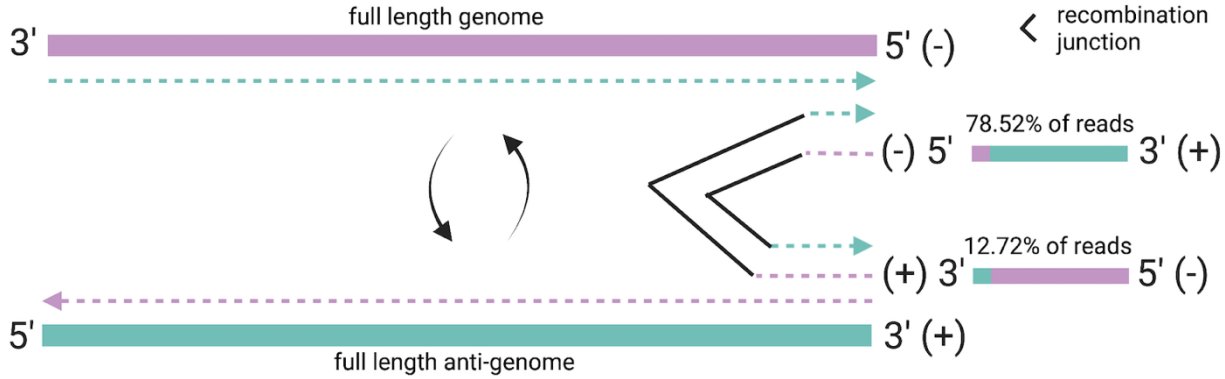


Figure 5.12: Schematic for formation of short secondary copy-back DVG where this is formed from replication of the original DVG. We suspect this as the percentage of reads that map to this region of the genome that have this recombination event is ~78%. Figure created by S. Sotcheff using BioRender under license, used here with permission from Sotcheff et al. *BioRxiv* 2022 (*232*).



AE006641.1) in parallel. As expected, the majority of the NGS reads mapped directly to the viral genomes. Interestingly however, we also observed a large number of 'virus-to-host' recombination events (reported in the Virus-to-Host_Recombinations.BEDPE file). The majority of these events where found at nucleotide position 7200 of the STIV genome, which is in a noncoding region between the genes encoding viral proteins C381 and D66 (**Figure 5.13**). Such recombination events are illustrated in **Figure 5.14**. Small (100-500 nt) portions of host genome appear to be inserted into the viral genome at sites containing flanking CCTAGG motifs, The column labeled 'Reads' describes the number of reads at that nucleotide position of STIV (which can be found in the 'cuttingsites' bedfiles and the final column was added to highlight the percentage of viral reads at that position containing that particular recombination event. Interestingly, STIV has previously been shown to undergo homologous recombination with various species of *Sulfolobus* at sites with the above-mentioned motifs (*246*). However, previous reports show this recombination occurring in the host genome as opposed to within the STIV genome.

This example demonstrates the ability of *ViReMa* to identify virus-to-host recombination events in the form of both insertions of host genome into the viral genome as well as the insertion of viral genome into the host genome. In addition, the updated BEDPE format makes these results easier to view and understand to the user.

Tiled-ClickSeq to investigate SARS-CoV-2 variants in patient samples

We can use high-throughput genomics to identify minority variants (MVs) and recombination events in SARS-CoV-2 that drive viral evolution and, as a result, pathogenicity and transmission. Whole genome targeted sequencing is typically accomplished by using primer pairs specific to the virus to synthesize cDNA amplicons

Figure 5.13: Host-to-virus recombination in *Sulfolobus* turreted icosohedral virus (STIV) as detected by *ViReMa*. Map of STIV and zoomed in on the region between C381 and D66 with a CCTAGG motif where host-to-virus recombination occurs. We also list the events found in our dataset at that site. Figure created by S. Sotcheff using BioRender under license, used here with permission from Sotcheff et al. *BioRxiv* 2022 (*232*).

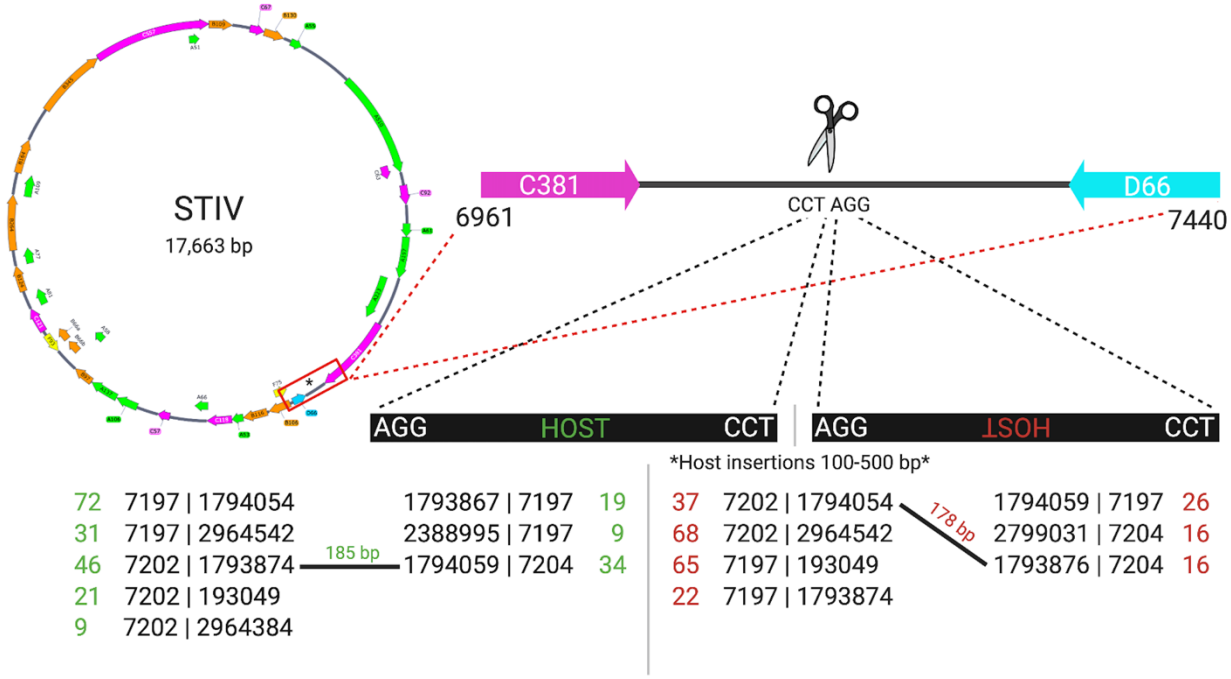


Figure 5.14: *ViReMa* output from the Virus-to-Host_Recombinations.BEDPE. Where 'Reference 1' is the first genome mapped to a read up until position 'Stop', 'Reference 2' is the second genome mapped from 'Start' and 'Sense 1'/'Sense 2' refers to the sense of the respective reference genome. The 'Count' is the number of reads containing the fusion event and 'Reads' is the total number of reads mapped to the virus at that position, which can be found in the 'cuttingsites' bedfile. 'Percent' is not part of the *ViReMa* output, but was rather calculated from comparing the 'Count' to 'Reads' at that position, used here with permission from Sotcheff et al. *BioRxiv* 2022 (*232*).

Reference 1	Stop	...	Reference 2	Start	...	Event Type	Count	Sense 1	Sense 2	Reads	Percent
STIV	7197	7198	SSP2 (Host)	1794054	1794055	Virus-Host-Fusion	72	-	-	7481	0.96
SSP2 (Host)	2964376	2964377	STIV	7196	7197	Virus-Host-Fusion	35	-	-	7551	0.46
SSP2 (Host)	1793867	1793868	STIV	7197	7198	Virus-Host-Fusion	19	-	-	7481	0.25
SSP2 (Host)	2388995	2388996	STIV	7197	7198	Virus-Host-Fusion	9	-	-	7481	0.12
STIV	7197	7198	SSP2 (Host)	2964542	2964543	Virus-Host-Fusion	7	-	-	7481	0.09
STIV	7202	7203	SSP2 (Host)	2964542	2964543	Virus-Host-Fusion	41	+	-	6371	0.64
STIV	7202	7203	SSP2 (Host)	1794054	1794055	Virus-Host-Fusion	37	+	-	6371	0.58
SSP2 (Host)	1794059	1794060	STIV	7197	7198	Virus-Host-Fusion	26	+	-	7481	0.35
SSP2 (Host)	2964548	2964549	STIV	7196	7197	Virus-Host-Fusion	16	+	-	7551	0.21
STIV	7197	7198	SSP2 (Host)	193049	193050	Virus-Host-Fusion	65	-	+	7481	0.87
STIV	7197	7198	SSP2 (Host)	1793874	1793875	Virus-Host-Fusion	22	-	+	7481	0.29
SSP2 (Host)	1793867	1793868	STIV	7204	7205	Virus-Host-Fusion	16	-	+	5586	0.29
SSP2 (Host)	2799031	2799032	STIV	7204	7205	Virus-Host-Fusion	16	-	+	5586	0.29
STIV	7202	7203	SSP2 (Host)	1793874	1793875	Virus-Host-Fusion	46	+	+	6371	0.72
SSP2 (Host)	1794059	1794060	STIV	7204	7205	Virus-Host-Fusion	34	+	+	5586	0.61
STIV	7202	7203	SSP2 (Host)	193049	193050	Virus-Host-Fusion	21	+	+	6371	0.33
STIV	7202	7203	SSP2 (Host)	2964384	2964385	Virus-Host-Fusion	9	+	+	6371	0.14

that can be used to produce NGS libraries. However, these primer pairs introduce bias that may hinder the detection of MVs or recombination events, for example if an event occurs outside of a primer pair or where a primer binds. The ARTIC protocol attempts to reduce this bias by using two primer sets (**Figure 2.4** – see Appendix B for primer sequences) (*65*). Although this certainly helps, this approach maintains the constraint of primer pairing in the RT step. Therefore, we found a way to combine the ARTIC method with ClickSeq, a click-chemistry based library preparation method which uses azido-NTPs to stochastically terminate the RT reaction, with sequence specific primers to SARS-CoV-2 (*63*). These are unpaired, or individual, primers (Appendix C). In addition, with Tiled-ClickSeq we include UMIs in the p5 adapter which allows for computational removal of PCR duplicates, but this is also controlled by the use of fewer PCR cycles (~18 compared to 35 in ARTIC).

In our eLife paper, we compared the use of ClickSeq and Tiled-ClickSeq to prepare libraries from 12 clinical samples that could be used to reconstruct the SARS-CoV-2 genome. We used *fastp* to process and quality filter the resulting fastq files and *bowtie2* to map them to the viral genome. We found that the use of sequence specific primers (Tiled-ClickSeq) greatly improved the percentage of reads mapping to SARS-CoV-2: where the use of random primers (ClickSeq) resulted in a maximum of 2.8% of reads mapping to virus and Tiled-ClickSeq resulted in 87.65% of reads mapping to virus on average (range 78.5-93.4%) (*63*). This is for libraries sequenced on an Illumina Next-Seq with ~2-5M paired end (2x150) reads per sample. We used this data to reconstruct the SARS-CoV-2 genome for each isolate using *pilon* and found 5-12 MVs per patient sample. Additionally, when comparing MVs identified in the isolates with each library preparation method we found the results were identical except for one MV found in a single isolate (WRCEVA_000510: T168C).

We also used the RNA from these 12 isolates to generate ARTIC libraries, which were sequenced on the ONT MinION (as were >600 bp fragments from the Tiled-ClickSeq libraries for comparison). The MinION data was mapped to SARS-CoV-2 using *minimap2*. The Tiled-ClickSeq libraries had at least 100X coverage over 99.6% of the genome. Although the overall coverage of the ARTIC libraries was similar, the coverage of the 5' UTR was greatly increased by the use of Tiled-ClickSeq as opposed to ARTIC (also shown in Illumina data in **Figure 5.15**). This is because ARTIC's use of both forward and reverse primers hinders sequencing of the absolute ends of the genome. Otherwise, MVs identified by preparing libraries using ClickSeq and Tiled-ClickSeq and sequencing on a Next Seq (Illumina) were virtually identical for each sample, except for C14220T found in WRCEVA_000514 MinION results (both Tiled-ClickSeq and ARTIC). We documented all MVs for each isolate in **Table 5.1.**

With these isolates and an additional 4 others we also wanted to investigate the occurrence of recombination events. Using *ViReMa* (described above) we identified a number of micro-insertion and deletion events that occurred in more than 2% of reads (**Table 5.2**). In seven isolates we also found a structural variant (SV) that was present in 2-50% of reads (Δ23585^23599), as well as a novel SV (Δ27619^27642) in WRCEVA_000504 that results in deletion of eight amino acids in ORF7a. Additionally, *ViReMa* identified thousands of recombination events corresponding to defective viral genomes with insertions or deletions that are primarily enriched in the 3' UTR (**Table 5.3**) – consistent with previous characterization of recombination in coronaviruses such as MERS, MHV, and SARS-CoV-2. Also, large deletions spanning from nucleotide position ~6,000-7,000 to the 3' UTR were also observed, again consistent with previously reported coronavirus recombination.

Figure 5.15: Comparison of coverage of the SARS-CoV-2 genome from a patient isolate with libraries generated using either ARTIC (magenta) or Tiled-ClickSeq (TCS, cyan) and sequenced on an Illumina NextSeq 550 with similar read counts. TCS coverage is smoother than ARTIC coverage. Figure created by S. Sotcheff using Microsoft Excel.

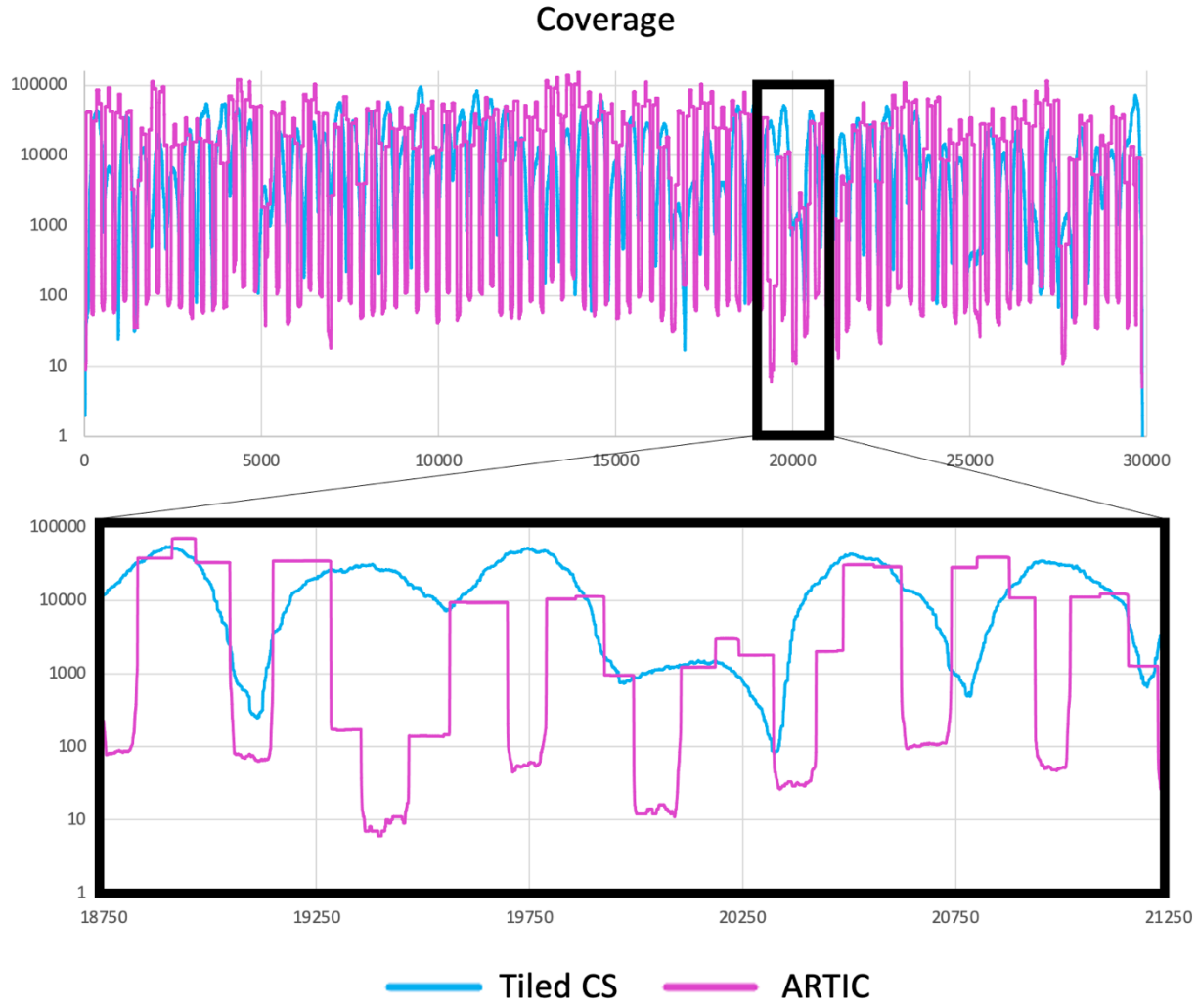


Table 5.1: Minority variants (MVs) found in the SARS-CoV-2 patient isolates after pylon was used to reconstruct the viral genome from sequencing data across all samples.

Nucleotides are color coded (C=blue, U=green, A=orange, G=yellow) and bolded values indicate a deviation from the reference nucleotide. Variant rate is calculated by adding all bolded values in a row and dividing that sum by the read depth at that nucleotide position times 100 to get the percentage.

Sample	Nucleotide Position	Reference Nucleotide	Read Depth	A	U	G	C	Variant Rate	Location	Result
WRCEVA_000501	12,049	C	2,116	**0**	**95**	**1**	2,020	4.50%	ORF1ab	N3928K
WRCEVA_000502	10,207	C	2,240	**0**	**118**	**0**	2,122	5.30%	-	-
WRCEVA_000502	16,050	U	3,853	**0**	3,322	**0**	**531**	13.80%	-	-
WRCEVA_000502	17,489	A	4,597	4,433	**162**	**1**	**1**	3.60%	ORF1ab	E5742V
WRCEVA_000502	21,526	A	8,749	6,508	**0**	**2,240**	**1**	25.60%	ORF1ab	I7088V
WRCEVA_000503	14,220	C	1,638	**1**	**463**	**0**	1,174	28.30%	-	-
WRCEVA_000504	1,556	A	2,828	2,499	**0**	**328**	**1**	11.60%	ORF1ab	I431V
WRCEVA_000504	27,925	C	2,857	**0**	**134**	**0**	2,723	4.70%	ORF8	T11I
WRCEVA_000507	19,515	A	2,393	2,295	**1**	**97**	**0**	4.10%	-	-
WRCEVA_000508	9,756	G	1,376	**28**	**0**	1,348	**0**	2.10%	ORF1ab	R3164H
WRCEVA_000508	26,056	G	2092	**0**	**86**	2,006	**0**	4.10%	ORF3a	D222Y
WRCEVA_000508	27,556	G	2066	**128**	**0**	1,938	**0**	6.20%	ORF7a	A55T
WRCEVA_000509	11,956	C	1962	**0**	**199**	**0**	1,763	10.10%	-	-
WRCEVA_000509	17,245	C	4,062	**2**	**470**	**0**	3,590	11.60%	ORF1ab	R5661C
WRCEVA_000509	18,005	U	5,408	**1**	4,949	**458**	**0**	8.50%	ORF1ab	L5915R
WRCEVA_000509	25,569	U	3,448	**4**	3,326	**113**	**5**	3.50%	-	-
WRCEVA_000509	27,919	U	839	**0**	809	**0**	**30**	3.60%	ORF8	I9T
WRCEVA_000509	28,767	C	2011	**0**	**109**	**0**	1,902	5.40%	N	T165I
WRCEVA_000511	3,003	U	2,880	**79**	2,787	**1**	**13**	2.70%	ORF1ab	V913E
WRCEVA_000511	10,738	U	4,580	**0**	4,440	**0**	**140**	3.10%	-	-
WRCEVA_000511	25,892	U	133	**0**	130	**0**	**3**	2.30%	ORF3a	I167T
WRCEVA_000511	28,001	G	1,414	**1**	**29**	1,384	**0**	2.10%	-	-
WRCEVA_000513	27,046	C	5,539	**0**	**138**	**0**	5,401	2.50%	M	T175M
WRCEVA_000514	11,603	A	5,405	5,075	**0**	**330**	**0**	6.10%	ORF1ab	M3780V
WRCEVA_000514	26,526	G	525	**0**	**20**	505	**0**	3.80%	M	A2S

Table 5.2: Micro- insertion and deletion events found in >2% of reads mapping to SARS-CoV-2 (in their respective genomic region) in patient isolates.

This were found using *ViReMa* to look at recombination events in these samples. Lines highlighted in blue indicate the same microindel event found in greater than 2% of reads at that position across 4 isolates.

Sample	Micro-Insertion/Deletion	Nucleotides	Variant Rate	Location	Result
WRCEVA_000502	Δ519^523	UGGUU	2.20%	ORF1AB	Frameshift
WRCEVA_000504	Δ29,686^29,693	CAGUGUGU	3.50%	3'UTR	-
WRCEVA_000505	Δ519^523	UGGUU	2.90%	ORF1AB	Frameshift
WRCEVA_000506	Δ519^523	UGGUU	3.80%	ORF1AB	Frameshift
WRCEVA_000509	Δ1,237^1,239	UCA	2.90%	ORF1AB	ΔH325
WRCEVA_000510	Δ686^694	AAGUCAUUU	5.10%	ORF1ab	ΔLSF141-143
WRCEVA_000511	Δ519^523	UGGUU	3.70%	ORF1AB	Frameshift
WRCEVA_000511	Δ10,811^10,813	CUU	3.10%	ORF1AB	ΔL3516
WRCEVA_000512	Δ29,750^29,759	GAUCGAGUG	10.00%	3'UTR	-

Table 5.3: Large deletion events found in SARS-CoV-2 in patient isolates, consistent with previously identified recombination events in coronaviruses.

Dark orange rows indicate events found in greater than 2% of reads mapping to those nucleotide positions, light orange rows indicate events found in 1-2% of reads mapping to those nucleotide positions.

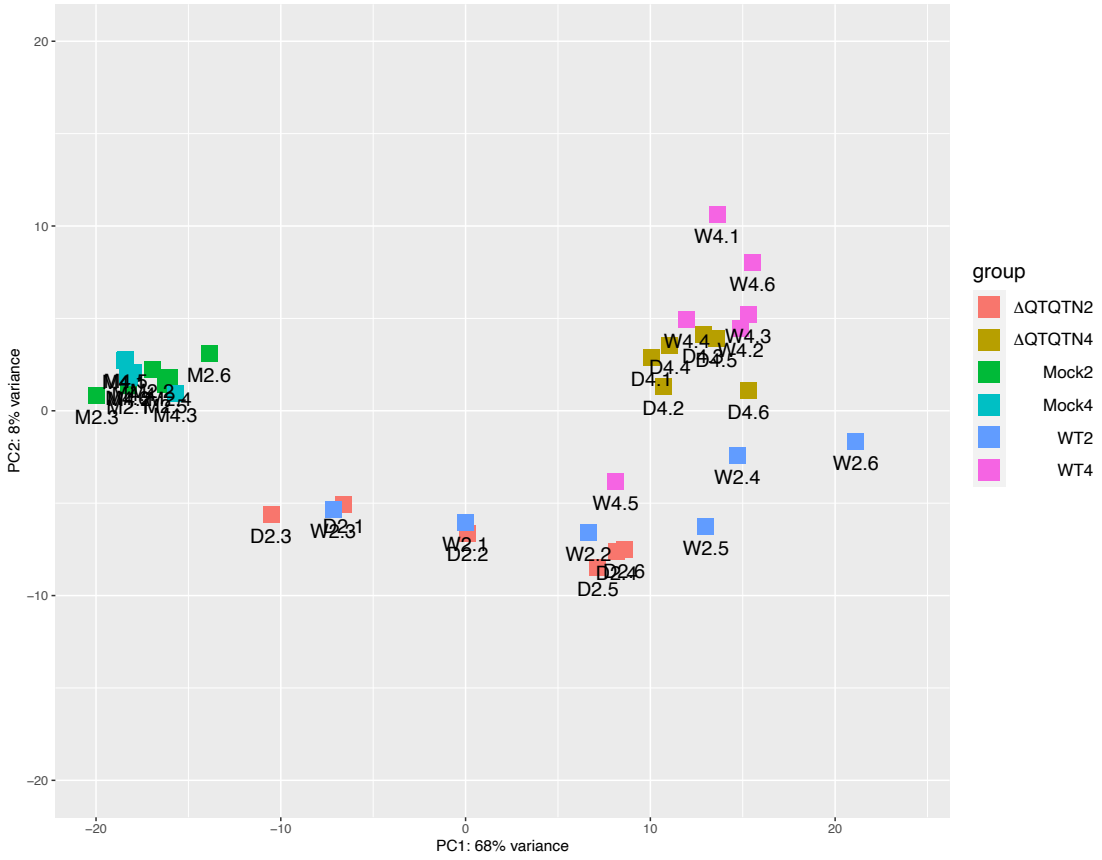
Genome	Stop	Start	Type	Count	Sense	Reads at Stop	Reads at Start	Recombination Rate (%)
SARS-CoV-2	70	26474	Deletion	206	+	541	7817	4.93
SARS-CoV-2	72	25388	Deletion	81	+	548	3601	3.90
SARS-CoV-2	71	27390	Deletion	37	+	541	5618	1.20
SARS-CoV-2	70	28261	Deletion	32	+	541	1581	3.02
SARS-CoV-2	73	26241	Deletion	26	+	547	3682	1.23
SARS-CoV-2	76	26496	Deletion	3	+	545	5893	0.09
SARS-CoV-2	75	27766	Deletion	2	+	549	1199	0.23
SARS-CoV-2	76	26491	Deletion	2	+	545	6373	0.06
SARS-CoV-2	75	26487	Deletion	2	+	549	6688	0.06
SARS-CoV-2	76	26483	Deletion	2	+	545	6906	0.05
SARS-CoV-2	76	26482	Deletion	2	+	545	7338	0.05
SARS-CoV-2	72	26477	Deletion	2	+	548	7622	0.05
SARS-CoV-2	68	26473	Deletion	2	+	542	7912	0.05
SARS-CoV-2	72	27044	Deletion	1	+	548	244	0.25
SARS-CoV-2	76	26503	Deletion	1	+	545	4910	0.04
SARS-CoV-2	76	26484	Deletion	1	+	545	6891	0.03
SARS-CoV-2	76	26481	Deletion	1	+	545	7358	0.03
SARS-CoV-2	77	25399	Deletion	1	+	537	5421	0.03

PAC-seq confirms QTQTN motif in SARS-CoV-2 aids pathogenesis

The coronavirus spike is a trimer of S1 and S2 (spike protein subunits) dimers. When a virion binds a host receptor, proteolytic cleavage between these two subunits facilitates fusion (*229, 247, 248*). The SARS-CoV-2 genome includes a furin cleavage site (PRRAR motif) within the spike protein upstream of the canonical site between S1 and S2 (*248*). This cleavage occurs in virions prior to release from a host cell and this plays a key role in spike processing, host infectivity and pathogenesis. Interestingly this furin cleavage site is not typically found in group 2B coronaviruses (*247*). Another motif found in SARS-CoV-2 that is novel is QTQTN just upstream of that furin cleavage site (*249*). In group 2B coronaviruses, this sequence is usually deleted, but this motif has been found in alpha, beta, and delta variants of SARS-CoV-2. Deletions of this motif have been found in patient samples, but how this motif may impact viral replication and virulence is not understood.

In collaboration with Dr. Vineet Menachery and his graduate student Michelle Vu, we synthesized 36 PAC-seq libraries from RNA extracted from hamster lung tissue, evenly spread across two time points (2 and 4 days post SARS-CoV-2 infection) and 3 conditions (WT, ΔQTQTN, and mock-infected) (*228*). The libraries were sequenced on the NextSeq 550 at UTMB with an average of ~13M reads/sample. The principal component analysis (PCA) plot shown in **Figure 5.16** shows clear clustering of the mock samples for both time points, and clustering of 2 DPI and 4 DPI for all infected samples. Of note, there is more variability at 2 DPI than at 4 DPI, and at both time points ΔQTQTN seems to shift towards mock. Further *DESeq2* analysis provided 16,297 genes to compare WT to ΔQTQTN SARS-CoV-2. Of those, 41 genes were differentially expressed (32 up-regulated in WT and 9 up-regulated in ΔQTQTN) at 2 DPI

Figure 5.16: Principal component analysis (PCA) plot for the PAC-seq data generated from golden hamster mRNAs extracted from mock-, WT-, or ΔQTQTN-SARS-CoV-2 infection. The PCA plot shows clear clustering of mock-infected samples from the infected co-horts and clustering by day for the WT- and ΔQTQTN infections. Within these days post infection (DPI) clusters it is evident that the mutant shifts towards the mock-infected samples. Plot created by S. Sotcheff using DESeq2 in R.



with a p-adjusted value (p-adj) of <0.1 and an absolute value of log2 fold change (|log2FC|) greater than 0.58 (or minimum of 50% increase/decrease). At 4 DPI, there were fewer differentially expressed genes (6 total, 4 up-regulated in WT and 2 in ΔQTQTN). Distributions of all genes based on their p-adj and log2FC are shown in the volcano plots in **Figures 5.17 and 5.18**. We confirmed the data quality was sufficient by generating and loading bedgraph files into the UCSC Genome Browser (*161*). We expected that reads would 1) map to the 3' end of genes and 2) reads would trail off as they were further from the 3' end because different reads would be of varying lengths (a block would potentially indicate PCR duplication of a single strand/read). These criteria were met as shown in **Figure 5.19** for LOC101842405. *Enrichr* was used to identify enrichment of different pathways, ontologies, or phenotypes in the 41 and 6 genes differentially expressed *(167-169)*. The *Enrichr* results suggest that the immune response and IL-9 regulation are up-regulated in the WT compared to the deletion mutant. Overall, these results suggest that the deletion of QTQTN (ΔQTQTN) attenuates coronavirus infection, indicating that inclusion of this motif in SARS-CoV-2 helps explain why this coronavirus has had greater global impact compared to SARS-CoV and MERS-CoV.

DISCUSSION

Using a collection of 'case studies', we demonstrate how *ViReMa* provides a versatile and robust platform for the detection and reporting of a diverse range of RNA recombination events found in viral genetic materials. These include straight-forward recombination events such as deletions in FHV; short duplications in the HIV genome, -ve to +ve sense copy-back RNAs in SeV; as well as virus to host fusion events in STIV. The flexibility of this platform derives from the strategy whereby read segments are mapped entirely independently from other

Figure 5.17: Volcano plot to display differentially expressed genes between WT- and ΔQTQTN SARS-CoV-2 infection in hamster lung 2 days post infection (DPI). WT (purple) and deletion mutant (orange). Plot created by S. Sotcheff using R.

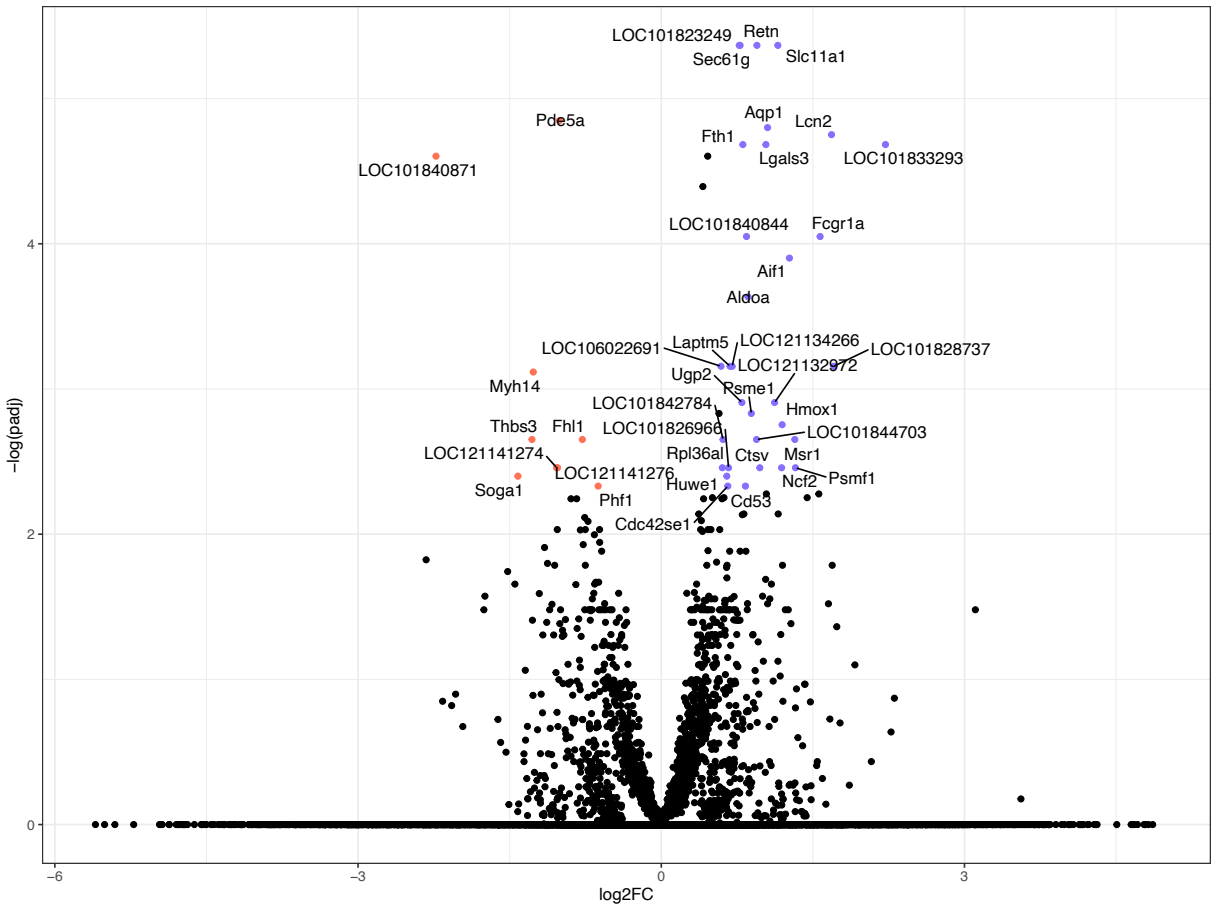


Figure 5.18: Volcano plot to display differentially expressed genes between WT- and ΔQTQTN SARS-CoV-2 infection in hamster lung 4 days post infection (DPI). WT (pink) and deletion mutant (gold). Plot created by S. Sotcheff using R.

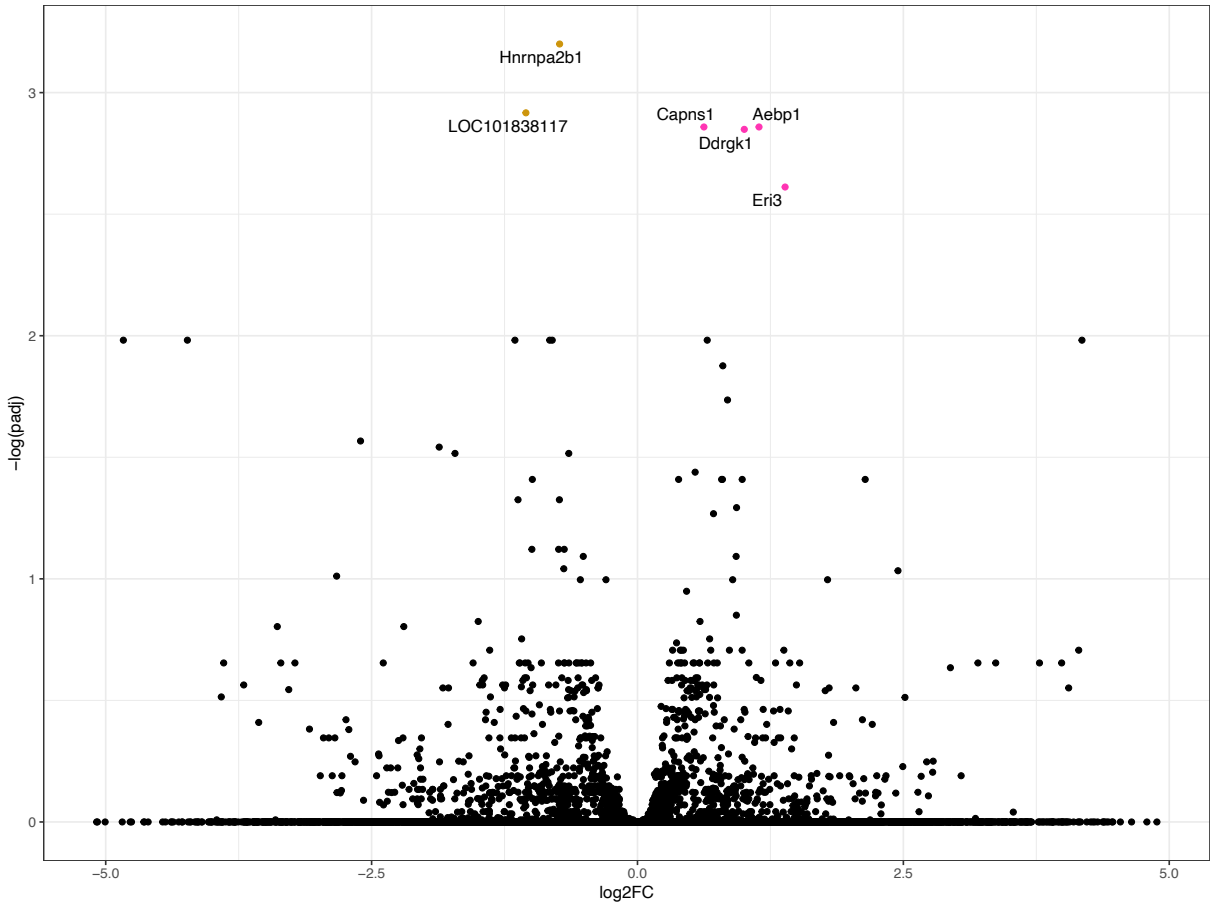
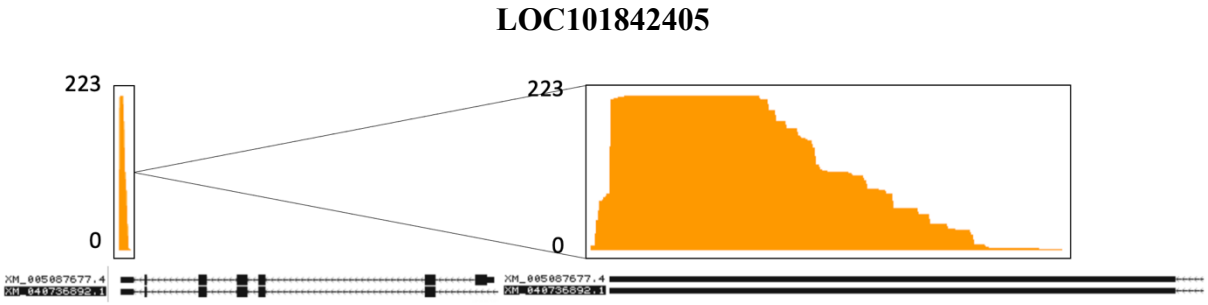


Figure 5.19: Bedgraph alignment of LOC101842405 showing that we have higher read density or coverage at the poly-A site and this coverage tapers off as we extend into the 3' of the gene. As PAC-seq primes at the poly-A tail and the RT reaction is stochastically terminated this confirms that the quality of the libraries was sufficient.



mapped read segments. *ViReMa* strictly requires a fixed number of mapped nucleotides to be mapped for each read segment, improving confidence in the identity of the recombination event reported, though at the expense of sensitivity. Here, we present a renovated handling of mismatches found in the read segments appropriate for the longer read lengths now commonly acquired in NGS platform. We also provide standardized outputs (.bed and .bedpe files) for both the read alignment and discovered recombination junctions that allow integration of the outputs of *ViReMa* with other sequence visualization software and bioinformatic packages. The versatility of *ViReMa* places it as useful tool all-in-one tool for the discovery, mapping and annotations of viral recombination events, particularly when *a priori* knowledge is lacking.

For situations where the predominant form of recombination event is a deletion, canonical pipelines for splice detection such as *HiSat2* (*250*), *STAR* (*251*) etc would provide a quicker and more sensitive route to detect these events as they do not rely on multiple iterations of bowtie alignment and use a dynamic seed-based heuristic. Additionally, where recombination/deletion events are known *a priori*, annotated junction events can be utilized in the index building steps of *HiSat2* and *STAR*. However, these aligners may be overly permissive in requiring mapped nucleotides on either side of a recombination junction resulting in the reporting of artifactual events. For eukaryotic splicing, this permissibility is acceptable as splice events are restricted to a small number of possible sites. However, this is not the case for viral recombination where junctions are often diverse and unpredictable (as previously noted for RNA viruses such as FHV). To reflect this, *ViReMa* enforces strict parameters for read mapping either side of a junction for all identified recombination events.

Recently, due to the increased interest in copy-back and snap-back defective viral genomes (DVGs) in negative-strand RNA viruses such as Respiratory Syncytial virus (*252*) and Ebola virus,

and number of platforms have been proposed to specifically map these species including DI-Tector (*70*) and *VODKA* (*72*). The *VODKA* algorithm uses a 'greedy' algorithm that will generate numerous reference sequences based upon putative copy/snap-back viral genome rearrangements and align short reads to this 'pseudo-library'. This constitutes a brute-force approach and requires prior information to focus the pseudo-reference library to regions suspected or known to form DVGs. Without this knowledge, the pseudo-reference library can become exponentially large with increased genome size. However, by providing a range of putative sequences with pre-defined recombination junctions, this constitutes a more sensitive approach as unambiguous mappings can be found across these junctions' sites whilst requiring a smaller seed region. Therefore, similar to finding known deletion/splice-like events using *HISAT2/STAR* aligners, in scenarios where information about the types of putative the copy/snap-back DVGs is known, VODKA may provide a more sensitive readout. In contrast, when no prior information is available, *ViReMa* provides an 'agnostic' approach, but at the expense of sensitivity.

Here, we present a novel method for sequencing SARS-CoV-2. Tiled-ClickSeq allowed us to prepare libraries that could be sequenced using an ONT MinION or an Illumina platform, so we were able to compare this method to both ClickSeq and ARTIC. We found that using sequence specific primers in the reverse transcription step greatly increased the percentage of reads mapping to the genome and therefore reduces the total number of reads necessary to get adequate coverage to detect MVs and recombination events relative to randomly primed library preparation methods (ClickSeq). Additionally, compared to other targeted priming methods such as ARTIC, the use of only reverse primers as opposed to primer pairs to generate cDNA amplicons increases coverage in the 5' UTR. Using all three methods we were able to find 5-12 MVs in each of the reconstructed genomes with only two MV deviations across different

methods. Additionally, we used our Tiled-ClickSeq data to analyze recombination events occurring in 16 patient samples using *ViReMa*. There were a handful of micro- insertion and deletion events found in these isolates at greater than 2% of reads, and a known SV present in 7 samples between 2 and 50% of reads. We also identified a novel SV and there were large deletions detected that were consistent with previously annotated defective SARS-CoV-2 genomes. Overall, this shows that Tiled-ClickSeq improves our ability to sequence genomes in patient samples by increasing coverage as well as maintaining the ability to accurately detect MVs and recombination events.

One motif that sets SARS-CoV-2 apart from other coronaviruses that infect humans is the QTQTN motif upstream of the furin cleavage site. Prior to our sequencing studies, the Menachery lab computationally modeled the spike protein in the absence of this motif and found that the exclusion of the motif may hinder furin cleavage. They also evaluated replication kinetics for WT SARS-CoV-2 and ΔQTQTN and found that in Vero E6 cells the deletion mutant had a fitness advantage although there was no difference in replication kinetics. Consistent with their findings for SARS-CoV-2 without a furin cleavage site, they also found the deletion mutant was attenuated in respiratory (Calu-3) cells. *In vivo* (male golden Syrian hamsters) they found that hamsters infected with WT SARS-CoV-2 steadily lost weight from 2 to 5 days post infection, when they began to recover they returned to their starting weight by 7 days post infection. In contrast, hamsters infected with ΔQTQTN SARS-CoV-2 experienced minimal weight loss, and in fact gained weight over the course of the infection with no obvious signs of disease. Both infections resulted in pulmonary lesions, but to a lesser extent in the deletion mutant infections. Interestingly, viral replication of ΔQTQTN was 10X greater in vivo compared

to WT as seen in nasal wash titers 1, 2 and 4 days post infection. In the lung however, replication was equivalent between both at 2 and 4 days post infection.

RNA was extracted from hamster lungs for transcriptomics analysis. We used PAC-seq to identify DEGs in response to WT SARS-CoV-2 and a deletion mutant in the lungs. We found that at two and four days post-infection the deletion mutant samples had expression patterns somewhere between that of the mock-infected and WT-infected samples. Additionally, the immune response was up-regulated in WT compared to the ΔQTQTN mutant. The Menachery lab also purified virus particles produced from infection of both viruses in Vero E6 and Calu-3 cells and found that ΔQTQTN showed reduced cleavage at S1/S2 – consistent with their computational modeling. Our PAC-seq results combined with the work done by Dr. Menachery's group suggests that the inclusion of the QTQTN motif increases pathogenicity of SARS-CoV-2 but not due to increased replication capacity but rather by more efficient processing of the spike protein allowing an increase in viral entry.

Interestingly, small deletions (ΔQTQTN as well as ΔPRRAR noted above) near the furin cleavage site was not found in any of the patient samples sequenced for the previous section. This suggests that although SARS-CoV-2 seems to generate various structural variants around the furin cleavage site in cell culture, these variants are not found in primary human samples and are therefore do not appear to be selected for in human infections.

Chapter 6 Perspectives

My studies have centered on the use and development of novel methods to generate and analyze NGS data to study RNA viruses. This includes the investigation of mutations and recombination events, determining changes in the host transcriptome in response to viral infection, and combining crosslinking with NGS to study capsid:vRNA interactions within virus particles. Truthfully, NGS can be used to study every phase of the replication cycle. As sequencing data is hypothesis forming in many cases, this section also allows for the discussion of additional studies, both NGS and others, that can be utilized in the future to validate and investigate further. Additionally, this section also allows us to contemplate the advantages and shortcomings of the various ClickSeq derived methods used in the studies highlighted above.

FUTURE DIRECTIONS

My main project was to determine transcriptional changes in human placental cells upon ZIKV infection. To accomplish this, we used PAC-seq to generate cDNA libraries that could be sequenced on an Illumina platform and the *DPAC* pipeline to identify differentially expressed and alternatively poly-adenylated genes. Here I discuss additional studies that may support or expand upon my findings.

Transcriptional Changes in JEG3 Cells Upon ZIKV Infection

The results suggested that there were more genes undergoing APA than differential expression based on normalized read counts. Additionally, we found that one of the up-regulated genes (SRSF11) appears to have the ability to bind near the poly-A signal of nascently transcribed RNAs. We found its binding motif (AGRRR) to be enriched just downstream of this

signal in our up-regulated PACs, but it was not enriched in the sequences surrounding down-regulated PACs. Therefore, we posited that the over-expressed SRSF11 may be binding near the poly-A signal and may aid in the recruitment or stabilization of core poly-adenylation factors (illustrated in **Figure 3.19**) at these sites, resulting in preferential poly-adenylation at these PACs.

We would like to determine if these APA events are relevant to the pathogenesis of ZIKV. First, we would need to determine if APA of various genes actually changes their expression levels. This could involve conducting the above mentioned study with an additional time point and cross-referencing APA at 16 hpi to the DEGs in the second time point. It could also include western blot or qRT-PCR analysis. Of course, there were hundreds of genes undergoing changes in poly-adenylation patterns, so the best choice would be to start with genes from pathways we found enriched in our APA genes; for example: virion assembly pathways were enriched for shortened 3' UTRs and T cell receptor signaling pathways were enriched for lengthened 3' UTRs. In addition to investigating if expression of these genes is altered it would also be pertinent to determine which of these genes, if any, include a binding site for SRSF11 by the poly-A signal.

Then we would like to determine if the over-expression of SRSF11 is necessary to cause these APA events. That being said, we would need to determine if the amount of SRSF11 protein is actually diminished upon ZIKV infection. This can be confirmed via western blot analysis. Once this is validated the studies to determine the effect of SRSF11 on APA can commence. These studies will include the knockdown (KD) of SRSF11 using a method like expressing short hairpin RNAs or similar to target the SRSF11 transcripts and either target them for degradation or prevent their translation. KD will be validated by RT-PCR and western blot analysis. With

successful SRSF11 KD, cells can be infected with ZIKV and PAC-seq can be used to determine any changes in alternative poly-adenylation. If up-regulation of SRSF11 in response to ZIKV was the primary cause of APA upon ZIKV infection, we would expect that KD of SRSF11 would hinder the up-regulation of the specific PACs found in the study presented above that contain the AGRRR motif just downstream of their poly-A signal.

Flavivirus Capsid and Non-packaging Roles in Pathogenesis

As noted in the introduction, various studies have found the flavivirus capsid protein outside of the replication factories at the ER membrane. Capsid has been found in the nucleus, nucleolus, associating with lipid droplets, interacting with various proteins in each region (see **Table 1.1**). In addition to these interactions with host proteins capsid, a small positively charged protein, has also been shown to interact with all types of nucleic acids *in vitro* regardless of being DNA or RNA, single or double-stranded, with no sequence specificity. Likely, this indicates a electrostatic interaction occurring between the negatively charged phosphate backbone of nucleic acids and the positively charged surface of capsid protein. Even packaging of the viral genome does not appear to occur via direct recognition of vRNA by capsid, but rather vRNA is chaperoned to capsid by viral NS2A. Other studies have shown that expression of capsid protein without the rest of the viral genome or proteins wreaks havoc on ribosome biogenesis and other host pathways. Considering all these features of capsid, it is reasonable to consider that capsid has roles in flavivirus pathogenesis outside of packaging the viral genome.

The transcriptomic study highlighted in Chapter 3 serves as a good control study to investigate additional roles of capsid. I have designed 15 capsid mutants, changing all positively charged sections to either swap one basic residue for another (R/K or K/R), or changing basic

residues to acidic or non-polar ones of similar molecular weight. These constructs (as well as a WT capsid sequence) are in mammalian expression plasmids. In the future these plasmids can be transfected into a human cell line (JEG3, Huh7, HEK293) and transcriptomic studies can be conducted to determine which changes to capsid elicit changes in host gene expression relative to WT.

For capsid mutants whose expression result in drastic changes in host expression we would be interested in a) determining how capsid may be eliciting these changes and b) how these changes in capsid affect viral fitness. To investigate how capsid may be eliciting these changes in host gene expression we would aim to identify protein and nucleic acid interactions of the capsid mutants. This can be accomplished by doing co-immunoprecipitation of capsid or CLIP-Seq (or a variation such as PAR-CLIP, etc) respectively. There is a commercially available antibody for capsid protein (GeneTex, GTX134186) that can be used for these studies. To investigate how each of these changes impacts viral fitness we can introduce these mutations with InFusion cloning into vectors for *in vitro* transcription of full-length virus. After transcription, these RNAs can be electroporated into human cells and we can measure changes in viral fitness using plaque assays or conducting qRT-PCR to determine changes in the number of infectious particles or amount of viral RNA produced respectively. Additionally, we can do western blots of capsid and envelope within cells and the supernatant to ensure viral proteins are made. This will allow us to ascertain if the reduction in virus particles is due to issues with packaging, replication, or translation respectively.

In the event that viral fitness is decreased due to packaging (i.e. viral genome is replicated and all viral proteins are present) we may also use v-PAR-CL to determine capsid:vRNA interactions in virions. We would expect to see less PAR-CL signal or more noise

if this were the case, as we would anticipate the formation of empty particles. We could also send purified virus for electron microscopy to confirm these results.

CLICKSEQ AND DERIVED METHODS TO STUDY RNA VIRUSES

Throughout my dissertation there has been use of various techniques centered on investigating RNA viruses and their impact on the host using NGS technologies. Each method involves some form of RNA extraction and may vary in their library preparation, sequencing platform, or data analysis, but overall the process is largely similar. Here I would like to discuss the various methods that were derived from ClickSeq to highlight their strengths relative to traditional RNAseq methods but also provide an idea for when the use of various methods may be appropriate or even inappropriate.

ClickSeq utilizes azido-NTPs during reverse transcription with a random (3' 6N) primer (see Appendix A) to stochastically terminate cDNA synthesis. This removes the need for fragmentation which is typically done in library preparation for traditional RNAseq, and also provides a site for copper catalyzed cyclo-addition ("clicking") an i5 adaptor to. The click reaction can only occur at the azido terminated 3' ends of cDNAs and therefore reduces the artefactual recombination of this method compared to the enzymatic ligation step found in other protocols. ClickSeq has been used to prepare libraries for viral genomes, but could also be used to study transcriptomics if sufficient read depth is achieved (similar to traditional RNAseq each sample would need ~30M reads for DGE studies, and over 80M reads/sample to look at AS) and ribosomal RNA is removed or mRNA is selected for using poly-A enrichment methods. Additionally, we have combined ClickSeq with the use of photo-activateable ribonucleosides and

crosslinking to study capsid:vRNA interactions within FHV particles. Indeed, ClickSeq as a method of library preparation is very versatile.

By simply changing the primer used in the reverse transcription step you can use ClickSeq to study both changes in gene expression and APA. An oligo-dT (21T primer) is used to bind the poly-A tail of mRNAs, removing the need for ribo-depletion or poly-A enrichment. This method is called Poly(A)-ClickSeq or PAC-seq and is coupled with *DPAC* analysis which also provides information on the use of alternative terminal exons. PAC-seq is a great library preparation method to use when investigating DGE and APA as the total number of reads necessary is reduced to ~10M per sample as we do not need coverage of entire genes, just their 3' UTR. Unfortunately, this method would not be suitable for AS studies for this same reason – we are only sequencing the 3' ends of mRNA transcripts.

Utilizing a pool of sequence specific primers, similar to ARTIC, allows the reliable sequencing of viral genomes in low quantities. We have used *primalseq* to design tiled primers across the SARS-CoV-2 genome for use in a method we have published called Tiled-ClickSeq. This increases the number of reads that map to virus, effectively lowering the total number of reads needed to achieve adequate coverage of the genome for reconstruction. Additionally, slight modifications in the amount of azido-NTPs can change the length of cDNAs produced in the RT step and allow us to sequence on an ONT platform in addition to Illumina. On either platform, resulting data is sufficient to find MVs, SVs, and various recombination events in patient isolates. Although the development of this method with SARS-CoV-2 is timely, we have also designed primers for a number of other viruses including Zika, dengue, Mayaro, chikungunya, and others. In fact, one of the graduate students in our lab has worked on mitigating the limitations of this technique by optimizing primers, number of PCR cycles, and modifying

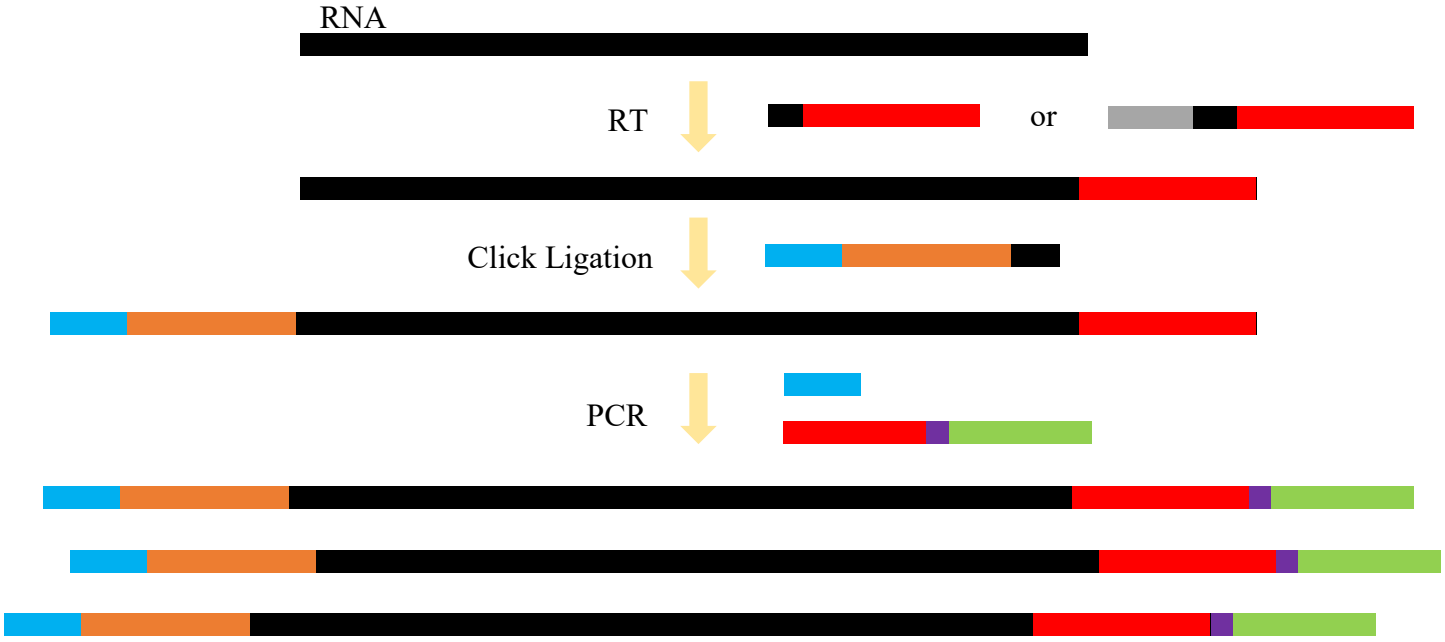
ethanol ratios in wash steps to be able to use this method to identify viruses in mosquito populations in Texas as surveillance and is involved with the West African Center for Emerging Infectious Diseases (WACEID). In the future we may consider designing primers for multi-partite viral genomes as well.

Overall, ClickSeq and derived methods have proven to be a cost-effective and efficient means of producing libraries from RNA (both viral and cellular) that has allowed the study of viruses by allowing the identification of variations in genome sequence, protein:RNA interactions, and host response. As analysis pipelines evolve and sequencing platforms advance we may expect to see an even larger breadth of studies that can be conducted using ClickSeq. Additionally, other modifications to the protocol may broaden the horizons of molecular virology as well.

Appendix A

Primer and oligo sequences for ClickSeq and Poly-A Click Seq where the same colors either base pair or are identical

Name	**Method (C/P/Both)**	**Sequence**
3' 6N primer for RT	C	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**NNNNNN**
Oligo-dT primer for RT	P	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**NNNN**TT TTTTTTTTTTTTTTTTTT
Hexynyl-Reverse-Complement Universal Illumina Primer	Both	5'Hexynyl-**NNNN**AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA GATCTCGGTGGTCGCCGTATCATT
Universal Primer Short (UPS)	Both	AATGATACGGCGACCACCGAG
Indexing primer example	Both	CAAGCAGAAGACGGCATACGAGAT*CGTGAT*GTGACTGGA GTTCAGACGTGT



Appendix B

ARTIC v3 primer set from: https://github.com/artic-network/artic-ncov2019/

name	pool	sequence
nCoV-2019_1_LEFT	nCoV-2019_1	ACCAACCAACTTTCGATCTCTTGT
nCoV-2019_1_RIGHT	nCoV-2019_1	CATCTTTAAGATGTTGACGTGCCTC
nCoV-2019_2_LEFT	nCoV-2019_2	CTGTTTTACAGGTTCGCGACGT
nCoV-2019_2_RIGHT	nCoV-2019_2	TAAGGATCAGTGCCAAGCTCGT
nCoV-2019_3_LEFT	nCoV-2019_1	CGGTAATAAAGGAGCTGGTGGC
nCoV-2019_3_RIGHT	nCoV-2019_1	AAGGTGTCTGCAATTCATAGCTCT
nCoV-2019_4_LEFT	nCoV-2019_2	GGTGTATACTGCTGCCGTGAAC
nCoV-2019_4_RIGHT	nCoV-2019_2	CACAAGTAGTGGCACCTTCTTTAGT
nCoV-2019_5_LEFT	nCoV-2019_1	TGGTGAAACTTCATGGCAGACG
nCoV-2019_5_RIGHT	nCoV-2019_1	ATTGATGTTGACTTTCTCTTTTTGGAGT
nCoV-2019_6_LEFT	nCoV-2019_2	GGTGTTGTTGGAGAAGGTTCCG
nCoV-2019_6_RIGHT	nCoV-2019_2	TAGCGGCCTTCTGTAAAACACG
nCoV-2019_7_LEFT	nCoV-2019_1	ATCAGAGGCTGCTCGTGTTGTA
nCoV-2019_7_LEFT_alt0	nCoV-2019_1	CATTTGCATCAGAGGCTGCTCG
nCoV-2019_7_RIGHT	nCoV-2019_1	TGCACAGGTGACAATTTGTCCA
nCoV-2019_7_RIGHT_alt5	nCoV-2019_1	AGGTGACAATTTGTCCACCGAC
nCoV-2019_8_LEFT	nCoV-2019_2	AGAGTTTCTTAGAGACGGTTGGGA
nCoV-2019_8_RIGHT	nCoV-2019_2	GCTTCAACAGCTTCACTAGTAGGT
nCoV-2019_9_LEFT	nCoV-2019_1	TCCCACAGAAGTGTTAACAGAGGA
nCoV-2019_9_LEFT_alt4	nCoV-2019_1	TTCCCACAGAAGTGTTAACAGAGG
nCoV-2019_9_RIGHT	nCoV-2019_1	ATGACAGCATCTGCCACAACAC
nCoV-2019_9_RIGHT_alt2	nCoV-2019_1	GACAGCATCTGCCACAACACAG
nCoV-2019_10_LEFT	nCoV-2019_2	TGAGAAGTGCTCTGCCTATACAGT
nCoV-2019_10_RIGHT	nCoV-2019_2	TCATCTAACCAATCTTCTTCTTGCTCT
nCoV-2019_11_LEFT	nCoV-2019_1	GGAATTTGGTGCCACTTCTGCT
nCoV-2019_11_RIGHT	nCoV-2019_1	TCATCAGATTCAACTTGCATGGCA
nCoV-2019_12_LEFT	nCoV-2019_2	AAACATGGAGGAGGTGTTGCAG
nCoV-2019_12_RIGHT	nCoV-2019_2	TTCACTCTTCATTTCCAAAAAGCTTGA
nCoV-2019_13_LEFT	nCoV-2019_1	TCGCACAAATGTCTACTTAGCTGT
nCoV-2019_13_RIGHT	nCoV-2019_1	ACCACAGCAGTTAAAACACCCT
nCoV-2019_14_LEFT	nCoV-2019_2	CATCCAGATTCTGCCACTCTTGT
nCoV-2019_14_LEFT_alt4	nCoV-2019_2	TGGCAATCTTCATCCAGATTCTGC
nCoV-2019_14_RIGHT	nCoV-2019_2	AGTTTCCACACAGACAGGCATT

nCoV-2019_14_RIGHT_alt2	nCoV-2019_2	TGCGTGTTTCTTCTGCATGTGC
nCoV-2019_15_LEFT	nCoV-2019_1	ACAGTGCTTAAAAAGTGTAAAAGTGCC
nCoV-2019_15_LEFT_alt1	nCoV-2019_1	AGTGCTTAAAAAGTGTAAAAGTGCCT
nCoV-2019_15_RIGHT	nCoV-2019_1	AACAGAAACTGTAGCTGGCACT
nCoV-2019_15_RIGHT_alt3	nCoV-2019_1	ACTGTAGCTGGCACTTTGAGAGA
nCoV-2019_16_LEFT	nCoV-2019_2	AATTTGGAAGAAGCTGCTCGGT
nCoV-2019_16_RIGHT	nCoV-2019_2	CACAACTTGCGTGTGGAGGTTA
nCoV-2019_17_LEFT	nCoV-2019_1	CTTCTTTCTTTGAGAGAAGTGAGGACT
nCoV-2019_17_RIGHT	nCoV-2019_1	TTTGTTGGAGTGTTAACAATGCAGT
nCoV-2019_18_LEFT	nCoV-2019_2	TGGAAATACCCACAAGTTAATGGTTTAAC
nCoV-2019_18_LEFT_alt2	nCoV-2019_2	ACTTCTATTAAATGGGCAGATAACAACTGT
nCoV-2019_18_RIGHT	nCoV-2019_2	AGCTTGTTTACCACACGTACAAGG
nCoV-2019_18_RIGHT_alt1	nCoV-2019_2	GCTTGTTTACCACACGTACAAGG
nCoV-2019_19_LEFT	nCoV-2019_1	GCTGTTATGTACATGGGCACACT
nCoV-2019_19_RIGHT	nCoV-2019_1	TGTCCAACTTAGGGTCAATTTCTGT
nCoV-2019_20_LEFT	nCoV-2019_2	ACAAAGAAAACAGTTACACAACAACCA
nCoV-2019_20_RIGHT	nCoV-2019_2	ACGTGGCTTTATTAGTTGCATTGTT
nCoV-2019_21_LEFT	nCoV-2019_1	TGGCTATTGATTATAAACACTACACACCC
nCoV-2019_21_LEFT_alt2	nCoV-2019_1	GGCTATTGATTATAAACACTACACACCCT
nCoV-2019_21_RIGHT	nCoV-2019_1	TAGATCTGTGTGGCCAACCTCT
nCoV-2019_21_RIGHT_alt0	nCoV-2019_1	GATCTGTGTGGCCAACCTCTTC
nCoV-2019_22_LEFT	nCoV-2019_2	ACTACCGAAGTTGTAGGAGACATTATACT
nCoV-2019_22_RIGHT	nCoV-2019_2	ACAGTATTCTTTGCTATAGTAGTCGGC
nCoV-2019_23_LEFT	nCoV-2019_1	ACAACTACTAACATAGTTACACGGTGT
nCoV-2019_23_RIGHT	nCoV-2019_1	ACCAGTACAGTAGGTTGCAATAGTG
nCoV-2019_24_LEFT	nCoV-2019_2	AGGCATGCCTTCTTACTGTACTG
nCoV-2019_24_RIGHT	nCoV-2019_2	ACATTCTAACCATAGCTGAAATCGGG
nCoV-2019_25_LEFT	nCoV-2019_1	GCAATTGTTTTTCAGCTATTTTGCAGT
nCoV-2019_25_RIGHT	nCoV-2019_1	ACTGTAGTGACAAGTCTCTCGCA
nCoV-2019_26_LEFT	nCoV-2019_2	TTGTGATACATTCTGTGCTGGTAGT
nCoV-2019_26_RIGHT	nCoV-2019_2	TCCGCACTATCACCAACATCAG
nCoV-2019_27_LEFT	nCoV-2019_1	ACTACAGTCAGCTTATGTGTCAACC
nCoV-2019_27_RIGHT	nCoV-2019_1	AATACAAGCACCAAGGTCACGG
nCoV-2019_28_LEFT	nCoV-2019_2	ACATAGAAGTTACTGGCGATAGTTGT
nCoV-2019_28_RIGHT	nCoV-2019_2	TGTTTAGACATGACATGAACAGGTGT
nCoV-2019_29_LEFT	nCoV-2019_1	ACTTGTGTTCCTTTTTGTTGCTGC
nCoV-2019_29_RIGHT	nCoV-2019_1	AGTGTACTCTATAAGTTTTGATGGTGTGT
nCoV-2019_30_LEFT	nCoV-2019_2	GCACAACTAATGGTGACTTTTTGCA

nCoV-2019_30_RIGHT	nCoV-2019_2	ACCACTAGTAGATACACAAACACCAG
nCoV-2019_31_LEFT	nCoV-2019_1	TTCTGAGTACTGTAGGCACGGC
nCoV-2019_31_RIGHT	nCoV-2019_1	ACAGAATAAACACCAGGTAAGAATGAGT
nCoV-2019_32_LEFT	nCoV-2019_2	TGGTGAATACAGTCATGTAGTTGCC
nCoV-2019_32_RIGHT	nCoV-2019_2	AGCACATCACTACGCAACTTTAGA
nCoV-2019_33_LEFT	nCoV-2019_1	ACTTTTGAAGAAGCTGCGCTGT
nCoV-2019_33_RIGHT	nCoV-2019_1	TGGACAGTAAACTACGTCATCAAGC
nCoV-2019_34_LEFT	nCoV-2019_2	TCCCATCTGGTAAAGTTGAGGGT
nCoV-2019_34_RIGHT	nCoV-2019_2	AGTGAAATTGGGCCTCATAGCA
nCoV-2019_35_LEFT	nCoV-2019_1	TGTTCGCATTCAACCAGGACAG
nCoV-2019_35_RIGHT	nCoV-2019_1	ACTTCATAGCCACAAGGTTAAAGTCA
nCoV-2019_36_LEFT	nCoV-2019_2	TTAGCTTGGTTGTACGCTGCTG
nCoV-2019_36_RIGHT	nCoV-2019_2	GAACAAAGACCATTGAGTACTCTGGA
nCoV-2019_37_LEFT	nCoV-2019_1	ACACACCACTGGTTGTTACTCAC
nCoV-2019_37_RIGHT	nCoV-2019_1	GTCCACACTCTCCTAGCACCAT
nCoV-2019_38_LEFT	nCoV-2019_2	ACTGTGTTATGTATGCATCAGCTGT
nCoV-2019_38_RIGHT	nCoV-2019_2	CACCAAGAGTCAGTCTAAAGTAGCG
nCoV-2019_39_LEFT	nCoV-2019_1	AGTATTGCCCTATTTTCTTCATAACTGGT
nCoV-2019_39_RIGHT	nCoV-2019_1	TGTAACTGGACACATTGAGCCC
nCoV-2019_40_LEFT	nCoV-2019_2	TGCACATCAGTAGTCTTACTCTCAGT
nCoV-2019_40_RIGHT	nCoV-2019_2	CATGGCTGCATCACGGTCAAAT
nCoV-2019_41_LEFT	nCoV-2019_1	GTTCCCTTCCATCATATGCAGCT
nCoV-2019_41_RIGHT	nCoV-2019_1	TGGTATGACAACCATTAGTTTGGCT
nCoV-2019_42_LEFT	nCoV-2019_2	TGCAAGAGATGGTTGTGTTCCC
nCoV-2019_42_RIGHT	nCoV-2019_2	CCTACCTCCCTTTGTTGTGTTGT
nCoV-2019_43_LEFT	nCoV-2019_1	TACGACAGATGTCTTGTGCTGC
nCoV-2019_43_RIGHT	nCoV-2019_1	AGCAGCATCTACAGCAAAAGCA
nCoV-2019_44_LEFT	nCoV-2019_2	TGCCACAGTACGTCTACAAGCT
nCoV-2019_44_LEFT_alt3	nCoV-2019_2	CCACAGTACGTCTACAAGCTGG
nCoV-2019_44_RIGHT	nCoV-2019_2	AACCTTTCCACATACCGCAGAC
nCoV-2019_44_RIGHT_alt0	nCoV-2019_2	CGCAGACGGTACAGACTGTGTT
nCoV-2019_45_LEFT	nCoV-2019_1	TACCTACAACTTGTGCTAATGACCC
nCoV-2019_45_LEFT_alt2	nCoV-2019_1	AGTATGTACAAATACCTACAACTTGTGCT
nCoV-2019_45_RIGHT	nCoV-2019_1	AAATTGTTTCTTCATGTTGGTAGTTAGAGA
nCoV-2019_45_RIGHT_alt7	nCoV-2019_1	TTCATGTTGGTAGTTAGAGAAAGTGTGTC
nCoV-2019_46_LEFT	nCoV-2019_2	TGTCGCTTCCAAGAAAAGGACG
nCoV-2019_46_LEFT_alt1	nCoV-2019_2	CGCTTCCAAGAAAAGGACGAAGA
nCoV-2019_46_RIGHT	nCoV-2019_2	CACGTTCACCTAAGTTGGCGTA

nCoV-2019_46_RIGHT_alt2	nCoV-2019_2	CACGTTCACCTAAGTTGGCGTAT
nCoV-2019_47_LEFT	nCoV-2019_1	AGGACTGGTATGATTTTGTAGAAAACCC
nCoV-2019_47_RIGHT	nCoV-2019_1	AATAACGGTCAAAGAGTTTTAACCTCTC
nCoV-2019_48_LEFT	nCoV-2019_2	TGTTGACACTGACTTAACAAAGCCT
nCoV-2019_48_RIGHT	nCoV-2019_2	TAGATTACCAGAAGCAGCGTGC
nCoV-2019_49_LEFT	nCoV-2019_1	AGGAATTACTTGTGTATGCTGCTGA
nCoV-2019_49_RIGHT	nCoV-2019_1	TGACGATGACTTGGTTAGCATTAATACA
nCoV-2019_50_LEFT	nCoV-2019_2	GTTGATAAGTACTTTGATTGTTACGATGGT
nCoV-2019_50_RIGHT	nCoV-2019_2	TAACATGTTGTGCCAACCACCA
nCoV-2019_51_LEFT	nCoV-2019_1	TCAATAGCCGCCACTAGAGGAG
nCoV-2019_51_RIGHT	nCoV-2019_1	AGTGCATTAACATTGGCCGTGA
nCoV-2019_52_LEFT	nCoV-2019_2	CATCAGGAGATGCCACAACTGC
nCoV-2019_52_RIGHT	nCoV-2019_2	GTTGAGAGCAAAATTCATGAGGTCC
nCoV-2019_53_LEFT	nCoV-2019_1	AGCAAAATGTTGGACTGAGACTGA
nCoV-2019_53_RIGHT	nCoV-2019_1	AGCCTCATAAAACTCAGGTTCCC
nCoV-2019_54_LEFT	nCoV-2019_2	TGAGTTAACAGGACACATGTTAGACA
nCoV-2019_54_RIGHT	nCoV-2019_2	AACCAAAAACTTGTCCATTAGCACA
nCoV-2019_55_LEFT	nCoV-2019_1	ACTCAACTTTACTTAGGAGGTATGAGCT
nCoV-2019_55_RIGHT	nCoV-2019_1	GGTGTACTCTCCTATTTGTACTTTACTGT
nCoV-2019_56_LEFT	nCoV-2019_2	ACCTAGACCACCACTTAACCGA
nCoV-2019_56_RIGHT	nCoV-2019_2	ACACTATGCGAGCAGAAGGGTA
nCoV-2019_57_LEFT	nCoV-2019_1	ATTCTACACTCCAGGGACCACC
nCoV-2019_57_RIGHT	nCoV-2019_1	GTAATTGAGCAGGGTCGCCAAT
nCoV-2019_58_LEFT	nCoV-2019_2	TGATTTGAGTGTTGTCAATGCCAGA
nCoV-2019_58_RIGHT	nCoV-2019_2	CTTTTCTCCAAGCAGGGTTACGT
nCoV-2019_59_LEFT	nCoV-2019_1	TCACGCATGATGTTTCATCTGCA
nCoV-2019_59_RIGHT	nCoV-2019_1	AAGAGTCCTGTTACATTTTCAGCTTG
nCoV-2019_60_LEFT	nCoV-2019_2	TGATAGAGACCTTTATGACAAGTTGCA
nCoV-2019_60_RIGHT	nCoV-2019_2	GGTACCAACAGCTTCTCTAGTAGC
nCoV-2019_61_LEFT	nCoV-2019_1	TGTTTATCACCCGCGAAGAAGC
nCoV-2019_61_RIGHT	nCoV-2019_1	ATCACATAGACAACAGGTGCGC
nCoV-2019_62_LEFT	nCoV-2019_2	GGCACATGGCTTTGAGTTGACA
nCoV-2019_62_RIGHT	nCoV-2019_2	GTTGAACCTTTCTACAAGCCGC
nCoV-2019_63_LEFT	nCoV-2019_1	TGTTAAGCGTGTTGACTGGACT
nCoV-2019_63_RIGHT	nCoV-2019_1	ACAAACTGCCACCATCACAACC
nCoV-2019_64_LEFT	nCoV-2019_2	TCGATAGATATCCTGCTAATTCCATTGT
nCoV-2019_64_RIGHT	nCoV-2019_2	AGTCTTGTAAAAGTGTTCCAGAGGT
nCoV-2019_65_LEFT	nCoV-2019_1	GCTGGCTTTAGCTTGTGGGTTT

nCoV-2019_65_RIGHT	nCoV-2019_1	TGTCAGTCATAGAACAAACACCAATAGT
nCoV-2019_66_LEFT	nCoV-2019_2	GGGTGTGGACATTGCTGCTAAT
nCoV-2019_66_RIGHT	nCoV-2019_2	TCAATTTCCATTTGACTCCTGGGT
nCoV-2019_67_LEFT	nCoV-2019_1	GTTGTCCAACAATTACCTGAAACTTACT
nCoV-2019_67_RIGHT	nCoV-2019_1	CAACCTTAGAAACTACAGATAAATCTTGGG
nCoV-2019_68_LEFT	nCoV-2019_2	ACAGGTTCATCTAAGTGTGTGTGT
nCoV-2019_68_RIGHT	nCoV-2019_2	CTCCTTTATCAGAACCAGCACCA
nCoV-2019_69_LEFT	nCoV-2019_1	TGTCGCAAAATATACTCAACTGTGTCA
nCoV-2019_69_RIGHT	nCoV-2019_1	TCTTTATAGCCACGGAACCTCCA
nCoV-2019_70_LEFT	nCoV-2019_2	ACAAAAGAAAATGACTCTAAAGAGGGTTT
nCoV-2019_70_RIGHT	nCoV-2019_2	TGACCTTCTTTTAAAGACATAACAGCAG
nCoV-2019_71_LEFT	nCoV-2019_1	ACAAATCCAATTCAGTTGTCTTCCTATTC
nCoV-2019_71_RIGHT	nCoV-2019_1	TGGAAAAGAAAGGTAAGAACAAGTCCT
nCoV-2019_72_LEFT	nCoV-2019_2	ACACGTGGTGTTTATTACCCTGAC
nCoV-2019_72_RIGHT	nCoV-2019_2	ACTCTGAACTCACTTTCCATCCAAC
nCoV-2019_73_LEFT	nCoV-2019_1	CAATTTTGTAATGATCCATTTTTGGGTGT
nCoV-2019_73_RIGHT	nCoV-2019_1	CACCAGCTGTCCAACCTGAAGA
nCoV-2019_74_LEFT	nCoV-2019_2	ACATCACTAGGTTTCAAACTTTACTTGC
nCoV-2019_74_RIGHT	nCoV-2019_2	GCAACACAGTTGCTGATTCTCTTC
nCoV-2019_75_LEFT	nCoV-2019_1	AGAGTCCAACCAACAGAATCTATTGT
nCoV-2019_75_RIGHT	nCoV-2019_1	ACCACCAACCTTAGAATCAAGATTGT
nCoV-2019_76_LEFT	nCoV-2019_2	AGGGCAAACTGGAAAGATTGCT
nCoV-2019_76_LEFT_alt3	nCoV-2019_2	GGGCAAACTGGAAAGATTGCTGA
nCoV-2019_76_RIGHT	nCoV-2019_2	ACACCTGTGCCTGTTAAACCAT
nCoV-2019_76_RIGHT_alt0	nCoV-2019_2	ACCTGTGCCTGTTAAACCATTGA
nCoV-2019_77_LEFT	nCoV-2019_1	CCAGCAACTGTTTGTGGACCTA
nCoV-2019_77_RIGHT	nCoV-2019_1	CAGCCCCTATTAAACAGCCTGC
nCoV-2019_78_LEFT	nCoV-2019_2	CAACTTACTCCTACTTGGCGTGT
nCoV-2019_78_RIGHT	nCoV-2019_2	TGTGTACAAAAACTGCCATATTGCA
nCoV-2019_79_LEFT	nCoV-2019_1	GTGGTGATTCAACTGAATGCAGC
nCoV-2019_79_RIGHT	nCoV-2019_1	CATTTCATCTGTGAGCAAAGGTGG
nCoV-2019_80_LEFT	nCoV-2019_2	TTGCCTTGGTGATATTGCTGCT
nCoV-2019_80_RIGHT	nCoV-2019_2	TGGAGCTAAGTTGTTTAACAAGCG
nCoV-2019_81_LEFT	nCoV-2019_1	GCACTTGGAAAACTTCAAGATGTGG
nCoV-2019_81_RIGHT	nCoV-2019_1	GTGAAGTTCTTTTCTTGTGCAGGG
nCoV-2019_82_LEFT	nCoV-2019_2	GGGCTATCATCTTATGTCCTTCCCT
nCoV-2019_82_RIGHT	nCoV-2019_2	TGCCAGAGATGTCACCTAAATCAA
nCoV-2019_83_LEFT	nCoV-2019_1	TCCTTTGCAACCTGAATTAGACTCA

nCoV-2019_83_RIGHT	nCoV-2019_1	TTTGACTCCTTTGAGCACTGGC
nCoV-2019_84_LEFT	nCoV-2019_2	TGCTGTAGTTGTCTCAAGGGCT
nCoV-2019_84_RIGHT	nCoV-2019_2	AGGTGTGAGTAAACTGTTACAAACAAC
nCoV-2019_85_LEFT	nCoV-2019_1	ACTAGCACTCTCCAAGGGTGTT
nCoV-2019_85_RIGHT	nCoV-2019_1	ACACAGTCTTTTACTCCAGATTCCC
nCoV-2019_86_LEFT	nCoV-2019_2	TCAGGTGATGGCACAACAAGTC
nCoV-2019_86_RIGHT	nCoV-2019_2	ACGAAAGCAAGAAAAAGAAGTACGC
nCoV-2019_87_LEFT	nCoV-2019_1	CGACTACTAGCGTGCCTTTGTA
nCoV-2019_87_RIGHT	nCoV-2019_1	ACTAGGTTCCATTGTTCAAGGAGC
nCoV-2019_88_LEFT	nCoV-2019_2	CCATGGCAGATTCCAACGGTAC
nCoV-2019_88_RIGHT	nCoV-2019_2	TGGTCAGAATAGTGCCATGGAGT
nCoV-2019_89_LEFT	nCoV-2019_1	GTACGCGTTCCATGTGGTCATT
nCoV-2019_89_LEFT_alt2	nCoV-2019_1	CGCGTTCCATGTGGTCATTCAA
nCoV-2019_89_RIGHT	nCoV-2019_1	ACCTGAAAGTCAACGAGATGAAACA
nCoV-2019_89_RIGHT_alt4	nCoV-2019_1	ACGAGATGAAACATCTGTTGTCACT
nCoV-2019_90_LEFT	nCoV-2019_2	ACACAGACCATTCCAGTAGCAGT
nCoV-2019_90_RIGHT	nCoV-2019_2	TGAAATGGTGAATTGCCCTCGT
nCoV-2019_91_LEFT	nCoV-2019_1	TCACTACCAAGAGTGTGTTAGAGGT
nCoV-2019_91_RIGHT	nCoV-2019_1	TTCAAGTGAGAACCAAAAGATAATAAGCA
nCoV-2019_92_LEFT	nCoV-2019_2	TTTGTGCTTTTTAGCCTTTCTGCT
nCoV-2019_92_RIGHT	nCoV-2019_2	AGGTTCCTGGCAATTAATTGTAAAAGG
nCoV-2019_93_LEFT	nCoV-2019_1	TGAGGCTGGTTCTAAATCACCCA
nCoV-2019_93_RIGHT	nCoV-2019_1	AGGTCTTCCTTGCCATGTTGAG
nCoV-2019_94_LEFT	nCoV-2019_2	GGCCCCAAGGTTTACCCAATAA
nCoV-2019_94_RIGHT	nCoV-2019_2	TTTGGCAATGTTGTTCCTTGAGG
nCoV-2019_95_LEFT	nCoV-2019_1	TGAGGGAGCCTTGAATACACCA
nCoV-2019_95_RIGHT	nCoV-2019_1	CAGTACGTTTTTGCCGAGGCTT
nCoV-2019_96_LEFT	nCoV-2019_2	GCCAACAACAACAAGGCCAAAC
nCoV-2019_96_RIGHT	nCoV-2019_2	TAGGCTCTGTTGGTGGGAATGT
nCoV-2019_97_LEFT	nCoV-2019_1	TGGATGACAAAGATCCAAATTTCAAAGA
nCoV-2019_97_RIGHT	nCoV-2019_1	ACACACTGATTAAAGATTGCTATGTGAG
nCoV-2019_98_LEFT	nCoV-2019_2	AACAATTGCAACAATCCATGAGCA
nCoV-2019_98_RIGHT	nCoV-2019_2	TTCTCCTAAGAAGCTATTAAAATCACATGG

Appendix C

Tiled-ClickSeq reverse transcription primers for SARS-CoV-2 with the nucleotide positions annotated.

Stop	Start	Sequence
43	68	AGAGAACAGATCTACAAGAGATCGA
147	172	GTTACTCGTGTCCTGTCAACGACAG
196	221	TCGGCTGCAACACGGACGAAACCGT
251	276	GGCTCTCCATCTTACCTTTCGGTCA
303	328	AAAACAGGCAAACTGAGTTGGACGT
340	365	GTCTCCAAAGCCACGTACGAGCACG
392	417	AAGTGCCATCTTTAAGATGTTGACG
491	516	CATGACCATGAGGTGCAGTTCGAGC
554	579	CAAGTGTCTCACCACTACGACCGTA
616	641	CTTACGAAGAAGAACCTTGCGGTAA
670	695	AAATGACTTTAGATCGGCGCCGTAA
784	809	TGCCCCTCCGTTAAGCTCACGCATG
821	846	CAGGGCCACAGAAGTTGTTATCGAC
884	909	AAGTGCATGAAGCTTTACCAGCACG
913	938	AGTGTCAATAAAGTCCAGTTGTTCG
956	981	CAATTTCATGCTCATGTTCACGGCA
986	1011	AGCTCTTTTCAGAACGTTCCGTGTA
1059	1084	AAATTTGGACATTCCCCATTGAAGG
1181	1206	GGTTGCATTCATTTGGTGACGCAAC
1288	1313	CAAATTCTCAGTGCCACAAAATTCG
1337	1362	CAACAGCATTTTGGGGTAAGTAACC
1416	1441	CCAGATTCATTATGGTATTCGGCAA
1469	1494	CACAGCCTCCAAAGGCAATAGTGCG
1546	1571	ATGGTTACAACCTATGTTAGCGCTA
1594	1619	AAGAAGGTTGTCATTAAGACCTTCG
1696	1721	ACTTGTGGAAGCAGAAAAAGATGCC
1726	1751	ATCCAAACCTTTCACAGTTTCCACA
1888	1913	GGAGAAAATTGATCGTACAACACGA
1913	1938	AATTTTGAGCAGTTTCAAGAGTGCG
2054	2079	CACCACCTGTAATGTAGGCCATTAC
2092	2117	AAAGATGTTAGTTAGCCACTGCGAA

2244	2269	GCACAGGTGACAATTTGTCCACCGA
2397	2422	TTTCTGTACAATCCCTTTGAGTGCG
2512	2537	GACAACTTCCTCTGTTAACACTTCT
2635	2660	CTTTTCTGTGTCTTTGATTTCGAGC
2731	2756	CACAGTGTCATCACCAAAAGTAACC
2874	2899	GCATCTGCCACAACACAGGCGAACT
2940	2965	CACTCATCTAAATCAATGCCCAGTG
3107	3132	GGTAATCATCTTCAGTACCATACTC
3237	3262	GTCTGATTGTCCTCACTGCCGTCTT
3311	3336	CAATAGTCTGAACAACTGGTGTAAG
3430	3455	AACATTGGCTGCATTAACAACCACT
3552	3577	ACACAACTACCACCCACTTTAAGTG
3615	3640	CCTTTGTTAACATTTGGGCCGACAA
3773	3798	AGACAGCTAAGTAGACATTTGTGCG
3872	3897	CTTCCTCTTTAGGAATCTCAGCGAT
4044	4069	GCAGAATCTGGATGAAGATTGCCAT
4163	4188	CACCAGCCTTTTTAGTAGGTATAAC
4202	4227	GCACTTTTCTCAAAGCTTTCGCTAG
4271	4296	TCTTTGCCTCCTCTACAGTGTAACC
4388	4413	CTGCATGTGCAAGCATTTCTCGCAA
4420	4445	CACACAGACAGGCATTAATTTGCGT
4539	4564	GACGCTACAGTTGTTTTACTGGTGT
4655	4680	GCACTTTGAGAGATCTCATATACCG
4786	4811	GGACCAATCTTTATAGGAACCAGCA
4900	4925	AAAGGTGATAACTTCACCATCTAGG
5006	5031	TCATTGACATGTCCACAACTTGCGT
5164	5189	ACTAGGATCAGTTGTGTGGTAGTAC
5281	5306	CAATGCAGTGGCAAGATAACAGTTG
5384	5409	AGATAAGTGCACAAAAGTTAGCAGC
5512	5537	TTGTCCACAAGTTTTACACACCACG
5657	5682	TAACAAAAGGTGACTCCTGTTGTAC
5741	5766	CACACTGGTAATTACCAGTGTACTC
5805	5830	TTTGTAAGTAAAGCACCGTCTATGC
5985	6010	ACAAGATCAATTGGTTGCTCTGTGA
6101	6126	CTCTTGAAGCAGGTTTCTTATAACC
6155	6180	AATCAATAGCCACCACATCACCATT
6296	6321	CTGGTTTTGTGCTCCAAAGACAACG
6449	6474	CGGTAGTTTTCACATTACACTCAAG
6554	6579	GACTAGAATTGTCTACATAAGCAGC

6658	6683	AGCTATAGTATCCCAAGGGACACTA
6736	6761	ACAAACACGGTTTAAACACCGTGTA
6882	6907	GAAGCCTCTAGACAAAATTTACCGA
7002	7027	ATTAAAACACCTAAAGCAGCGGTTG
7058	7083	AGTTCAAATAGCCTTCTCTGTAACC
7100	7125	GTATAGAACCAGTACAGTAGGTTGC
7226	7251	AAAACCACTCTGCAACTAAGCCAAA
7375	7400	TGAAATCGGGGCCATTTGTACAAGA
7508	7533	ATTCGACTCTTGTTGCTCTATTACG
7671	7696	AACTGTAGTGACAAGTCTCTCGCAA
7741	7766	GGAACCATTCTTCACTGTAACACTA
7829	7854	TAGCTCTCAGGTTGTCTAAGTTAAC
7947	7972	ATAGGTTGACACATAAGCTGACTGT
8011	8036	CATTTTAACTGCAACTTCCGCACTA
8104	8129	TGCAAGTTCAGCTTCTGCAGTTGCA
8251	8276	ACAACTATCGCCAGTAACTTCTATG
8321	8346	AGTCAATACAAGCACCAAGGTCACG
8365	8390	GTTGTGACTTTTTGCTACCTGCGCA
8418	8443	CGTAGTTGTTCAGACAATGACATGA
8490	8515	ACTTGTCTAGTAGTTGCACATGTCA
8594	8619	CAGCAACAAAAAGGAACACAAGTGT
8717	8742	AAGTATCTGTAGATGCTATGTCACG
8796	8821	GGGCAAGCTTTGTCATTAGTATAAC
8856	8881	GTGCCAGGCAAACCAGGCACGACAA
8888	8913	GCAAAAAGTCACCATTAGTTGTGCG
8932	8957	GTAACAGATGTTACCAACTGCACTA
8975	9000	CTGATGTTGCAAAGTCAGTGTACTC
9077	9102	CATAAGCAACAGAACCTTCTAGTAC
9209	9234	CACAAGTGCCGTGCCTACAGTACTC
9302	9327	CACCACAGAAAACTCCTGGTAAAGA
9428	9453	AGTAGGCAAGGCATGTTACTACGAT
9548	9573	CAGGTAAGAATGAGTAAACTGGTGT
9652	9677	AGGTACTAAAGGTGTGAACATAACC
9755	9780	AGGAAACACCATTAAAGACTACACG
9799	9824	ATTTAACAAAAAGGTGCACAGCGCA
9844	9869	AAGAGGTAATAGCACATCACTACGC
9966	9991	CTGAAGTCATTGAGAGCCTTTGCGA
10010	10035	TAGAGGTTTGTGGTGGTTGGTAAAG
10154	10179	CATGTCTTGGACAGTAAACTACGTC

10266	10291	CCAATAACCCTGAGTTGAACATTAC
10366	10391	AAAAGTCTGTCCTGGTTGAATGCGA
10396	10421	TGGTGAACCATTGTAACAAGCTAAC
10437	10462	ATAGTGAAATTGGGCCTCATAGCAC
10538	10563	CAGTTGGTAATTCCATATGGTGCAT
10637	10662	TAACTGTAATAGTTGTGTCCGTACC
10688	10713	GAAACCACCTGTCTCCATTTATAAC
10831	10856	TGAAGCACACATATCTAAAACGGCA
10888	10913	TAAAGCACTACCCAATATGGTACGT
10996	11021	GAGTAACAACCAGTGGTGTGTACCC
11108	11133	CAGACATAGCAATAATACCCATAGC
11234	11259	TCATAATACGCATCACCCAACTAGC
11317	11342	TAGTAACACTACAGCTGATGCATAC
11376	11401	ATAAGTGTCCACACTCTCCTAGCAC
11487	11512	GTAACTACACCTGAGTAGTTAGAAG
11527	11552	ACACATAAAAACAATACCTCTGGCC
11665	11690	AAGAGTCAGTCTAAAGTAGCGGTTG
11766	11791	ATGTTGAGTTTGAAGGCATCTATGC
11936	11961	GAATGTCATTGTGTAACTGGACACA
12102	12127	TATGATGGAAGGGAACTAAACTCTG
12258	12283	TCAGCCATCTTTTCCAACTTACGTT
12345	12370	GTGAAAAGCATTGTCTGCATAGCAC
12429	12454	GGTATTATGTTCAAGGGAACACAAC
12545	12570	CAACCTGTTGGATTTCCCACAATGC
12623	12648	CAATAAGAGGCCATGCTAAATTAGG
12735	12760	TCAGTGCAAGCAGTTTGTGTAGTAC
12773	12798	CTCCCTTTGTTGTGTTGTAGTAAGC
12873	12898	GGTTCCAGTTCTGTATAGATAGTAC
12983	13008	CAGCTAAACTACCAAGTACCATACC
13029	13054	TTGGCAGGCACTTCTGTTGCATTAC
13118	13143	TAGTGATTGGTTGTCCCCCACTAGC
13253	13278	TTGGATGATCTATGTGGCAACGGCA
13332	13357	GTAAAACCCACAGGGTCATTAGCAC
13397	13422	GTTGATCACAACTACAGCCATAACC
13473	13498	AAGACGGGCTGCACTTACACCGCAA
13527	13552	GATGTCAAAAGCCCTGTATACGACA
13600	13625	CTTCGTCCTTTTCTTGGAAGCGACA
13706	13731	GCAACAGCTGGACAATCCTTAAGTA
13815	13840	ATGCCTTAAAGCATAGACGAGGTCT

13981	14006	TTTTTAACAAAGCTTGGCGTACACG
14029	14054	GTACACCAACAATACCAGCATTTCG
14096	14121	GGCGTGGTTTGTATGAAATCACCGA
14264	14289	AAGAGTTTTAACCTCTCTTCCGTGA
14306	14331	TTTGGGTGGTATGTCTGATCCCAAT
14335	14360	TGCATCTGTCATCCAAACAGTTAAC
14460	14485	GAAGTGGTATCCAGTTGAAACTACA
14559	14584	GTGCATAGCAGGGTCAGCAGCATAC
14582	14607	AGTAATAGATTACCAGAAGCAGCGT
14621	14646	GTAAGTGCAGCTACTGAAAAGCACG
14777	14802	TCATAATCGCTGATAGCAGCATTAC
14922	14947	AGCTGATTTGTCTAGGTTGTTGACG
14966	14991	TCATAATAAAGTCTAGCCTTACCCC
14997	15022	AAGTGCATCTTGATCCTCATAACTC
15037	15062	GAGTTATAGTAGGGATGACATTACG
15084	15109	GGTGCGAGCTCTATTCTTTGCACTA
15168	15193	TCCTCTAGTGGCGGCTATTGATTTC
15257	15282	AGGTGAGGGTTTTCTACATCACTAT
15368	15393	CGGTGTGACAAGCTACAACACGTTG
15427	15452	CGCCACACATGACCATTTCACTCAA
15539	15564	AAAAGTGCATTAACATTGGCCGTGA
15584	15609	AAATTGCGGACATACTTATCGGCAA
15656	15681	GCGTAAAACTCATTCACAAAGTCTG
15761	15786	AAGTTCTTTATGCTAGCCACTAGAC
15915	15940	ATCTGGGTAAGGAAGGTACACATAA
16011	16036	ATCTATAGCTAAAGACACGAACCGT
16036	16061	TAGGATGTTTAGTAAGTGGGTAAGC
16122	16147	CATGTCTAACATGTGTCCTGTTAAC
16184	16209	GCCTCATAAAACTCAGGTTCCCAAT
16221	16246	CCCAACAGCCTGTAAGACTGTATGC
16292	16317	CAACATAAGAATGGTCTACGTATGC
16376	16401	CAACCTGGAGCATTGCAAACATACG
16484	16509	CCAAAAACTTGTCCATTAGCACACA
16632	16657	AGCTTTGAGCGTTTCTGCTGCAAAA
16729	16754	TAGGTTTACCAACTTCCCATGAAAG
16825	16850	AGTCACCTTTTTCAAAGGTGTACTC
16930	16955	GTGTAGGTGCACTTAATGGCATTAC
16955	16980	CTAACATAGTGCTCTTGTGGCACTA
17053	17078	GGAGTGTAGAATACTTTTGCATACC

17147	17172	GCATGAGAGCAAGCTGTATACACTA
17172	17197	CTTCTCACATAGTGCATCAACAGCG
17245	17270	TATCAAAACACTCTACACGAGCACG
17333	17358	TCAAAGACAACTATATCTGCTGTCG
17411	17436	CCAATGTACACATAGTGCTTAGCAC
17461	17486	GTGTGCCCTTAGTTAGCAATGTGCG
17563	17588	TGTCAACAATTTCAGCAGGACAACG
17674	17699	TTGCAGATGAAACATCATGCGTGAT
17736	17761	TTTTCTCCAAGCAGGGTTACGTGTA
17842	17867	CATAGTCATATTCTGAGCCCTGTGA
17997	18022	CCTACGTGGAATTTCAAGACTTGTA
18089	18114	GTAGGTGCCTGTGTAGGATGTAACC
18170	18195	CTATAGGTCATGTCCTTAGGTATGC
18241	18266	GGGTGATAAACATGTTAGGGTAACC
18309	18334	TCTAGTAGCATGACACCCCTCGACA
18382	18407	CTGTAGGTACAGCAACTAGGTTAAC
18449	18474	TGATCTCCAGGCGGTGGTTTAGCAC
18593	18618	GTCAACTCAAAGCCATGTGCCCATA
18648	18673	ATCACATAGACAACAGGTGCGCTCA
18784	18809	GGTTGCTTTGTAGGTTACCTGTAAA
18847	18872	TAGTCATGATTGCATCACAACTAGC
18875	18900	ACAAAGCACTCGTGGACAGCTAGAC
19029	19054	AATAGCTTTAGGGTTACCAATGTCG
19148	19173	GTGAATTTGTCAGAATGTGTGGCAT
19204	19229	TGGAATTAGCAGGATATCTATCGAC
19252	19277	CAGGCAAGTTAAGGTTAGATAGCAC
19385	19410	TTTCCATGAGACTCACATGGACTGT
19464	19489	GACAGCACCACCTAAATTGCAACGT
19507	19532	AAGCATCGAGATACAATCTGTACTC
19667	19692	ACTGGTACTTCACCCTGTTGTCCAT
19801	19826	CCTCTGGTACTGGTTTAATGTTGCG
19952	19977	AGTGGTGCACAAATCGTTTCAGTTG
20070	20095	TTTGGGACCTACAGATGGTTGTAAA
20165	20190	GGTAATTGTTGGACAACACCATCAA
20290	20315	AGGCATAGCCTTCTAATTTATACCG
20357	20382	AGTCCAATCAGTAGATGTAAACCAC
20466	20491	ACACTTAGATGAACCTGTTTGCGCA
20605	20630	GGCCATCTTTACACCAAAGCATAAA
20675	20700	AGATTAGGCATAGCAACACCCGGTT

20788	20813	GACACAGTTGAGTATATTTTGCGAC
20857	20882	CAGAACCAGCACCAAAATGTATAAC
20935	20960	TAAGATCTGAATCGACAAGCAGCGT
20997	21022	AGCTGTATGTACAGTTGCACAATCA
21138	21163	AGCCACGGAACCTCCAAGAGCTAGC
21218	21243	GTAACAAAGGCTGTCCACCATGCGA
21293	21318	TCTATTTGTTCGCGTGGTTTGCCAA
21422	21447	TTTAAAGACATAACAGCAGTACCCC
21581	21606	ACTGACTAGAGACTAGTGGCAATAA
21641	21666	CACGTGTGAAAGAATTAGTGTATGC
21662	21687	CTTTGTCAGGGTAATAAACACCACG
21786	21811	ACAGGGTTATCAAACCTCTTAGTAC
21893	21918	TAAGTAGGGACTGGGTCTTCGAATC
22075	22100	AAGGTCCATAAGAAAAGGCTGAGAG
22197	22222	AAACCCTGAGGGAGATCACGCACTA
22222	22247	ATCTACCAATGGTTCTAAAGCCGAA
22364	22389	GAAAAGTCCTAGGTTGAAGATAACC
22448	22473	TCAACGTACACTTTGTTTCTGAGAG
22584	22609	GATGCAAATCTGGTGGCGTTAAAAA
22630	22655	ATCAGCAACACAGTTGCTGATTCTC
22790	22815	TCTTTCCAGTTTGCCCTGGAGCGAT
22964	22989	CGGCCTGATAGATTTCAGTTGAAAT
23072	23097	CTACTACTCTGTATGGTTGGTAACC
23111	23136	AAACAGTTGCTGGTGCATGTAGAAG
23247	23272	GCAATGTCTCTGCCAAATTGTTGGA
23405	23430	CAACAGGGACTTCTGTGCAGTTAAC
23573	23598	TAGTCTGAGTCTGATAACTAGCGCA
23624	23649	TGTAGGCAATGATGGATTGACTAGC
23745	23770	ACTGATGTCTTGGTCATAGACACTG
23855	23880	CAACAGCTATTCCAGTTAAAGCACG
23873	23898	GGGTGTTTTTGTCTTGTTCAACAGC
24043	24068	TTTGATGAAGCCAGCATCTGCAAGT
24199	24224	GGTCCAACCAGAAGTGATTGTACCC
24348	24373	GAGTCTTGAATTTTGCCAATAGCAC
24442	24467	GGAGCTAAGTTGTTTAACAAGCGTG
24506	24531	CAGCCTCAACTTTGTCAAGACGTGA
24542	24567	GAAGTCTGCCTGTGATCAACCTATC
24651	24676	TTTGATTGTCCAAGTACACACTCTG
24744	24769	GGGACATAAGTCACATGCAAGAAGA

24849	24874	ACAAACCAGTGTGTGCCATTTGAAA
24967	24992	AGGTTGCAAAGGATCATAAACTGTG
25111	25136	CTTGGCAACCTCATTGAGGCGGTCA
25155	25180	TACTTTCCAAGTTCTTGGAGATCGA
25191	25216	AGCCAAATGTACCATGGCCATTTTA
25339	25364	TCCTTTGAGCACTGGCTCAGAGTCG
25489	25514	GTGAGGCTTGTATCGGTATCGTTGC
25521	25546	TGCAACGCCAACAATAAGCCATCCG
25565	25590	TTGAGGGTTATGATTTTGGAAGCGC
25674	25699	TTCAAGGCCAGCAGCAACGAGCAAA
25789	25814	AAAGTAATGGGTTTTTGGAACGGCA
25890	25915	GCCATCACCTGAAGTAATGACAATT
26049	26074	ATGTTCAACACCAGTGTCTGTACTC
26147	26172	ATTACTGGATTAACAACTCCGGATG
26204	26229	GCTTGTGCTTACAAAGGCACGCTAG
26233	26258	CGAATGAGTACATAAGTTCGTACTC
26256	26281	TAACGTACCTGTCTCTTCCGAAACG
26292	26317	CACGAAAGCAAGAAAAAGAAGTACG
26325	26350	AGTAAGGATGGCTAGTGTAACTAGC
26366	26391	ACGTTAACAATATTGCAGCAGTACG
26453	26478	TTCGTTTAGACCAGAAGATCAGGAA
26538	26563	TAAGCTCTTCAACGGTAATAGTACC
26622	26647	TGTTGGCATAGGCAAATTGTAGAAG
26767	26792	AAGCCTACAAGACAAGCCATTGCGA
26839	26864	GGATTGAATGACCACATGGAACGCG
26882	26907	AATAGTGCCATGGAGTGGCACGTTG
26940	26965	GTCCACGAAGGATCACAGCTCCGAT
26966	26991	TAGATGGTGTCCAGCAATACGAAGA
26994	27019	TAGGCAGGTCCTTGATGTCACAGCG
27078	27103	CAAAACCTGAGTCACCTGCTACACG
27105	27130	TGCCAATCCTGTAGCGACTGTATGC
27212	27237	GCTATAGTAACCTGAAAGTCAACGA
27366	27391	CGTTTAATCAATCTCCATTGGTTGC
27412	27437	AAGTAGCGAGTGTTATCAGTGCCAA
27510	27535	ATGAAATGGTGAATTGCCCTCGTAT
27622	27647	TAGGTGAAACTGATCTGGCACGTAA
27703	27728	GTGTTATAAACACTATTGCCGCAAC
27789	27814	GGAATAGCAGAAAGGCTAAAAAGCA
27876	27901	ATTTCATGTTCGTTTAGGCGTGACA

27928	27953	CATTCTTGGTGAAATGCAGCTACAG
28071	28096	TAGAACCAGCCTCATCCACGCACAA
28189	28214	TCTTCATAGAACGAACAACGCACTA
28319	28344	TTGAATCTGAGGGTCCACCAAACGT
28388	28413	TGGGTAAACCTTGGGGCCGACGTTG
28431	28456	TTGCCATGTTGAGTGAGAGCGGTGA
28473	28498	TTAATTGGAACGCCTTGTCCTCGAG
28556	28581	CTTTCATTTTACCGTCACCACCACG
28665	28690	AAGGCTCCCTCAGTTGCAACCCATA
28745	28770	TTGTTCCTTGAGGAAGTTGTAGCAC
28844	28869	GAGTTGAATTTCTTGAACTGTTGCG
28920	28945	AGCAGCAGCAAAGCAAGAGCAGCAT
29072	29097	CGAAAGCTTGTGTTACATTGTATGC
29102	29127	TTCCTTGGGTTTGTTCTGGACCACG
29229	29254	GAAGGTGTGACTTCCATGCCAATGC
29259	29284	ATGGCACCTGTGTAGGTCAACCACG
29348	29373	CTGTTGGTGGGAATGTTTTGTATGC
29417	29442	GCTGTTTCTTCTGTCTCTGCGGTAA
29571	29596	ATCGTAAACGGAAAAGCGAAAACGT
29601	29626	GAATTCATTCTGCACAAGAGTAGAC
29698	29723	GGTGGCTCTTTCAAGTCCTCCCTAA
29753	29778	CTAGCATTGTTCACTGTACACTCGA

Appendix D

```
#!/bin/bash

#NOTES on how to run:
#
# place this script into the folder you will be running the command from
#
# make sure your MinION fastq files are already compiled by barcode (use cat in previous folder)
# copy the compiled fastq files into the folder you will be running this command from
#
# use a for loop to run the command for all of your compiled fastq files
# example:
# for i in $(ls /[path]/*.fastq); do ./Flair_A $i; done
#
# .gtf file downloaded from UCSC Genome Browser
#
# hg19.chrom.sizes are indexed

File=$1
Folder=$2
Root=${File##*/}
Root=${Root%%_S*}

echo "Working on "$Root" now....."

#align
```

```
python3 flair.py align -r $File -g [GENOME] -o $Root -m minimap2

#correct

python3 flair.py correct -q $Root'.bed' -g [GENOME] -o 'Corrected_'$Root -f human_hg19_Ensembl.gtf

#convert to psl

python3 bed_to_psl.py hg19.chrom.sizes 'Corrected_'$Root'_all_corrected.bed' $Root'_corrected.psl'
```

Appendix E

```
#!/bin/bash


#NOTES on how to run:
#
# place this script into the folder you will be running the command from
#
# be sure to have made $j_reads_manifest.tsv and $j-comparisons.txt files
# and placed them in this folder
#
# note that $j should be the name of the folder you're running from - if I ran this in the folder JEG3_PRV, then my file names would need to be JEG3_PRV_reads_manifest.tsv and JEG3_PRV_comparisons.txt
#
# how to make $j'_reads_manifest.tsv' file
# in a text file note the sample name, condition, batch, and fastq file for each sample
# example:
# sample1	conditionA	batch1	sample1.fastq
# sample2	conditionA	batch1	sample2.fastq
# sample3	conditionB	batch1	sample3.fastq
# sample4	conditionB	batch1	sample4.fastq
# sample5	conditionA	batch2	sample5.fastq
# sample6	conditionA	batch2	sample6.fastq
# sample7	conditionB	batch2	sample7.fastq
# sample8	conditionB	batch2	sample8.fastq
# sample9	conditionC	batch1	sample9.fastq
```

```
# sample10	conditionC	batch1	sample10.fastq
# sample11	conditionC	batch2	sample11.fastq
# sample12	conditionC	batch2	sample12.fastq
#
# ** if you do not have batches include the column but denote all as "batch1"
#
# how to make $j'-comparisons.txt' file
# in a text file make a line for each comparison based on the conditions listed in the reads manifest file
# one column per condition
# example:
# conditionA	conditionB
# conditionA	conditionC
# conditionB	conditionC
#
# to run this script simply use the following command in the directory you want to run from:
# ./Flair_B
#
# .gtf files downloaded from UCSC Genome Browser


i=$(pwd)
j=${i##*/}
echo $j


#concatenate
echo "Concatenating corrected .psl and .fastq files"
```

```
cat *.psl > $j'_All_Corrected.psl'

cat *.fastq > $j'_All.fastq'

#collapse

echo "Performing Flair collapse"

python3 flair.py collapse -g [GENOME] -r $j'_All.fastq' -q $j'_All_Corrected.psl' -f human_hg19_Ensembl.gtf -m minimap2

#quantify

echo "Quantifying splice variants"

python3 flair.py quantify -r $j'_reads_manifest.tsv' -i flair.collapse.isoforms.fa -m minimap2 -o $j'_counts_matrix.tsv'

#diffExp

echo "Comparing splice variant counts across sample groups"

python3 flair.py diffExp -q $j'_counts_matrix.tsv' -o $j'_diffExp' -t 4

#diffSplice

echo "Writing all comparisons"
```

```
while read cond1 cond2; do
        echo "$cond1'_vs_'$cond2"

        python3 flair.py diffSplice -i flair.collapse.isoforms.psl -q $j'_counts_matrix.tsv' -t 4 --
conditionA $cond1 --conditionB $cond2 -o $j'_diffSplice_'$cond1'_vs_'$cond2

        done < "$j-comparisons.txt"

#convert psl to bed

echo "Converting isoforms.psl to .bed"

convert2bed -i psl -o bed <flair.collapse.isoforms.psl> flair.collapse.isoforms.psl.bed

#replace accessions with gene names

echo "Replacing accessions with gene names in the isoforms.psl.bed file"

while read input output; do
        echo "$input'_to_'$output"

        sed -i "s/$input/$output/g" flair.collapse.isoforms.psl.bed

done < "human_hg19_Ensembl_genenames.gtf"

echo "Replacing accessions with gene names in the counts_matrix.tsv file"
```

```
while read input output; do
        echo "$input'_to_'$output"

        sed -i "s/$input/$output/g" $j'_counts_matrix.tsv'

done < "human_hg19_Ensembl_genenames.gtf"
```

Appendix F

```
#load necessary packages

if(!require(devtools)){install.packages("devtools")}
devtools::install_github("xnnba1984/DoubletCollection")

library(Seurat)
library(cowplot)
library(scater)
library(Matrix)
library(dplyr)
library(patchwork)
library(DESeq2)
library(SeuratObject)
library(dittoSeq)
library(DoubletCollection)
library(magrittr)
library(data.table)

methods=c("scDblFinder")

# Load data and quality filter to only include cells with:
# any gene must be expressed in at least 10 cells to be included
# expression between 700 and 2500 genes
# less than 5% genes expressed in a given cell are mitochondrial
# a doublet score less than 0.2
```

```
Sample1 <- Read10X(data.dir = "~/[PATH]/Sample1/filtered_feature_bc_matrix/")
Sample1_so <- CreateSeuratObject(counts = Sample1, project = "Sample1", min.cells = 10, min.features = 700)
Sample1_so[["percent.mt"]] <- PercentageFeatureSet(Sample1_so, pattern = "^Mt-")
Sample1_so[["doublet.score"]]=FindScores(Sample1, methods)
Sample1.qc<-subset(Sample1_so, subset = nFeature_RNA > 700 & nFeature_RNA < 2500 & percent.mt<5 & doublet.score<0.2)

Sample2 <- Read10X(data.dir = "~/[PATH]/Sample2/filtered_feature_bc_matrix/")
Sample2_so <- CreateSeuratObject(counts = Sample2, project = "Sample2", min.cells = 10, min.features = 700)
Sample2_so[["percent.mt"]] <- PercentageFeatureSet(Sample2_so, pattern = "^Mt-")
Sample2_so[["doublet.score"]]=FindScores(Sample2, methods)
Sample2.qc<-subset(Sample2_so, subset = nFeature_RNA > 700 & nFeature_RNA < 2500 & percent.mt<5 & doublet.score<0.2)

Sample3 <- Read10X(data.dir = "~/[PATH]/Sample3/filtered_feature_bc_matrix/")
Sample3_so <- CreateSeuratObject(counts = Sample3, project = "Sample3", min.cells = 10, min.features = 700)
Sample3_so[["percent.mt"]] <- PercentageFeatureSet(Sample3_so, pattern = "^Mt-")
Sample3_so[["doublet.score"]]=FindScores(Sample3, methods)
Sample3.qc<-subset(Sample3_so, subset = nFeature_RNA > 700 & nFeature_RNA < 2500 & percent.mt<5 & doublet.score<0.2)

Sample4 <- Read10X(data.dir = "~/[PATH]/Sample4/filtered_feature_bc_matrix/")
```

```
Sample4_so <- CreateSeuratObject(counts = Sample4, project = "Sample4", min.cells = 10, min.features = 700)
Sample4_so[["percent.mt"]] <- PercentageFeatureSet(Sample4_so, pattern = "^Mt-")
Sample4_so[["doublet.score"]]=FindScores(Sample4, methods)
Sample4.qc<-subset(Sample4_so, subset = nFeature_RNA > 700 & nFeature_RNA < 2500 & percent.mt<5 & doublet.score<0.2)

#Create a merged list
MergedList<- list(Sample1.qc, Sample2.qc, Sample3.qc,Sample4.qc)

#first perform standard normalization and variable feature selection.
MergedList <- lapply(X = MergedList, FUN = function(x) {
 x <- NormalizeData(x, verbose = FALSE)
 x <- FindVariableFeatures(x, verbose = FALSE)
        })

#Next, select features for downstream integration, and run PCA on each object in the list, which is required for running the alternative reciprocal PCA workflow.
features <- SelectIntegrationFeatures(object.list = MergedList)
MergedList <- lapply(X = MergedList, FUN = function(x) {
x <- ScaleData(x, features = features, verbose = FALSE)
x <- RunPCA(x, features = features, verbose = FALSE)
        })

anchors <- FindIntegrationAnchors(object.list = MergedList, reference = c(1), reduction = "rpca",dims = 1:50)
List.integrated <- IntegrateData(anchorset = anchors, dims = 1:50)
```

```
List.integrated <- ScaleData(List.integrated, verbose = FALSE)
List.integrated <- RunPCA(List.integrated, verbose = FALSE)
List.integrated = RunTSNE(List.integrated, dims = 1:12)
List.integrated <- RunUMAP(List.integrated, dims = 1:50)
List.integrated <- FindNeighbors(List.integrated, reduction = "pca", dims = 1:30)
List.integrated<- FindClusters(List.integrated, resolution = 0.45)

#To avoid having to run all of that again, save an R data file after finding clusters.
saveRDS(List.integrated,file="Clustered.rds")

my_cols = c('Sample1'="firebrick4", 'Sample2'="firebrick", 'Sample3'="cadetblue3",
'Sample4'="cadetblue1")

#uMAP plots & markers

png("UMAP_samples.png", width = 12, height = 10, units = "in", res = 300)
DimPlot(List.integrated, group.by = "orig.ident", pt.size = 0.5, cols=my_cols)
dev.off()

png("UMAP_Clusters.png", width = 12, height = 10, units = "in", res = 300)
DimPlot(List.integrated, pt.size = 0.5, label=TRUE, label.size=5) + NoLegend()
dev.off()

md <- List.integrated@meta.data %>% as.data.table
md[, .N, by = c("orig.ident", "seurat_clusters")]
```

```
CPC_set1=md <- md[, .N, by = c("orig.ident", "seurat_clusters")] %>% dcast(., orig.ident ~ seurat_clusters, value.var = "N")
write.csv(CPC_set1, file="CellsPerCluster.csv")

allmarkers <- FindAllMarkers(object = List.integrated, only.pos = T, min.pct = 0.25, thresh.use = 0.35)
write.csv(allmarkers, file="allmarkers.csv")

#Determination of cell type using featureplots (note this looks at relative expression of a gene across all cells)
#color scheme
#Neuron = blue
#Oligodendrocyte = purple
#Microglia = orange
#Ng2 = light green
#Astrocyte = pink
#Cholinergic = yellow
#Glutamatergic = green
#Grm8 = red
#Sst = red

#Neuron
pdf(file="Syt6_Neuron.pdf", width = 7, height = 7)
Syt6=FeaturePlot(List.integrated, c("Syt6"),cols=c("grey", "lightblue","blue"))
Syt6+ggtitle("Syt6")
dev.off()
pdf(file="Snhg11_Neuron.pdf", width = 7, height = 7)
```

```
Snhg11=FeaturePlot(List.integrated, c("Snhg11"),cols=c("grey", "lightblue","blue"))
Snhg11+ggtitle("Snhg11")
dev.off()

#Oligodendrocyte
pdf(file="Mog_Oligodendrocyte.pdf", width = 7, height = 7)
Mog=FeaturePlot(List.integrated, c("Mog"), cols=c("grey",
"mediumpurple1","mediumpurple4"))
Mog+ggtitle("Mog")
dev.off()
pdf(file="Sox10_Oligodendrocyte.pdf", width = 7, height = 7)
Sox10=FeaturePlot(List.integrated, c("Sox10"),cols=c("grey",
"mediumpurple1","mediumpurple4"))
Sox10+ggtitle("Sox10")
dev.off()

#Microglia
pdf(file="Cx3cr1_Microglia.pdf", width = 7, height = 7)
Cx3cr1=FeaturePlot(List.integrated, c("Cx3cr1"),cols=c("grey", "orange","darkorange3"))
Cx3cr1+ggtitle("Cx3cr1")
dev.off()
pdf(file="Csf1r_Microglia.pdf", width = 7, height = 7)
Csf1r=FeaturePlot(List.integrated, c("Csf1r"),cols=c("grey", "orange","darkorange3"))
Csf1r+ggtitle("Csfr1")
dev.off()

#Astrocyte
```

```
pdf(file="Gfap_Astrocyte.pdf", width = 7, height = 7)
Gfap=FeaturePlot(List.integrated, c("Gfap"),cols=c("grey", "lightpink","maroon2"))
Gfap+ggtitle("Gfap")
dev.off()
pdf(file="Aqp4_Astrocyte.pdf", width = 7, height = 7)
Aqp4=FeaturePlot(List.integrated, c("Aqp4"),cols=c("grey", "lightpink","maroon2"))
Aqp4+ggtitle("Aqp4")
dev.off()

#Ng2
pdf(file="Pdgfra_Ng2.pdf", width = 7, height = 7)
Pdgfra=FeaturePlot(List.integrated, c("Pdgfra"),cols=c("grey", "chartreuse","chartreuse4"))
Pdgfra+ggtitle("Pdgfra")
dev.off()
pdf(file="Cspg4_Ng2.pdf", width = 7, height = 7)
Cspg4=FeaturePlot(List.integrated, c("Cspg4"),cols=c("grey","chartreuse","chartreuse4"))
Cspg4+ggtitle("Cspg4")
dev.off()

#Cholinergic
pdf(file="Ache_Cholinergic.pdf", width = 7, height = 7)
Ache=FeaturePlot(List.integrated, c("Ache"),cols=c("grey", "goldenrod1","goldenrod3"))
Ache+ggtitle("Ache")
dev.off()
pdf(file="Chat_Cholinergic.pdf", width = 7, height = 7)
Chat=FeaturePlot(List.integrated, c("Chat"),cols=c("grey","goldenrod1","goldenrod3"))
Chat+ggtitle("Chat")
```

```
dev.off()

#Glutamatergic
pdf(file="Slc17a7_Glutamatergic.pdf", width = 7, height = 7)
Slc17a7=FeaturePlot(List.integrated, c("Slc17a7"),cols=c("grey", "lightgreen","green"))
Slc17a7+ggtitle("Slc17a7")
dev.off()

#Grm8
pdf(file="Grm8.pdf", width = 7, height = 7)
Grm8=FeaturePlot(List.integrated, c("Grm8"),cols=c("grey", "red","firebrick4"))
Grm8+ggtitle("Grm8")
dev.off()

#Sst
pdf(file="Sst.pdf", width = 7, height = 7)
Sst=FeaturePlot(List.integrated, c("Sst"),cols=c("grey", "red","firebrick4"))
Sst+ggtitle("Sst")
dev.off()

#Markers by cluster where [#] can be replaced with the number of the cluster you are interested in
cluster[#].markers <- FindMarkers(List.integrated, ident.1 = [#], logfc.threshold = 0.25, test.use =
"roc")
write.csv(cluster[#].markers, file='cluster[#].markers.csv')

#Make list of markergenes for dotplots
```

```
markergenes=c("Tpx2","Cenpf","Ache","Chat","Pdyn","Ebf1","Drd2","Penk","Drd3","Grm8","
Elavl2","Kit","Sst","Slc17a7","Mbp","Hapln2","Gja1","Pdgfra","Arhgap15","Rgs5","Ppp1r1b","
Foxp2","Bcl11b","Gad1","Syt1")

#Make dotplot for all clusters before labelling
pdf("Dotplot_Clusters.pdf", width=12,height=10)
DotPlot(List.integrated, assay= "RNA", markergenes) + theme(axis.text.x = element_text(angle =
90, vjust = 0.5, hjust=1)) + ggtitle("Marker Genes")
dev.off()

#Annotation Colors for Heatmaps
annotation_colors <- list("Treatments"= c("Fentanyl" = "blue", "Saline" = "red"),"orig.ident" =
c('Sample1'="firebrick4", 'Sample2'="firebrick", 'Sample3'="cadetblue3",
'Sample4'="cadetblue1"))

#rename each cluster by creating a list and then swapping making that list a set of identities for the
seurat clusters, in this example cluster 0 is Drd2-MSN, cluster 1 is Drd1-MSN, cluster 3 is
Astrocytes, etc - this is based on your data from the featureplots and dotplot above#

new.cluster.ids=c("Drd2-MSN1","Drd1-MSN","Drd3-MSN","Astrocytes","Grm8-
MSN","Grm8-MSN","Oligodendrocytes","Ng2","Glutamatergic","Drd3-
MSN","Astro/Drd2","Oligo/Drd3","Sst-int","Drd2-MSN2","Microglia","Pvalb-int","Drd2-
MSN1","Astrocytes","Mitosis","Oligo/Micro","Oligodendrocytes",
"Astrocytes","Cholinergic","Ng2")
names(new.cluster.ids)=levels(List.integrated)
List.integrated=RenameIdents(List.integrated,new.cluster.ids)
```

```
#It takes some time to get to this point, so in the interest of not having to run all of the above again we recommend saving an R data file for later use. saveRDS(List.integrated,file="CellTypes.rds")

md <- List.integrated@meta.data %>% as.data.table
md[["new.clusters"]]=List.integrated@active.ident
md[, .N, by = c("orig.ident", "new.clusters")]
CPC_set2=md <- md[, .N, by = c("orig.ident", "new.clusters")] %>% dcast(., orig.ident ~ new.clusters, value.var = "N")
write.csv(CPC_set2, file="CellsPerType.csv")

markergenes2=c("Tpx2","Cenpf","Kit191","Stk331","Ache","Chat","Pdyn","Ebf1","Drd2","Penk","Drd3","Grm8","Elavl2","Kit","Sst","Slc17a7","Mbp","Hapln2","Gja1","Pdgfra","Arhgap15","Rgs5","Ppp1r1b","Foxp2","Bcl11b","Gad1","Syt1")

pdf("Dotplot_CellType.pdf", width=12,height=10)
DotPlot(List.integrated, assay= "RNA", markergenes2) + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) + ggtitle("Marker Genes and Cell Types")
dev.off()

png("~/Documents/Cunningham/SingleCell/Aug21/UMAP_celltype_set2.png", width = 12, height = 10, units = "in", res = 300)
DimPlot(List.integrated, pt.size = 0.5, label=TRUE, label.size=5) + NoLegend()
dev.off()

#Example of analysis for a single cell type
Microglia=subset(List.integrated,idents="Microglia")
DimPlot(Microglia, label=TRUE)
```

```
DimPlot(Microglia,group.by = "orig.ident", label=TRUE, cols=my_cols, label.size=5) + NoLegend()
Microglia=ScaleData(Microglia, verbose=FALSE, assay="RNA")
Microglia_Fentanyl=subset(x = Microglia, subset= (orig.ident == "Sample3" | orig.ident == "Sample4"))
Microglia_Saline=subset(x = Microglia, subset= (orig.ident == "Sample1" | orig.ident == "Sample2" ))
MF_Label="Fentanyl"
names(MF_Label)=levels(Microglia_Fentanyl)
Microglia_Fentanyl=RenameIdents(Microglia_Fentanyl,MF_Label)
MS_Label="Saline"
names(MS_Label)=levels(Microglia_Saline)
Microglia_Saline=RenameIdents(Microglia_Saline,MS_Label)
Microglia.combined <- merge(Microglia_Fentanyl, y = Microglia_Saline)
Microglia.combined <- ScaleData(Microglia.combined, verbose=FALSE, assay="RNA")
Microglia.combined = NormalizeData(Microglia.combined, assay="RNA")
Microglia.combined <- FindVariableFeatures(Microglia.combined, selection.method = "vst", nfeatures = 2000, assay="RNA")
Microglia.combined <- RunPCA(Microglia.combined, npcs = 12, verbose = FALSE, assay="RNA")
MicrogliaMarkersbyTreament=FindMarkers(Microglia.combined, ident.1="Fentanyl", ident.2="Saline")
write.csv(MicrogliaMarkersbyTreament,"MicrogliaMarkersbyTreatment.csv")

Microglia.combined=FindNeighbors(Microglia.combined,dims=1:10)
Microglia.markers=FindAllMarkers(Microglia.combined,min.pct=0.25,logfc.threshold = 0.35)
top10y=Microglia.markers %>% group_by(cluster) %>% top_n(n=10,wt=avg_log2FC)
```

```
DoHeatmap(Microglia.combined,features=top10y$gene,slot="data", group.colors = c("blue", "red"))
DoHeatmap(Microglia.combined,features=top10y$gene,group.by="orig.ident",slot="data", group.colors = c("blue", "red"))

genes=top10y$gene

# Note that the number of Fentanyl or Saline rows for the Treatments column is based on the subset numbers (Microglia_Fentanyl and Microglia_Saline, so this will change with each data set)#

TreatmentMeta=Microglia.combined@meta.data
Treatments=c(rep("Fentanyl",398), rep("Saline",402))
TreatmentMeta["Treatments"]=Treatments
TreatmentMetaTrim= subset(TreatmentMeta, select =c ("Treatments"))
Microglia.combined = AddMetaData(Microglia.combined, TreatmentMetaTrim)

pdf("Microglia_Heatmap.pdf", width = 12, height = 10)
dittoHeatmap(Microglia.combined,genes, annot.by = c("Treatments", "orig.ident"), cluster_rows=FALSE, annotation_colors=annotation_colors, main="Microglia")
dev.off()

pdf("Microglia_Heatmap2.pdf", width = 12, height = 10)
dittoHeatmap(Microglia.combined,genes, annot.by = "Treatments", cluster_rows=FALSE, cluster_cols=TRUE, annotation_colors=annotation_colors, main="Microglia")
dev.off()

pdf("Microglia_dotplot.pdf", width=12,height=10)
```

```
DotPlot(Microglia.combined, assay= "RNA", group.by = "orig.ident", top10y$gene) + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) + ggtitle("Treatment Markers in Microglia")
dev.off()

#To investigate the expression of covid markers featureplots like shown above can be used in addition to the dotplot below#

covidmarkers=c("Tmprss2","Ace2","Ctsb","Ctsl","Adam17","Adam10","Dpp4")

pdf("Dotplot_CellType_Covid.pdf", width=12,height=10)
DotPlot(List.integrated, assay= "RNA", covid markers) + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) + ggtitle("COVID-19 Genes and Cell Types")
dev.off()
```

Appendix G

Nucleic Acids Research, 2022 1
https://doi.org/10.1093/nar/gkab1259

Covariation of viral recombination with single nucleotide variants during virus evolution revealed by CoVaMa

Shiyi Wang [1,2,3], Stephanea L. Sotcheff[4], Christian M. Gallardo [1,2], Elizabeth Jaworski[4], Bruce E. Torbett [1,2,3,5,*] and Andrew L. Routh [4,6,7,*]

[1]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA, USA, [2]Center for Immunity and Immunotherapies, Seattle Children's Research Institute, Seattle, WA, USA, [3]Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA, USA, [4]Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, TX, USA, [5]Department of Pediatrics, University of Washington School of Medicine, Seattle, WA, USA, [6]Institute for Human Infections and Immunity, University of Texas Medical Branch, Galveston, TX, USA and [7]Sealy Center for Structural Biology and Molecular Biophysics, University of Texas Medical Branch, Galveston, TX, USA

Received September 14, 2021; Revised November 29, 2021; Editorial Decision November 30, 2021; Accepted December 09, 2021

ABSTRACT

Adaptation of viruses to their environments occurs through the acquisition of both novel single-nucleotide variants (SNV) and recombination events including insertions, deletions, and duplications. The co-occurrence of SNVs in individual viral genomes during their evolution has been well-described. However, unlike covariation of SNVs, studying the correlation between recombination events with each other or with SNVs has been hampered by their inherent genetic complexity and a lack of bioinformatic tools. Here, we expanded our previously reported CoVaMa pipeline (v0.1) to measure linkage disequilibrium between recombination events and SNVs within both short-read and long-read sequencing datasets. We demonstrate this approach using long-read nanopore sequencing data acquired from Flock House virus (FHV) serially passaged *in vitro*. We found SNVs that were either correlated or anti-correlated with large genomic deletions generated by nonhomologous recombination that give rise to Defective-RNAs. We also analyzed NGS data from longitudinal HIV samples derived from a patient undergoing antiretroviral therapy who proceeded to virological failure. We found correlations between insertions in the $p6^{Gag}$ and mutations in Gag cleavage sites. This report confirms previous findings and provides insights on novel associations between SNVs and specific recombination events within the viral genome and their role in viral evolution.

INTRODUCTION

Recombination within viruses, and particularly RNA viruses, is a powerful driving force behind their evolution and adaptation (1). Recombination may result in manifold outcomes, including the reshuffling of advantageous or deleterious mutations among homologous viral genomes, insertions or duplications of host or viral genomic segments, or deletion of genomic segments to generate defective viral genomes (DVGs). These genetic changes can have a dramatic impact upon viral evolution, fitness, viral intra-host diversity, and the development of resistance to antivirals (2–5). Due to the dense and compact nature of viral genomes, adaptive mutations rarely occur in isolation but rather mutations are often correlated with one another (6–8). Characterizing how or whether individual adaptions are correlated is essential to understand the mechanisms whereby viral strains emerge and adapt to their environment, for example, in response to anti-viral therapy or immune pressure.

A number of approaches have been documented to characterize the emergence of correlated viral adaptations (9). Classical approaches use consensus-level viral genomic data derived from multiple individuals/hosts deposited in curated databases such as GISAID, Stanford HIVdb, and others (10–12). More recently, approaches have focused on extracting evidence for mutational covariation in next-generation sequencing (NGS) data that reports on virus intra-host genetic diversity. Broadly, these methods either

*To whom correspondence should be addressed. Tel: +1 409 772 3663; Email: alrouth@utmb.edu
Correspondence may also be addressed to Bruce E. Torbett. Tel: +1 206 884 1140; Email: betorbet@uw.edu

directly assess the frequency of detected mutations found co-occurring within NGS reads, or infer correlated mutations using probabilistic or mathematical models (6,13–17). However, there are no current approaches aimed at determining the correlation of recombination events, such as insertions, deletions, or duplications, with each other or with single nucleotide variants (SNV). Characterizing these correlations may be necessary for understanding why certain recombination events might be selected for and their role in viral evolution.

We recently reported a method that measures linkage disequilibrium of SNVs within viral NGS datasets (16). This approach, called CoVaMa (co-variation mapper), built a large matrix of contingency tables corresponding to every possible pair-wise interaction of correlated nucleotides within a viral genome. Each contingency table was a 4 × 4 matrix with columns and rows corresponding to the bases 'ATGC' for each nucleotide position and was populated by extracting every possible pair of co-mapping nucleotides within individual reads. From each 4 × 4 table, all possible 2 × 2 tables were extracted to measure linkage disequilibrium (LD), thus reporting evidence of mutational covariation. This basic process is a non-heuristic and brute-force approach implemented in a series of python scripts. We demonstrated the utility of this approach by identifying correlated mutations within the genomic RNAs of Flock House virus (FHV) and within anti-viral inhibitor-resistant HIV patient samples.

However, this approach retained many limitations, including the requirement for a gapless alignment and the inability to consider the correlation of other types of mutations besides SNVs. As we and others have shown (4,18–20), recombination events and InDels comprise an important component of the viral intra-host diversity. However, while there is evidence of the co-evolution of SNVs and other types of mutations, high-throughput tools to assess the epistatic linkage between recombination events and/or SNVs remain lacking. The requirement for a gapless alignment also prevented the analysis of long-error prone read data, such as nanopore or PacBio data. These data types have the distinct advantage in that they can resolve epistatic associations over much greater distances than Illumina short-read data and can even resolve full-length viral genomes direct from RNA.

Here, we expanded the capabilities of CoVaMa to measure linkage disequilibrium between recombination events found within deep-sequencing reads and SNVs. Recombination events, such as insertions and deletions, can be extracted from gapped-alignments using *HISAT2* (21) and *Bowtie2* (22), or using other tools designed for mapping recombination in virus genomes such as *ViReMa* (23). Moreover, our improved algorithm can accept long-reads generated from Nanopore sequencing platforms. This approach used a similar principle employed by the original version of CoVaMa by generating large matrices of contingency tables corresponding to every pair-wise interaction of each nucleotide position with each detected recombination event in the NGS dataset. We developed a set of criteria that determined when a sequence read excluded the possibility of mapping across a recombination event. This allowed the populating of 4 × 2 contingency tables for each nucleotide (ATGC) at each genomic position with either the confirmed absence or presence of a recombination event from which the degree of covariation or Linkage Disequilibrium value was calculated.

We demonstrate the utility of this approach by re-analyzing long-read nanopore data that we generated in a recent longitudinal study of Flock House virus (FHV) evolution in cell-culture (24) and short-read Illumina data from a study of clinical HIV patient samples collected during the development of antiretroviral therapy resistance. In the FHV dataset, we identified multiple SNVs that were either correlated or anti-correlated with large deletion events that constituted Defective-RNAs (D-RNAs). We hypothesize that these mutations were adaptive mutations that either allowed the heightened replication of D-RNAs, or that allowed the wild-type full-length genome to escape attenuation by D-RNAs. From the HIV samples, we identified insertions in the PTAP region of the $p6^{Gag}$ that correlated with mutations found in Gag cleavage sites. The location of these linked SNVs and InDels proximal to Gag cleavage sites suggested a role for these adaptions to support drug-resistance development. Overall, CoVaMa provides a powerful tool to characterize the molecular details of viral adaptation and the impact of RNA recombination upon virus evolution. The CoVaMa (v0.7) script is publicly available at https://sourceforge.net/projects/covama/.

MATERIALS AND METHODS

Long-read data from serial passaged FHV analyzed using CoVaMa

We previously described a serial passaging experiment of Flock House virus (FHV) followed by the analysis of viral genetic changes by the parallel use of short-read Illumina and long-read nanopore sequencing (24). Briefly, pMT vectors containing cDNA of each FHV genomic RNA were transfected into S2 cell in culture and expression of genomic RNA was induced with copper sulphate. After 3 days, 10 ml of supernatant was retained. 1ml of supernatant was passaged directly onto fresh S2 cells in culture, and FHV virions were purified from the remaining supernatant using sucrose cushions and sucrose gradients as described. This process was repeated for a total of nine passages. Encapsidated RNA was extracted from each stock of purified virions and reverse transcribed using sequence-specific primers cognate to the 3′ ends of FHV RNA1 and FHV RNA2. Full-length cDNAs were then PCR amplified in 19 cycles with NEB Phusion using primers targeting the 5′ and 3′ end of each genomic segment. Final PCR amplicons were used as input for Oxford Nanopore Amplicon sequencing protocol by ligating on native barcodes and then the ONT adaptor. Pooled libraries were sequenced on an ONT MinION MkIB with R9 (2017) flowcells. Data was demultiplexed and basecalled using the Metrichor software and using *poretools* (25). Datasets are publicly available in NCBI SRA under the accession code SRP094723. Basecalled data were mapped to the FHV Genomic RNAs (FHV RNA1: NC_004146, 3107 nts and FHV RNA2: NC_00414, 1400 nts) using *BBMap*. SAM files of the aligned data were passed to the CoVaMa (Ver 0.7) pipeline using the command lines shown in Table 1.

Table 1. Command lines used to analyze nanopore sequencing reads acquired from each passage of FHV

(1) Making matrices containing contingency table for each association using CoVaMa_Make_Matrices script: Output_Tag, the name for the output pickle file [Output_Tag].Total_Matrices.py.pi. Data_directory, the folder contains the FHV reference sequence in FASTA file format and the aligned sequences in SAM file format. Contingency tables populated by over 100 reads and the mutant frequency higher than 0.05 were generated and passed to the output pickle file.
python2 CoVaMa_Make_Matrices.py [Output_Tag] Data_directory/FHV_Genome_corrected.txt –Mode2 All –SAM1 Data_directory/FHV_mapping.sam –PileUp_Fraction 0.05 –Min_Fusion_Coverage 10 NT
(2) Calculating linkage disequilibrium values for each contingency table using CoVaMa_Analyse_Matrices script: [Output_Tag].Total_Matrices.py.pi generated by CoVaMa_Make_Matrices script was used as the input. Linkage disequilibrium information was stored in the output file in TXT file format. A minimum coverage of over 10 pairs of associated nucleotides and recombination events was required for the contingency table to be analyzed. The linkage disequilibrium values (LD) and R square values for each association were normalized by the number of reads populating that contingency table.
python2 CoVaMa_Analyse_Matrices.py [Output_Tag].Total_Matrices.py.pi CoVaMa_output.txt –Min_Coverage 10 –Min_Fusion_Coverage 10 -OutArray -Weighted NT

Table 2. Command lines used to analyze NGS sequencing reads acquired from each longitudinal HIV sample

(1) Making matrices containing contingency table for each association using CoVaMa_Make_Matrices script: Output_Tag, the name for the output pickle file [Output_Tag].Total_Matrices.py.pi. Data_directory, the folder contains the HIV reference sequence for each sample in FASTA file format and the aligned sequences in SAM file format. Contingency tables populated by over 100 reads and the mutant frequency higher than 0.01 were generated and passed to the output pickle file.
a. For the 15nt insertion event at nt 2158, 15 nucleotides were used to exclude negative recombination event. The analysis region started at nt 1958 and ended at nt 2358.
b. For the 6nt insertion event at nt 1149, 6 nucleotides were used to exclude negative recombination event. The analysis region started at nt 949 and ended at nt 1349.
python2 CoVaMa_Make_Matrices.py [Output_Tag] Data_directory/HIV_reference_seq.txt –Mode2 Recs –SAM1 Data_directory/HIV_mapping.sam –PileUp_Fraction 0.01 –Rec_Exclusion 15 –NtStart 1958 –NtFinish 2358 NT
(2) Calculating linkage disequilibrium values for each contingency table using CoVaMa_Analyse_Matrices script: [Output_Tag].Total_Matrices.py.pi generated by CoVaMa_Make_Matrices script was used as the input. Linkage disequilibrium information was stored in the output file in TXT file format. A minimum coverage of over 100 pairs of associated nucleotides and recombination events was required for the contingency table to be analyzed. The linkage disequilibrium values (LD) and R square values for each association were normalized by the number of reads populating that contingency table.
python2 CoVaMa_Analyse_Matrices.py [Output_Tag].Total_Matrices.py.pi CoVaMa_output.txt –Min_Coverage 100 –Min_Fusion_Coverage 100 -OutArray -Weighted NT

Short-read data from longitudinal HIV patient samples analyzed using CoVaMa

We previously described an RT-PCR approach to amplify the *gag-pol* regions of HIV from 93 clinical specimens for Next-Generation Sequencing (6,26). Briefly, two overlapping cDNA amplicons were RT-PCR amplified using two pairs of primers targeting *gag-pol* (F1: NT 754, R1: 1736; F2: 1569, R2: 2589). Full-length cDNA amplicons were sheared to approximately 175bp and submitted for paired-end sequencing (2 × 150 bp) on an Illumina HiSeq. We chose five longitudinal samples derived from a single patient over six years. Paired-end reads were merged using *BBMerge* and mapped to the HIV genome (HXB2: K03455.1, 9719 bp) using ViReMa v0.21 (23) with default settings plus the following parameters (*–X 3 –MicroInDel 5 –BackSplice_limit 25*). As the cDNA amplicon strategy for Illumina sequencing was not directional, we only used the reads mapping to the positive sense viral genome for further analysis. These SAM files were passed to CoVaMa (Ver 0.7) as described in Table 2.

Statistical analysis in CoVaMa outcomes

Linkage disequilibrium was measured for every pair-wise interaction of each nucleotide position with each detected recombination event within the viral genome using the parameters set in the command line in each CoVaMa analysis (Tables 1 and 2). To distinguish the significant associations from the background noise, the three-sigma rule for comparing a single data point to a very large distribution of other data points was applied (27). CoVaMa automatically generated the standard deviation and mean of all LD values, based on which the three-sigma could be calculated. Associations passing the three-sigma in each CoVaMa outcome were considered significant.

RESULTS

Description of algorithm

CoVaMa requires alignment information from standardized SAM files generated by any typical read mapper such as *bowtie2*, *hisat2* for short reads or *bwa* and *minimap2* for long reads. Specialized virus-focused read alignment software such as ViReMa can also be used if they output an alignment in SAM format. CoVaMa also requires information on the virus genome that was used in the alignment in FASTA format. The FASTA file provides information on the length of the viral genome and the number of genome segments. Using this information, CoVaMa builds three types of data matrices composed of numerous contingency tables. The largest is a nucleotide-vs-nucleotide matrix enumerating the pairwise nucleotide identities of every possible combination of mapped nucleotides found in every sequence read. The matrix comprises $N*N$ potential contingency tables for each genome segment, where N is the length of the viral genome. Each contingency table is a 4 × 4 matrix where each row corresponds to the number of mapped A, T, G or Cs in each read at the first genome coordinate, and each column corresponds to the number of mapped A,

T, G, or Cs in each read at the second read coordinate, as we previously described (16).

In this manuscript, we introduced a novel feature in CoVaMa to correlate the presence or absence of recombination events including insertions, duplications, and deletions. We built matrices containing 4 × 2 Nucleotide-vs-Recombination contingency tables and 2 × 2 Recombination-vs-Recombination contingency tables. CoVaMa scrutinizes the provided SAM file for evidence of any recombination event (reported as an 'N', 'D' or 'I' in the CIGAR string) and adds the recombination event to the 4 × 2 and 2 × 2 matrices if it is mapped with at least a user-defined number of reads (10 by default). Similar to the nucleotide-vs-nucleotide tables, the columns of each 4 × 2 or 2 × 2 contingency table correspond to either the presence or the *confirmed* absence of a recombination event in each mapped read. These matrices and contingency tables are depicted in Figure 1A.

The contingency tables within each matrix are populated using all the individual reads from the provided SAM file using the '*CoVaMa_Make_Matrices.py*' python script. Linkage disequilibrium analyses rely on the measurement of both the presence of the associated alleles and their absence. For the 4 × 4 nucleotide matrices, the presence of one nucleotide at a locus automatically infers the absence of the other three. However, to measure linkage disequilibrium between recombination events, or between recombination events and nucleotides, we must build a pipeline that determines whether a mapped read can confirm that a recombination event is absent. To this end, we devised the following set of processes and parameters (Figure 1D):

1) Deletions/splicing: To confirm the absence of a deletion event, a read must map inside the putative deletion site with at least a minimum number of nucleotides defined in the command-line (–Rec_Exclusion X; default is 10 nts). These mapped nucleotides can map anywhere within the deletion and/or overlap with the recombination junction. With a large number of nucleotides required, the exclusion of the recombination event has high confidence. If too few nucleotides are required, ambiguity may arise due to sequence similarity between deleted nucleotides and the nucleotides upstream of the recombination event. For instance, if only one nucleotide is required to map inside the recombination event, a negative mapping will be scored erroneously for one out of every four recombination events as the first deleted nucleotide has a one in four chance of also being the same as the nucleotide upstream of the recombination event. Care also must be taken in repetitive regions in case the nucleotides preceding the recombination 5′ site are close or identical to those preceding the recombination 3′ site. In cases like these, a large 'X' value is required.
2) Micro-deletions: If a deletion event is very small (i.e. it is a micro-deletion smaller than the –MicroInDel_Length Y parameter specified in the command line) then it is not possible for –InDel_Exclusion X nts (default is 5) to map inside the recombination event. In this case, the aligned read should map over the possible micro-deletion event and have X number of mapped nucleotides on each side of the event to confirm the absence of the putative deletion.
3) Insertions: A similar strategy is required to confirm the absence of a putative insertion event. Insertion events can either correspond to inserted nucleotides, insertions of fragments of host or other viral genes, or small duplications. To confirm the absence of an insertion event, an aligned read should have an appropriate number of mapped nucleotides on each side of the insert site. The number of mapped nucleotides on each side is controlled by the –Rec_Exclusion parameter (default is 10) in the command line for large insertions or –Indel_Exclusion parameter (default is 5) for micro-insertions smaller than the –MicroInDel_Length parameter (default is 5). An inappropriate setting alters the number of events detected and might influence the results.

Linkage disequilibrium calculation

Once contingency tables have been generated, evidence for covariation is determined using linkage disequilibrium (LD) by the same principle as we previously described, using the second python script called '*CoVaMa_Analyse_Matrices.py*'. From each 4 × 4, 4 × 2, and 2 × 2 contingency table, every possible 2 × 2 table is extracted. If the total number of reads for each 2 × 2 table exceeds a default value of x reads (which can be adjusted on the command-line), LD is calculated using the canonical formula: LD = (pAB * pab) – (pAb * paB) where 'A' and 'a' are the haplotypes for the nucleotide or recombination event at coordinate A, and 'B' and 'b' and the haplotypes for the nucleotide or recombination event at coordinate B, as depicted in Figure 1B. The range of LD is 0 to ±0.25, with higher values connoting disequilibrium and a value of 0 meaning there is no co-variation. In addition to the LD value, CoVaMa also calculates the R^2 value using the canonical formula: R^2 = (LD*LD)/ (pA * pa * pB * pb). The range of R^2 is 0 to 1, again with higher values connoting disequilibrium and a value of 0 meaning no covariation. CoVaMa also calculates the maximum possible LD that could have been obtained for each 2 × 2 contingency table given the frequencies of the variants at each coordinate. As a single 4 × 4 or 4 × 2 table can give rise to multiple possible LD reports if there is sufficient diversity in these contingency tables, as we described previously (16), the linkage disequilibrium values (LD) and R^2 values for each association are normalized by the proportion of reads populating the 2 × 2 contingency table from the entire 4 × 4 or 4 × 2 contingency table. This yields weighted LD values (wLD) and weighted R^2 values (wR^2). For a 2 × 2 table (recombination-vs-recombination), the LD or R2 will be unchanged after this normalization.

This information is reported in an output text file, as depicted in Figure 1C. This provides: (i) a description of the type of correlation being reported (e.g. nucleotide-vs-nucleotide, nucleotide-vs-recombination, or recombination-vs-recombination); (ii) the two events being tested for correlation (either the nucleotide position or the recombination event); (iii) the largest LD and R^2-value found the contingency table; (iv) and largest possible LD values that could have been found given the frequencies of each observed variants at each coordinate; (v) and a flat-

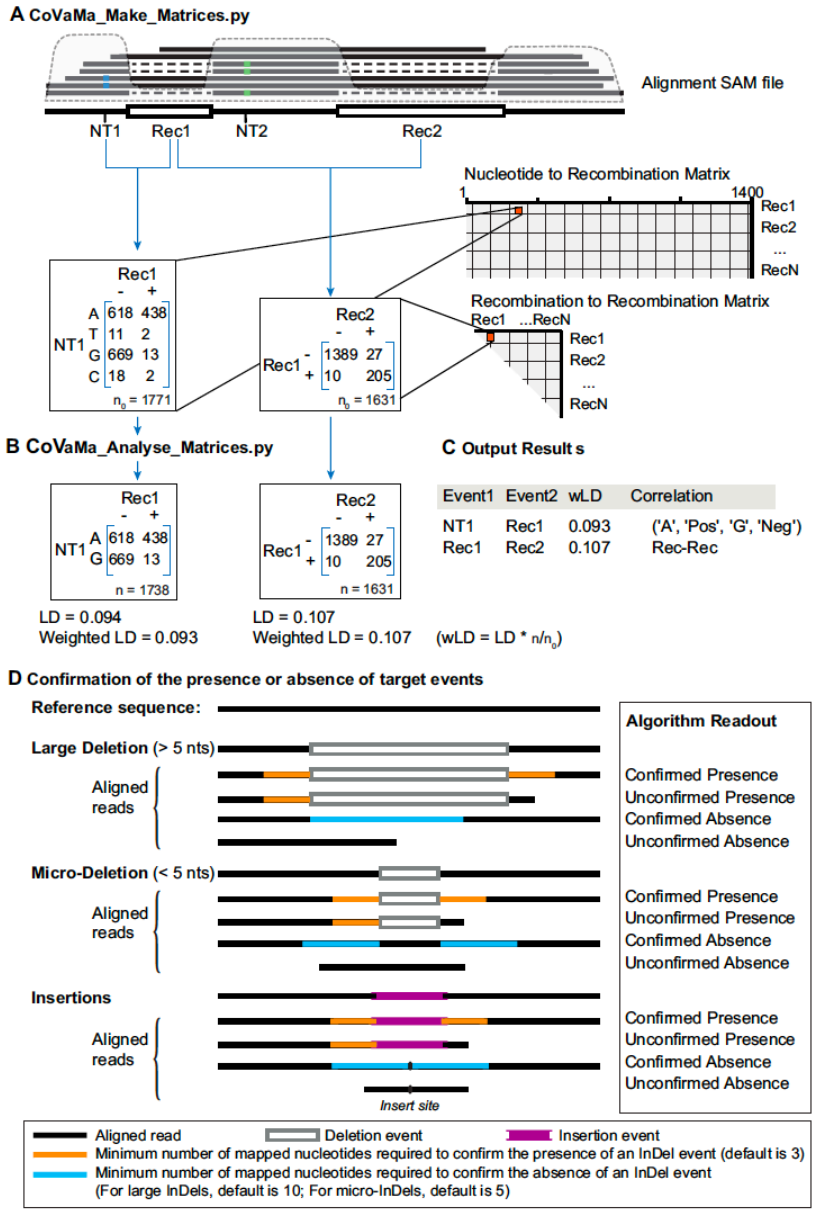


Figure 1. Schematic and flow chart of CoVaMa pipeline. (**A**) CoVaMa_Make_Matrices.py extracts information from each aligned read and generates large matrices containing 4 × 4 nucleotide-vs-nucleotide contingency tables, 4 × 2 nucleotide-vs-recombination contingency tables, and 2 × 2 recombination-vs-recombination contingency tables. In each contingency table, the columns correspond to either the mapped A, T, G and Cs, for nucleotides, or the presence and the confirmed absence for recombination events. Rec, Recombination; SNV, single-nucleotide variant. (**B**) CoVaMa_Analyse_Matrices.py analyses contingency tables from each matrix for evidence of linkage disequilibrium. From each 4 × 4, 4 × 2 and 2 × 2 contingency table, every possible 2 × 2 table populated by sufficient reads is extracted to calculate the LD value. The LD values and R^2 values are normalized by the proportion of reads populating the 2 × 2 contingency table. (**C**) An example of the CoVaMa output. (**D**) Schematic diagram showing the process to confirm the presence or the absence of an InDel event in aligned reads. To confirm the presence of the InDel event, three matched nucleotides are required at each end of that event in the aligned read. To confirm the absence of a large deletion event in a read, the minimum number of nucleotides required to map inside the putative deletion site is controlled by the *–Rec_Exclusion* (default is 10). To confirm the absence of a micro-deletion event in a read, the minimum number of mapped nucleotides required on each side of the putative event is controlled by *–InDel_Exclusion* (default is 5). To confirm the absence of an insertion event in a read, the minimum number of mapped nucleotides required on each side of the insert site is controlled by *–Rec_Exclusion* for large insertions or *–InDel_Exclusion* for micro-insertions. Insertions or deletions smaller than the *–MicroInDel_Length* value (default is 5) are considered as micro-insertions or micro-deletions.

tened 2D-array of the contingency table that was tested for LD.

Associations between recombination events and SNVs in defective FHV RNAs revealed by CoVaMa

Flock House virus (FHV) is an insect-specific small bipartite RNA virus and an ideal model system to study viral evolution and recombination (16,28). The viral genome consists of two segments, RNA1 (3.1 kb) and RNA2 (1.4 kb), which encode for the viral polymerase and viral capsid protein respectively (29). In a previous study (24), we serially passaged FHV in S2 Drosophila cells in culture to characterize the emergence, selection, and adaption of defective-RNAs (D-RNAs) *in vitro*, which arise through nonhomologous RNA recombination (30). We began with a clonally derived population of viral RNAs expressed from transfected pMT vectors, and blind-passaged the supernatant every three days onto fresh S2 cells for a total of nine passages. During this period, we extracted encapsidated RNA from purified virions and generated full-length cDNA copies of the FHV genomic segments using RT-PCR with primers designed for the first and last ∼25nts of each segment. Purified cDNA was prepared for long-read nanopore sequencing using the SQK-LSK107 2D kit to identify the emergence of insertions, deletions, and other RNA recombination events. We calculated an average error single-nucleotide mismatch rate of ∼7% in our 2D nanopore reads and we also robustly identified numerous recombination events including insertions and deletions. Interestingly, nanopore sequencing revealed that multiple deletions were most commonly found in individual reads, while reads containing only single-deletions were seldom seen. In that study, we postulated that the correlation of multiple deletion events within individual reads indicated a selective advantage for 'mature' D-RNAs over the 'immature' D-RNAs. The observation of these paired deletions was consistent with major defective RNA2 species that were previously characterized (24,31–33).

During passaging, we also noticed the emergence of multiple minorities and SNVs in the viral genome including A226G and G575A in the RNA2. However, the function of these mutations was not determined. A226G is a synonymous substitution, while G575A results in the Alanine to Threonine substitution at amino acid position 185 of the capsid protein. This is at the five-fold and quasi-three-fold symmetry axes of the T = 3 icosahedral virus particle (34) and might therefore interfere with the virus assembly (Supplementary Figure S1).

To determine whether these SNVs co-varied with the recombination events constituting D-RNA2 species, we passed the mapped FHV nanopore reads from each passage to our CoVaMa pipeline using the command-line parameters indicated in Table 1. This used the aligned sequence data to populate contingency tables corresponding to the association of all mapped recombination events (found in at least 10 reads) with other recombination events or with SNVs (with a mutant frequency higher than 5%). As most erroneous InDels generated by Nanopore sequencing were found to be shorter than 25nts for this dataset (24), only recombination events greater than 25nts in length were used for analysis. We found that recombination events were infrequent in the first two passages, but multiple recombination events emerged in Passage 3 (Figure 2A). Throughout passaging, diverse recombination events were observed to increase in frequency, reflecting the selective or replicative advantage of Defective RNA species. Interestingly, the most common recombination events were clustered into two regions with only small variance in the exact coordinates of the recombination junction: deletion events that excise nucleotides between nt 240 and nt 530 termed '*Group 1*' (such as 248^512 and 250^513), and deletion events that excised nucleotides between nt 720 and nt 1240 termed '*Group 2*' (such as 736^1219) (Figure 2B, Figure 2C). The fluctuation in their abundance over time could be due to the competition between the encapsulation of D-RNAs and the requirement of generating enough complete capsid proteins to form mature capsids (35). As expected, neither group removed the RNA2 packing motif or the RNA2 replication *cis*-acting motif (32,36).

CoVaMa detected strong associations between recombination events with each other and between recombination events and SNVs over passaging (Figure 3A, Figure 3B, Supplementary Table S1, Supplementary Figure S2, Supplementary Table S2), including between 248^512 and 736^1219, 250^513 and 736^1219, G272A or A226G and recombination events. The non-synonymous mutation G575A positively correlated with recombination events 250^513, 248^512 and 736^1219 (Figure 3B, Supplementary Figure S2B). This mutation was enriched over passaging along with the enrichment of D-RNA2, with a maintained significant LD between G575A and recombination events. This indicated that this SNV was predominantly found in the D-RNA2, but not full-length RNA2, suggesting a co-dependent evolution. In contrast to the G575A mutation, A226G mutation negatively correlated with recombination events 250^513, 248^512 and 736^1219 (Figure 3B). Consistent with these observations, CoVaMa also reported that A226G and G575A negatively correlated with one another (Figure 3B, Supplementary Table S3).

To confirm these results, we separated reads mapping to either the full-length RNA2 or D-RNA2 into two groups based on whether they mapped over the recombination event 736^1219, which was one of the major recombination events constituting D-RNA2. By enumerating the frequency of A226G and G575A in full-length RNA2 and D-RNA2, we found that A226G was enriched in full-length RNAs from 25% in Passage 4 to 50% in Passage 9, while the abundance of A226G in D-RNA2 remained low at about 2% in all passages (Figure 3C). The ratio of the abundance of this mutation in full-length RNA2 and D-RNA2 was 18.8 ± 6.0 (SD) over passaging (Supplementary Table S4). This SNV was therefore anti-correlated with the emergence of D-RNAs. The enrichment of A226G in the full-length RNA2 suggested a possible beneficial effect to the replication and/or packaging of the mutant full-length RNA2. In contrast, G575A was enriched in the D-RNA2 and relatively depleted in the full-length genomic RNA (Figure 3C). The ratio of the abundance of this mutation in D-RNA2 and full-length RNA2 was 16.2 ± 7.1 (SD) over passaging (Supplementary Table S5). This SNV was therefore posi-

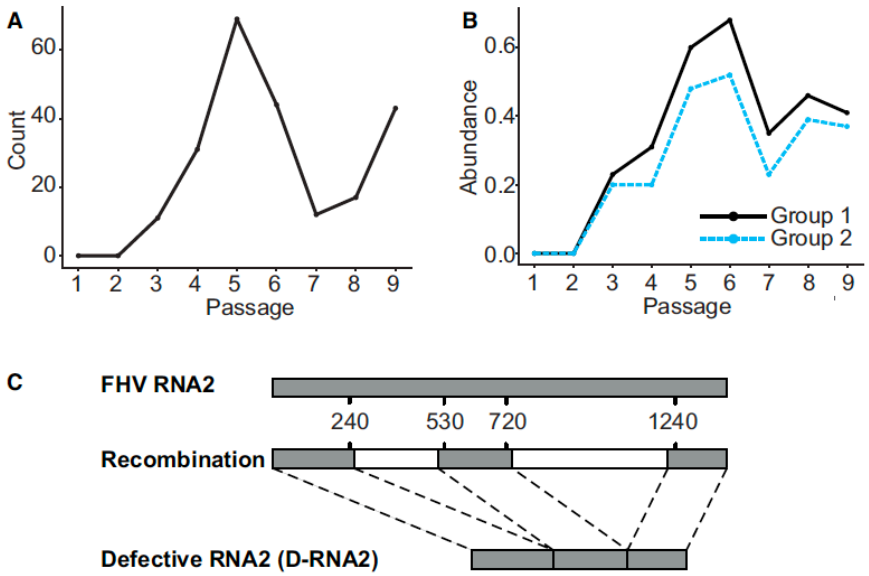


Figure 2. Recombination events detected in FHV RNA2 by CoVaMa. (**A**) The count of unique recombination events in FHV RNA2 in each passage. (**B**) The abundance of two groups of recombination events in the FHV RNA2 in each passage. (**C**) Schematic diagram of full-length FHV RNA2 and Defective RNA2 (D-RNA2).

tively correlated with the emergence of D-RNAs, suggesting a D-RNA-specific adaptation.

A226G in RNA2 is a synonymous substitution, thus we investigated its potential influence on the secondary structure of RNA2. We used Vienna RNA Website (37) to generate predicted RNA structures of both D-RNA2 and full-length RNA2, with and without A226G. The presence of A226G showed no significant influence on the adjacent RNA2 packaging motif (32), both in D-RNA2 and full-length RNA2 (Supplementary Figure S3). However, the A226G mutation was predicted to destabilize a region of secondary structure that formed a long-range interaction with residues 580–600, which in turn was predicted to alter additional long-range interactions even further downstream within the 50nts of the 3′ terminus. Interestingly, this region contains the cis-acting motif essential for RNA2 replication (38–40). This might indicate a possible deleterious effect of A226G in the replication of D-RNA2. Alternatively, the predicted structure of full-length RNA2 without the A226G mutation placed position 736 and position 1219 close to each other, between which one of the major recombination events in D-RNA2 was found. However, when there was an A226G mutation in RNA2, this distance was significantly increased, which might preclude the emergence of this recombination event (Supplementary Figure S4).

To validate the reported correlation of the non-synonymous substitution G575A in RNA2 with the recombination events found in D-RNA2, we engineered a pMT-FHVRNA2 expression vector with the G575A SNV using a reverse genetics approach that we have previously used extensively (41). However, we were unable to rescue this mutant virus. This suggests that this point mutation is a lethal mutation, rationalizing why it is not observed on the wild-type, full-length virus genome and only on the D-RNAs of FHV which are not constrained by a need to express FHV capsid.

Associations between insertion in p6^{gag} and mutations in *gag* cleavage sites revealed by CoVaMa

We reported on the covariation of amino acid and nucleotide variants within a large cohort of HIV-infected patients as part of the US Military HIV Natural History Longitudinal study using the previous version of CoVaMa (v0.1). There, we found evidence of correlated mutations in the HIV *gag* and *protease* regions as well as multiple correlated adaptations within *protease* itself, similar to previous reports (6,7,16,42,43). However, these studies were limited to assessing only SNVs, and did not report on the presence of insertions, deletion, or duplication events, which are common in the HIV genome and critical for viral evolution (44). Here, we used ViReMa (23) to detect and quantify unusual or unknown insertions, deletion, and duplication events *de novo*. This approach leverages the seed-based mapping algorithm of *bowtie* and is particularly well suited to the analysis of complex viral RNA recombination events that fail to follow strict, or characterized, rules.

Five longitudinal sera samples were collected from a patient who continuously failed two antiretroviral treatments (ARTs) over six years (6). At each time point, HIV genomic RNA was RT-PCR amplified as previously described and sequenced using NGS (26). The sequencing outputs were short reads 100nts long, which restricted the detection range of association to this length. We therefore merged paired reads using *BBMap*, extending the maximum detection range to 200 nucleotides. We analyzed the output NGS

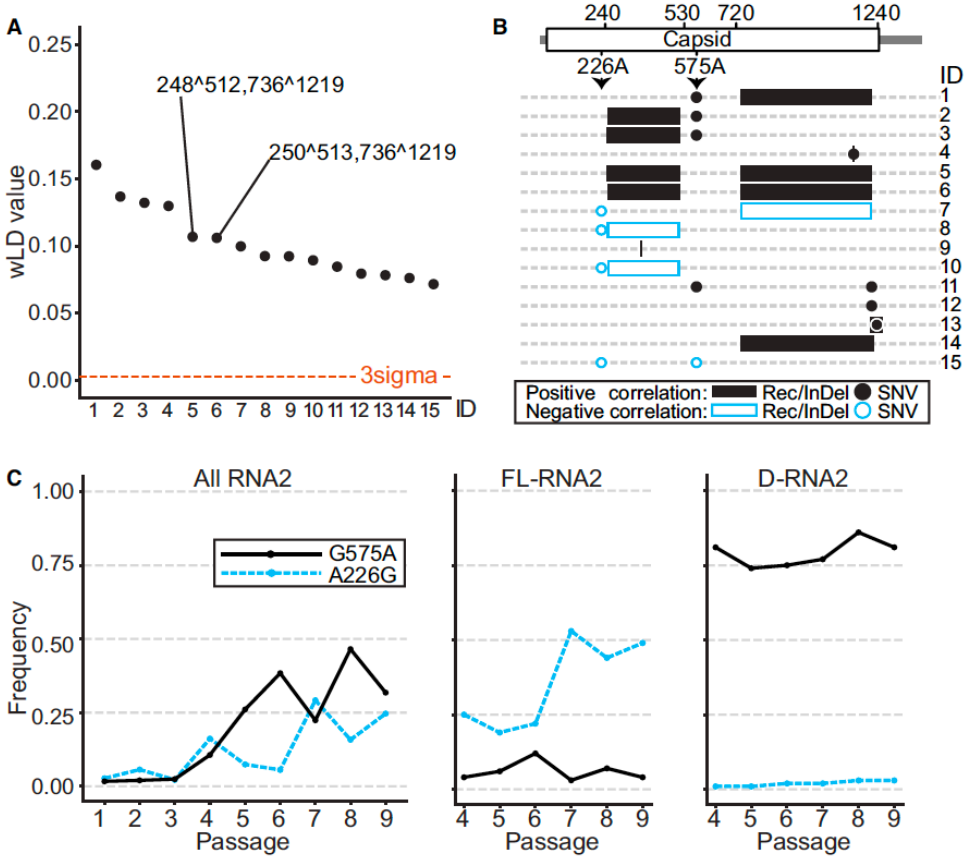


Figure 3. Associations between recombination events and SNVs in FHV RNA2 revealed by CoVaMa. The 15 associations with the highest wLD values in Passage 9 revealed by CoVaMa are labeled from 1 to 15. (**A**) The wLD values of the 15 associations are plotted from high to low. The associations between major recombination events are labeled in the plot. The three-sigma threshold is shown by the red dashed line. (**B**) This schematic diagram shows the distribution of the 15 associations on FHV RNA2. Positive associations are colored in black while negative associations are colored in blue. Recombination events and InDels are plotted using blocks and SNVs are plotted using dots. (**C**) The abundance of mutation A226G and G575A in all RNA2 reads (left), in the full-length RNA2 (middle), and in the D-RNA2 (right) in each passage.

data from these longitudinal samples by mapping the data to the HXB2-indexed consensus genome of each sample and used ViReMa to identify recombination events in addition to SNVs that arose during virological failure. The average coverage over HIV *gag* and part of the *protease* was 40466 overlapping reads, with first and third quantile values being 23 387 and 52 143 reads. With this depth, we could identify numerous minority variants (Supplementary Table S6) with at least 10 SNVs in each sample at a frequency of greater than 10%. In addition to SNVs, ViReMa reported a consistent insertion adding 'RPEPS' or 'RLEPS' in the P(T/S)AP region of the $p6^{gag}$, which also generated a 'RxEPS' duplication (x for P/L) right downstream of the P1/p6 cleavage site (Figure 4A, Table 3). Similar insertion or duplication events in the p6 region close to the p1/p6 cleavage site have been observed in other studies (45–47). Another consistent 6-nt insertion occurred close to the proteolysis sites between matrix and capsid protein. This insertion encoded an extra 'AA' between amino acid 120 and amino acid 121 (Figure 4A), which has also been observed in the HIV 1 group M subgroup B isolate ARV2/SF2 (48). Additionally, a 15-nt insertion in the MA/CA cleavage site was detected in Sample 2 and Sample 4.

The output SAM files from the ViReMa alignment were passed to CoVaMa to measure linkage disequilibrium. Significant associations were revealed between insertion events and SNVs. The C2146T mutation, which encodes the P453S substitution in the Gag P1/p6 cleavage site, emerged before sampling time 2, after which its abundance increased over time together with the enrichment of 'RxEPS' insertion in the $p6^{Gag}$ (Figure 4B). CoVaMa showed that this mutation positively correlated with 'RxEPS' insertion (Figure 4C, Supplementary Table S7), with a maintained elevated LD between them over time. Quantification of the abundance of P453S in reads with and without 'RxEPS' insertion showed that P453S was much more likely to be

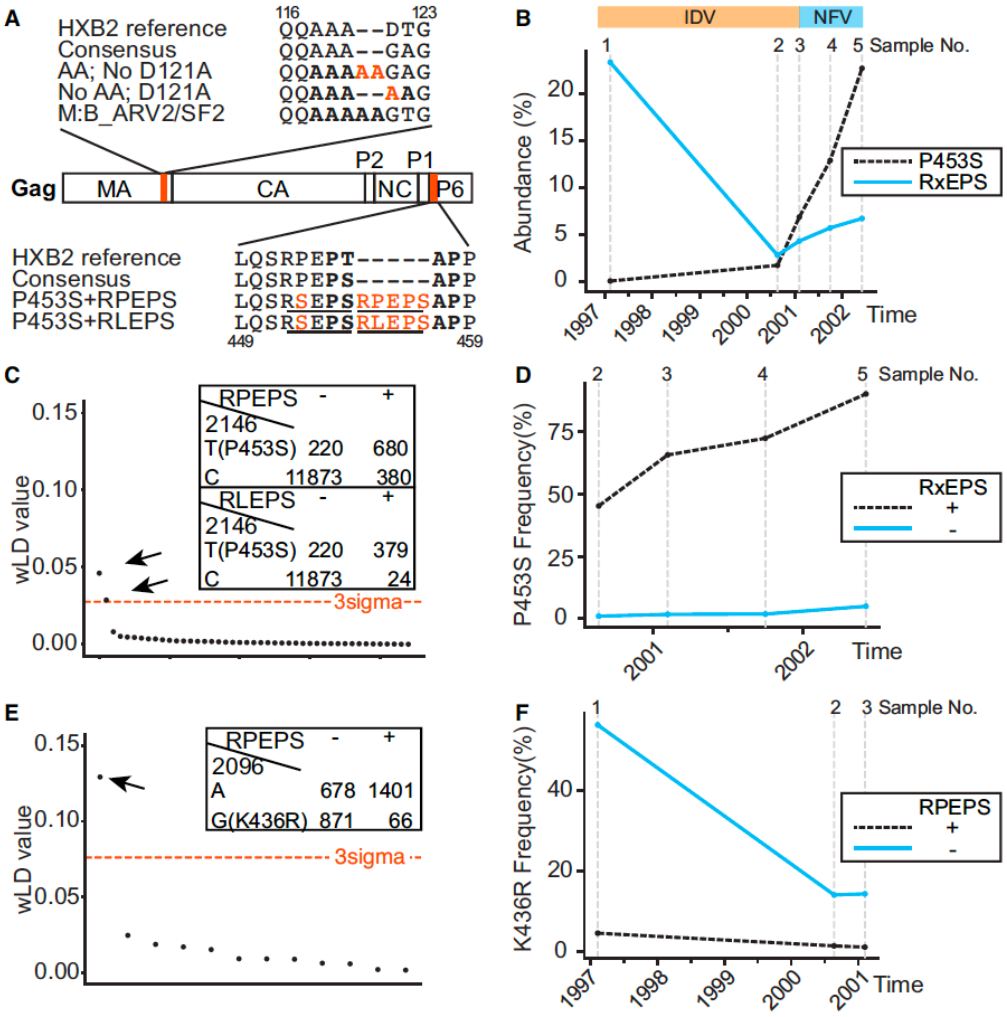


Figure 4. Associations between insertions in $p6^{Gag}$ and mutations in Gag cleavage sites revealed by CoVaMa. **(A)** Positions of two major insertions detected and their correlated SNVs are shown in the HIV Gag, highlighted in red. Gag domains are indicated (MA, CA, P2, NC, P1, $p6^{Gag}$). At the top are shown the HXB2 reference sequence, the consensus sequence of sample surrounding the 'AA' insertion between amino acid 120 and 121, the negative correlation between 'AA' and D121A, and the sequence for HIV 1 group M subgroup B isolate ARV2/SF2. The five-alanine motif is highlighted in bold. At the bottom are shown the HXB2 reference sequence, the consensus sequence of sample surrounding the 'RxEPS' insertion in $p6^{Gag}$, and the positive correlations between 'RxEPS' and P453S (x for R/L). The PTAP region is highlighted in bold, and the duplication is marked underlined. **(B)** The abundance of 'P453S' and 'RxEPS' insertion in the viral population over time. The two antiretroviral therapies used were composed of Lamivudine (3TC), Zidovudine (AZT), and Indinavir (IDV) or Nelfinavir (NFV). **(C)** Significant correlations between 'RxEPS' insertions and P453S are indicated by arrows among all associations involving 'RxEPS' in Sample 4, with their contingency tables shown at the top right. **(D)** The abundance of P453S in reads with and without 'RxEPS' insertions. **(E)** Significant correlation between 'RPEPS' insertion and K436R is indicated by arrow among all associations involving 'RPEPS' in Sample 1, with its contingency table shown at the top right. **(F)** The abundance of K436R in reads with and without 'RPEPS' insertions.

Table 3. InDels and recombination events detected in longitudinal HIV samples

Event	Region	Amino acid change	Sample
2157_AGACCAGAGCCATCA_2158, 2157_AGACTAGAGCCATCA_2158	$p6^{gag}$	RPEPS, RLEPS	1–5
1148_GGCAGC_1149, 1151_CAGCTG_1151	MA	AA, AA	1–5
1174_CCAGCAGCCAAGTCA_1175, 1174_CCAGCAGCCAGGTCA_1175	MA/CA cleavage site	TSSQV, TSSQV	2, 4

detected together with the 'RxEPS' insertion (Figure 4D). The ratio of P453S abundance in reads with and without 'RLEPS' was 27.74 ± 17.66 (SD) and the ratio of P453S abundance in reads with and without 'RPEPS' was 36.43 ± 16.67 (SD) (Supplementary Table S9). In the last sample, while only 5% of the reads without 'RxEPS' insertion carried the P453S substitution, this number increased to 99% and 69% in reads with 'RLEPS' and 'RPEPS', respectively (Supplementary Table S9). This finding supported cooperativity between P453S in the P1/p6 cleavage site and the 'RxEPS' insertion in the PTAP region. In contrast to C2146T, the A2096G mutation encoding the K436R substitution in the Gag NC/P1 cleavage site negatively correlated with the 'RPEPS' insertion (Figure 4E, Supplementary Table S8). The ratio of the K436R abundance in reads without and with 'RPEPS' was 12.27 ± 1.71 (SD) in the first three samples before its abundance decreasing to lower than 1% (Figure 4F, Supplementary Table S9). Consistent with the above, a combination of both K436R substitution and P453S substitution was not favored in the HIV genome (Supplementary Table S10).

In addition to the associations between 'RxEPS' in $p6^{Gag}$ and mutations in the Gag cleavages sites, the 'AA' insertion in the MA negatively correlated with the D121A substitution (Supplementary Figure S5A), resulting in a five-alanine motif in the C terminus of MA, which has been seen in the isolate ARV2/SF2. Additionally, the 'TSSQV' insertion in the MA/CA cleavage site negatively correlated with D121A (Supplementary Figure S5B) and positively correlated with the 'AA' insertion (wLD = 0.021805 in sample 4).

DISCUSSION

Adaption of viruses to their environments and to antiviral therapies occurs through both the acquisition of novel single-nucleotide variants (SNVs) as well as recombination events such as small structural variants or larger insertions and deletions (18–20). These adaptions seldom occur in isolation, rather, multiple adaptions work together to confer the virus with novel biological properties. The covariation of SNVs during viral evolution has been well-described for a range of viral systems (6,8). Bioinformatic tools that detect these co-varying SNVs are likewise readily available (reviewed in (9)). These report haplotypes using consensus-level data from large curated viral genomics databases as well as at the level of the viral intra-host diversity measured using Next-Generation Sequencing (13–15). However, computational tools that determine whether SNVs are correlated with recombination events or whether multiple recombination events are correlated with one another have to date been lacking. To address this gap, we therefore revised our previously reported CoVaMa (v0.1) pipeline to the new v0.7. We demonstrated the utility of this approach by studying two different viral systems characterized on different sequencing platforms.

Nanopore sequencing of FHV samples across passages revealed a strong linkage disequilibrium between the large deletion events that constituted D-RNA species that we previously characterized, consistent with the model that only '*mature*' D-RNAs efficiently replicate (24). Here we additionally found SNVs across the viral genome that were either positively correlated or negatively correlated with the deletions found in D-RNAs. These SNVs were previously too distantly spaced from each other or from deletion events to be correlated in the Illumina data, illustrating the value of performing long-read nanopore sequencing for this type of analysis.

The diversity and frequency changes of D-RNAs over passaging indicated the dynamic generation of D-RNAs from full-length RNAs and the competition among D-RNAs. Our data demonstrated that the D-RNA genomes acquired adaptions that were not found in the full-length wild-type genomic RNA. These adaptations in D-RNAs might be favored due to reduced genetic barriers of RNA2 mutation, due to not being responsible for functional viral expression, and could allow the formation of new secondary structures that confer D-RNAs with their-replicative and/or packaging advantage which have been granted the freedom to mutate by virtue of not being responsible for functional viral protein expression. Indeed, the G575A encodes an A185T mutation in the viral capsid protein at the five-fold and quasi-three-fold symmetry axes of the icosahedral virus particle. Such a substitution at this interface may prevent efficient virus assembly. Conversely, adaptations in the full-length viral genome that were rarely found in the D-RNAs and only started to enrich after the emergence of D-RNAs (such as the synonymous A226G SNV found here), suggested that the full-length '*helper*' virus may be adapting or escaping from the '*interfering*' properties of the D-RNAs. Although the mechanism of escape was not clear from these data, such a phenomenon was originally postulated by DePolo et al (49) who demonstrated that vesicular stomatitis virus (VSV) isolated from late viral passages was not subject to attenuation from defective interfering viral particles that arose in earlier viral passages. Further experimental characterization of these SNVs will reveal the precise molecular mechanisms driving their selection and competition.

Amino acid substitutions in cleavage sites and non-cleavage sites in HIV Gag have been shown to compensate for the compromised catalytic functions of protease with protease inhibitor (PI) drug resistant mutations (DRMs) and contribute to PI sensitivity (6,50–52). Besides the amino acid substitutions, insertions in Gag have been found to increase viral infectivity and drug resistance. Tamiya *et al.* reported that 'SRPE' duplication in $p6^{Gag}$ in multi-PI resistant HIV subtype G could increase the cleavage efficiency of protease with DRMs (46). They also showed that several insertions, mostly duplications, near Gag p1/p6 cleavage site (e.g. APP duplication) in different multi-PI resistant HIV-1 subtypes could restore the compromised catalytic functions of mutant protease (46). Full or partial PTAP duplications have been reported to be selected during anti-retroviral treatment (53,54). Martins *et al.* showed that, under PI pressure, full PTAP duplication in the Gag p6 significantly increases the cleavage efficiency of protease-bearing DRMs, thus increasing drug resistance and infectivity (55). In longitudinal serum samples collected from a patient who continuously failed antiretroviral therapies over six years, we detected an 'R(P/L)EPS' insertion in the P(T/S)AP region of the $p6^{Gag}$. Using CoVaMa v0.7, we found this insertion in a strong correlation with mu-

tations in Gag cleavage sites. The P453S mutation in the Gag P1/p6 CS was rarely detected in HIV strains without 'R(P/L)EPS' insertion, whereas it was enriched to over 90% in HIV isolates with 'R(P/L)EPS' insertion over time under drug pressure (Supplementary Table S9). 'RPEPS' belongs to the proline-rich motif 'RPEP(S/T)APP' in the N terminus of $p6^{Gag}$, which plays essential roles in the packaging of processed Pol proteins during late assembly (56,57). Replacing the P453 and P455 showed reduced viral replication in primary monocytes (57), which might account for the consistent low frequency of P453S in viral strains we analyzed. However, the duplication of 'RPEPS' restores the two prolines or adds more prolines in this motif, which could release the restriction and allow the P453S substitution. Together, this might account for the strong positive correlation detected by CoVaMa between 'RPEPS' and P453S. Additionally, CoVaMa reported a weak correlation between the 'RPEPS' insertion in $p6^{Gag}$ and A431V in the NC/P1 cleavage site (Supplementary Table S11), which improves the cleavage efficacy and is mostly selected in the presence of protease DRM V82A (58,59). The findings support the premise that for protease DRMs, duplications near the cleavage site and mutations in the cleavage site, may provide cooperative Gag cleavage functions and support drug-resistance development.

The final output of CoVaMa comprises a large table detailing the linkage disequilibrium found in each contingency table measured. As per the original report, CoVaMa also reports the LD values comprising the threshold values for three- and five-sigma. This value and the LD table can be used to provide a ranked list of genetic co-variations ranging from the most to least co-varying pairs of genetic adaptations. However, CoVaMa does not imply or provide a statistical framework with which to determine false discovery rates or degrees of significance with multiple hypothesis testing. CoVaMa can be applied to study diverse data types, including Illumina and Nanopore reads, which each have their own inherent error rates and profiles. These sequencing platforms are also applied in different manners to yield volumes of data with different number of technical and/or biological replicates that will vary according to the investigator's specifications. Furthermore, the underlying templates are naturally highly diverse, ranging from small multiple partite viruses (such as FHV) to long single-stranded RNA viruses, each also having their own error profile and different profiles of viral adaptions. As a result, different statistical frameworks that reflect the parameters used to detect and report covariance in the original sample must be deployed in each scenario to provide appropriate and robust estimates of statistical significance. These are not inherently provided by CoVaMa, but can be developed based upon the table of LD values reported in the CoVaMa output.

Overall, CoVaMa provides a simple and intuitive tool that probes both NGS datasets and nanopore datasets for evidence of the correlation between intra-host variants. Importantly, we here expanded this approach to detect and report the co-occurrence of SNVs with recombination events. While we focused here on viral intra-host diversity, the same approach and pipeline could equally be applied for NGS analysis of other organisms where diversity or correlation of sequence variants is anticipated, such as in bacterial or other complex mixtures of populations. We demonstrated the utility of this approach using both nanopore sequencing data acquired from the experimental evolution of the Flock House virus and using Illumina sequencing data acquired during the adaptation of HIV to antiviral therapies. In both cases, we observed novel associations of SNVs with specific recombination events. Knowledge of these associations is necessary to understand how viruses adapt to their environments and to characterize the distribution of specific genomic variants within viral intra-host diversity.

DATA AVAILABILITY

FHV datasets are publicly available in NCBI SRA under the accession code SRP094723. HIV datasets are publicly available in NCBI SRA under the accession code SRR15732332, SRR15732333, SRR15732334, SRR15732335, SRR15732336. The source code and supporting materials for CoVaMa (v0.7) are available at https://sourceforge.net/projects/covama/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Yiyang Zhou for providing comments and critically reading the manuscript.

FUNDING

National Institutes of Health [R21AI151725 to A.L.R.]; University of Texas System Rising STARs Award (to A.L.R.); National Institute of Allergy and Infectious Diseases [U54AI150472 to B.E.T., A.L.R]; National Human Genome Research Institute [R01HG009622 to B.E.T.]; Scripps Translational Science Institute [UL1TR001114-03 to B.E.T.]. Funding for open access charge: National Institute of Allergy and Infectious Diseases [U54AI150472 to B.E.T., A.L.R].

Conflict of interest statement. None declared.

REFERENCES

1. Simon-Loriere,E. and Holmes,E.C. (2011) Why do RNA viruses recombine? *Nat. Rev. Microbiol.*, **9**, 617–626.
2. Domingo,E., Sheldon,J. and Perales,C. (2012) Viral quasispecies evolution. *Microbiol. Mol. Biol. Rev.*, **76**, 159–216.
3. Vignuzzi,M. and López,C.B. (2019) Defective viral genomes are key drivers of the virus–host interaction. *Nat. Microbiol.*, **4**, 1075–1087.
4. Charpentier,C., Nora,T., Tenaillon,O., Clavel,F. and Hance,A.J. (2006) Extensive recombination among human immunodeficiency virus type 1 quasispecies makes an important contribution to viral diversity in individual patients. *J. Virol.*, **80**, 2472–2482.
5. Nora,T., Charpentier,C., Tenaillon,O., Hoede,C., Clavel,F. and Hance,A.J. (2007) Contribution of recombination to the evolution of human immunodeficiency viruses expressing resistance to antiretroviral treatment. *J. Virol.*, **81**, 7620–7628.
6. Flynn,W.F., Chang,M.W., Tan,Z., Oliveira,G., Yuan,J., Okulicz,J.F., Torbett,B.E. and Levy,R.M. (2015) Deep sequencing of protease inhibitor resistant HIV patient isolates reveals patterns of correlated mutations in Gag and protease. *PLoS Comput. Biol.*, **11**, e1004249.
7. Flynn,W.F., Haldane,A., Torbett,B.E. and Levy,R.M. (2017) Inference of epistatic effects leading to entrenchment and drug resistance in HIV-1 protease. *Mol. Biol. Evol.*, **34**, 1291–1306.

8. Aurora,R., Donlin,M.J., Cannon,N.A. and Tavis,J.E. (2009) Genome-wide hepatitis C virus amino acid covariance networks can predict response to antiviral therapy in humans. *J. Clin. Invest.*, **119**, 225–236.
9. Posada-Cespedes,S., Seifert,D. and Beerenwinkel,N. (2017) Recent advances in inferring viral diversity from high-throughput sequencing data. *Virus Res.*, **239**, 17–32.
10. Kuiken,C., Thurmond,J., Dimitrijevic,M. and Yoon,H. (2012) The LANL hemorrhagic fever virus database, a new platform for analyzing biothreat viruses. *Nucleic Acids Res.*, **40**, D587–D592.
11. Elbe,S. and Buckland-Merrett,G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.*, **1**, 33–46.
12. Vondrasek,J. and Wlodawer,A. (2002) HIVdb: a database of the structures of human immunodeficiency virus protease. *Proteins: Struct., Funct., Bioinf.*, **49**, 429–431.
13. Knyazev,S., Tsyvina,V., Shankar,A., Melnyk,A., Artyomenko,A., Malygina,T., Porozov,Y.B., Campbell,E.M., Switzer,W.M., Skums,P. *et al.* (2021) Accurate assembly of minority viral haplotypes from next-generation sequencing through efficient noise reduction. *Nucleic Acids Res.*, **49**, e102.
14. Macalalad,A.R., Zody,M.C., Charlebois,P., Lennon,N.J., Newman,R.M., Malboeuf,C.M., Ryan,E.M., Boutwell,C.L., Power,K.A., Brackney,D.E. *et al.* (2012) Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. *PLoS Comput. Biol.*, **8**, e1002417.
15. Yang,X., Charlebois,P., Macalalad,A., Henn,M.R. and Zody,M.C. (2013) V-Phaser 2: variant inference for viral populations. *BMC Genomics*, **14**, 674.
16. Routh,A., Chang,M.W., Okulicz,J.F., Johnson,J.E. and Torbett,B.E. (2015) CoVaMa: co-variation mapper for disequilibrium analysis of mutant loci in viral populations using next-generation sequence data. *Methods*, **91**, 40–47.
17. Töpfer,A., Zagordi,O., Prabhakaran,S., Roth,V., Halperin,E. and Beerenwinkel,N. (2013) Probabilistic inference of viral quasispecies subject to recombination. *J. Comput. Biol.*, **20**, 113–123.
18. Johnson,B.A., Xie,X., Bailey,A.L., Kalveram,B., Lokugamage,K.G., Muruato,A., Zou,J., Zhang,X., Juelich,T., Smith,J.K. *et al.* (2021) Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis. *Nature*, **591**, 293–299.
19. Gribble,J., Stevens,L.J., Agostini,M.L., Anderson-Daniels,J., Chappell,J.D., Lu,X., Pruijssers,A.J., Routh,A.L. and Denison,M.R. (2021) The coronavirus proofreading exoribonuclease mediates extensive viral recombination. *PLoS Pathog.*, **17**, e1009226.
20. Langsjoen,R.M., Muruato,A.E., Kunkel,S.R., Jaworski,E., Routh,A., Hardy,R.W. and Griffin,D.E. (2020) Differential Alphavirus Defective RNA Diversity between Intracellular and Extracellular Compartments Is Driven by Subgenomic Recombination Events. *mBio*, **11**, e00731-20.
21. Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.
22. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
23. Routh,A. and Johnson,J.E. (2014) Discovery of functional genomic motifs in viruses with ViReMa—a virus recombination mapper—for analysis of next-generation sequencing data. *Nucleic Acids Res.*, **42**, e11.
24. Jaworski,E. and Routh,A. (2017) Parallel ClickSeq and nanopore sequencing elucidates the rapid evolution of defective-interfering RNAs in Flock House virus. *PLoS Pathog.*, **13**, e1006365.
25. Loman,N.J. and Quinlan,A.R. (2014) Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics*, **30**, 3399–3401.
26. Chang,M.W., Oliveira,G., Yuan,J., Okulicz,J.F., Levy,S. and Torbett,B.E. (2013) Rapid deep sequencing of patient-derived HIV with ion semiconductor technology. *J. Virol. Methods*, **189**, 232–234.
27. Pukelsheim,F. (1994) The three sigma rule. *Am. Stat.*, **48**, 88–91.
28. Routh,A., Ordoukhanian,P. and Johnson,J.E. (2012) Nucleotide-resolution profiling of RNA recombination in the encapsidated genome of a eukaryotic RNA virus by next-generation sequencing. *J. Mol. Biol.*, **424**, 257–269.
29. Venter,P.A. and Schneemann,A. (2008) Recent insights into the biology and biomedical applications of Flock House virus. *Cell. Mol. Life Sci.*, **65**, 2675–2687.
30. Li,Y. and Ball,L.A. (1993) Nonhomologous RNA recombination during negative-strand synthesis of flock house virus RNA. *J. Virol.*, **67**, 3854–3860.
31. Dasgupta,R., Cheng,L.-L., Bartholomay,L.C. and Christensen,B.M. (2003) Flock house virus replicates and expresses green fluorescent protein in mosquitoes. *J. Gen. Virol.*, **84**, 1789–1797.
32. Zhong,W., Dasgupta,R. and Rueckert,R. (1992) Evidence that the packaging signal for nodaviral RNA2 is a bulged stem-loop. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 11146–11150.
33. Jovel,J. and Schneemann,A. (2011) Molecular characterization of Drosophila cells persistently infected with Flock House virus. *Virology*, **419**, 43–53.
34. Fisher,A.J. and Johnson,J.E. (1993) Ordered duplex RNA controls capsid architecture in an icosahedral animal virus. *Nature*, **361**, 176–179.
35. Rezelj,V.V., Levi,L.I. and Vignuzzi,M. (2018) The defective component of viral populations. *Curr. Opin. Virol.*, **33**, 74–80.
36. Ball,L.A. and Li,Y. (1993) cis-acting requirements for the replication of flock house virus RNA 2. *J. Virol.*, **67**, 3544–3551.
37. Gruber,A.R., Lorenz,R., Bernhart,S.H., Neubock,R. and Hofacker,I.L. (2008) The Vienna RNA Websuite. *Nucleic Acids Res.*, **36**, W70–W74.
38. Kaesberg,P., Dasgupta,R., Sgro,J.-Y., Wery,J.-P., Selling,B.H., Hosur,M.V. and Johnson,J.E. (1990) Structural homology among four nodaviruses as deduced by sequencing and X-ray crystallography. *J. Mol. Biol.*, **214**, 423–435.
39. Albariño,C.G., Eckerle,L.D. and Ball,L.A. (2003) The cis-acting replication signal at the 3′ end of Flock House virus RNA2 is RNA3-dependent. *Virology*, **311**, 181–191.
40. Rosskopf,J.J., Upton,J.H., Rodarte,L., Romero,T.A., Leung,M.-Y., Taufer,M. and Johnson,K.L. (2010) A 3′ terminal stem–loop structure in Nodamura virus RNA2 forms an essential cis-acting signal for RNA replication. *Virus Res.*, **150**, 12–21.
41. Zhou,Y. and Routh,A. (2020) Mapping RNA-capsid interactions and RNA secondary structure within virus particles using next-generation sequencing. *Nucleic Acids Res.*, **48**, e12.
42. Wu,T.D., Schiffer,C.A., Gonzales,M.J., Taylor,J., Kantor,R., Chou,S., Israelski,D., Zolopa,A.R., Fessel,W.J. and Shafer,R.W. (2003) Mutation patterns and structural correlates in human immunodeficiency virus type 1 protease following different protease inhibitor treatments. *J. Virol.*, **77**, 4836–4847.
43. Rhee,S.-Y., Liu,T.F., Holmes,S.P. and Shafer,R.W. (2007) HIV-1 subtype B protease and reverse transcriptase amino acid covariation. *PLoS Comput. Biol.*, **3**, e87.
44. Korber,B., Gaschen,B., Yusim,K., Thakallapally,R., Kesmir,C. and Detours,V. (2001) Evolutionary and immunological implications of contemporary HIV-1 variation. *Br. Med. Bull.*, **58**, 19–42.
45. Gallardo,C.M., Wang,S., Montiel-Garcia,D.J., Little,S.J., Smith,D.M., Routh,A.L. and Torbett,B.E. (2021) MrHAMER yields highly accurate single molecule viral sequences enabling analysis of intra-host evolution. *Nucleic Acids Res.*, **49**, e70.
46. Tamiya,S., Mardy,S., Kavlick,M.F., Yoshimura,K. and Mistuya,H. (2004) Amino acid insertions near Gag cleavage sites restore the otherwise compromised replication of human immunodeficiency virus type 1 variants resistant to protease inhibitors. *J. Virol.*, **78**, 12030–12040.
47. Marlowe,N., Flys,T., Hackett,J., Schumaker,M., Jackson,J.B. and Eshleman,S.H. (2004) Analysis of insertions and deletions in the gag p6 region of diverse HIV type 1 strains. *AIDS Res. Hum. Retroviruses*, **20**, 1119–1125.
48. Sanchez-Pescador,R., Power,M., Barr,P., Steimer,K., Stempien,M., Brown-Shimer,S., Gee,W., Renard,A., Randolph,A., Levy,J. *et al.* (1985) Nucleotide sequence and expression of an AIDS-associated retrovirus (ARV-2). *Science*, **227**, 484–492.
49. DePolo,N.J., Giachetti,C. and Holland,J.J. (1987) Continuing coevolution of virus and defective interfering particles and of viral genome sequences during undiluted passages: virus mutants exhibiting nearly complete resistance to formerly dominant defective interfering particles. *J. Virol.*, **61**, 454–464.
50. Dam,E., Quercia,R., Glass,B., Descamps,D., Launay,O., Duval,X., Kräusslich,H.-G., Hance,A.J., Clavel,F. and ANRS 109 Study Group (2009) Gag mutations strongly contribute to HIV-1 resistance to protease inhibitors in highly drug-experienced patients besides compensating for fitness loss. *PLoS Pathog.*, **5**, e1000345.

51. Kolli,M., Stawiski,E., Chappey,C. and Schiffer,C.A. (2009) Human immunodeficiency virus type 1 protease-correlated cleavage site mutations enhance inhibitor resistance. *J. Virol.*, **83**, 11027–11042.
52. Chang,M.W. and Torbett,B.E. (2011) Accessory mutations maintain stability in drug-resistant HIV-1 protease. *J. Mol. Biol.*, **410**, 756–760.
53. Martins,A.N., Arruda,M.B., Pires,A.F., Tanuri,A. and Brindeiro,R.M. (2010) Accumulation of P(T/S)AP late domain duplications in HIV type 1 subtypes B, C, and F derived from individuals failing ARV therapy and ARV drug-naive patients. *AIDS Res. Hum. Retroviruses*, **27**, 687–692.
54. Peters,S., Muñoz,M., Yerly,S., Sanchez-Merino,V., Lopez-Galindez,C., Perrin,L., Larder,B., Cmarko,D., Fakan,S., Meylan,P. *et al.* (2001) Resistance to nucleoside analog reverse transcriptase inhibitors mediated by human immunodeficiency virus type 1 p6 protein. *J. Virol.*, **75**, 9644–9653.
55. Martins,A.N., Waheed,A.A., Ablan,S.D., Huang,W., Newton,A., Petropoulos,C.J., Brindeiro,R.D.M. and Freed,E.O. (2015) Elucidation of the molecular mechanism driving duplication of the HIV-1 PTAP late domain. *J. Virol.*, **90**, 768–779.
56. Maguire,M.F., Guinea,R., Griffin,P., Macmanus,S., Elston,R.C., Wolfram,J., Richards,N., Hanlon,M.H., Porter,D.J.T., Wrin,T. *et al.* (2002) Changes in human immunodeficiency virus type 1 Gag at positions L449 and P453 are linked to I50V protease mutants in vivo and cause reduction of sensitivity to amprenavir and improved viral fitness in vitro. *J. Virol.*, **76**, 7398–7406.
57. Dettenhofer,M. and Yu,X.-F. (1999) Proline residues in human immunodeficiency virus type 1 p6Gag exert a cell type-dependent effect on viral replication and virion incorporation of Pol proteins. *J. Virol.*, **73**, 4696–4704.
58. Prabu-Jeyabalan,M., Nalivaika,E.A., King,N.M. and Schiffer,C.A. (2004) Structural basis for coevolution of a human immunodeficiency virus type 1 nucleocapsid-p1 cleavage site with a V82A drug-resistant mutation in viral protease. *J. Virol.*, **78**, 12446–12454.
59. Bally,F., Martinez,R., Peters,S., Sudre,P. and Telenti,A. (2000) Polymorphism of HIV Type 1 Gag p7/p1 and p1/p6 cleavage sites: clinical significance and implications for resistance to protease inhibitors. *AIDS Res. Hum. Retroviruses*, **16**, 1209–1213.

Appendix H

Number of nuclei per Seurat cluster from NAc snRNAseq experiment

orig.ident	0	1	2	3	4	5	6	7	8	9	10	11	12
Ctr.BK1S	996	996	575	352	619	242	787	199	14	27	143	107	50
Ctr.BK2S	188	162	174	228	192	797	349	186	1	82	8	47	99
Ctr.BK4S	234	247	98	284	170	517	246	175	1	29	27	56	152
Ctr.BL2S	484	397	419	286	283	313	227	131	1	280	663	49	79
Ctr.BL5S	772	842	355	534	386	391	136	247	34	135	63	86	153
Fen.BK11F	621	583	581	353	333	395	612	170	2	337	107	68	97
Fen.BK5F	646	699	317	343	383	447	134	161	14	32	38	67	62
Fen.BK6F	1316	1266	797	448	721	148	108	157	438	103	185	169	40
Fen.BK7F	320	182	179	526	284	564	60	201	96	268	30	62	207
Fen.BK8F	692	797	345	407	358	205	61	138	116	10	38	102	87
Fen.BL12F	710	719	597	341	476	167	912	252	NA	101	87	134	58
Fen.BL9F	429	278	313	415	297	151	203	139	1186	112	72	286	34

orig.ident	13	14	15	16	17	18	19	20	21	22	23	24	25
Ctr.BK1S	40	161	148	114	79	102	44	NA	26	46	16	35	6
Ctr.BK2S	11	31	177	33	16	24	26	3	12	9	6	6	13
Ctr.BK4S	39	10	76	12	19	25	21	12	15	2	1	4	8
Ctr.BL2S	50	61	110	52	18	55	12	2	10	19	5	9	7
Ctr.BL5S	66	42	20	33	78	43	24	15	20	13	2	4	13
Fen.BK11F	61	98	181	85	41	48	29	1	29	27	11	13	25
Fen.BK5F	88	44	22	68	73	42	19	2	11	11	13	14	9
Fen.BK6F	258	170	NA	146	127	66	33	3	38	36	2	13	10
Fen.BK7F	59	20	3	22	46	23	24	196	7	1	NA	21	14
Fen.BK8F	146	30	2	38	84	38	24	25	10	14	4	8	13
Fen.BL12F	81	131	67	91	90	62	48	5	30	41	89	20	7
Fen.BL9F	49	66	NA	61	66	43	19	42	29	8	59	16	7

Appendix I

Number of nuclei of each cell type in NAc snRNAseq experiment

orig.ident	Drd2-MSN	Drd1-MSN	Drd3-MSN	Astrocytes	Grm8-MSN	Oligo	Ng2	Glut	Astro/Drd2
Ctr.BK1S	1115	996	602	516	1406	321	199	14	143
Ctr.BK2S	215	162	256	411	541	829	186	1	8
Ctr.BK4S	292	247	127	361	416	542	175	1	27
Ctr.BL2S	552	397	699	401	510	334	131	1	663
Ctr.BL5S	916	842	490	556	522	419	247	34	63
Fen.BK11F	723	583	918	545	945	437	170	2	107
Fen.BK5F	807	699	349	378	517	480	161	14	38
Fen.BK6F	1701	1266	900	450	829	194	157	438	185
Fen.BK7F	425	182	447	529	344	609	201	96	30
Fen.BK8F	922	797	355	413	419	237	138	116	38
Fen.BL12F	881	719	698	497	1388	235	252	NA	87
Fen.BL9F	544	278	425	474	500	186	139	1186	72

orig.ident	Pvalb-int	Microglia	Oligo/Drd1	Micro/Drd2	Sst-int	Mitosis	Oligo/Astro	Cholinergic	Mural
Ctr.BK1S	107	50	161	114	102	NA	26	46	6
Ctr.BK2S	47	99	31	33	24	3	12	9	13
Ctr.BK4S	56	152	10	12	25	12	15	2	8
Ctr.BL2S	49	79	61	52	55	2	10	19	7
Ctr.BL5S	86	153	42	33	43	15	20	13	13
Fen.BK11F	68	97	98	85	48	1	29	27	25
Fen.BK5F	67	62	44	68	42	2	11	11	9
Fen.BK6F	169	40	170	146	66	3	38	36	10
Fen.BK7F	62	207	20	22	23	196	7	1	14
Fen.BK8F	102	87	30	38	38	25	10	14	13
Fen.BL12F	134	58	131	91	62	5	30	41	7
Fen.BL9F	286	34	66	61	43	42	29	8	7

REFERENCES

1. R. Perera, M. Khaliq, R. J. Kuhn, Closing the door on flaviviruses: entry as a target for antiviral drug design. *Antiviral Res* **80**, 11-22 (2008).
2. B. H. Song, S. I. Yun, M. Woolley, Y. M. Lee, Zika virus: History, epidemiology, transmission, and clinical presentation. *J Neuroimmunol* **308**, 50-64 (2017).
3. P. M. Armstrong, T. G. Andreadis, J. J. Shepard, M. C. Thomas, Northern range expansion of the Asian tiger mosquito (Aedes albopictus): Analysis of mosquito data from Connecticut, USA. *PLoS Negl Trop Dis* **11**, e0005623 (2017).
4. A. W. Bartlow *et al.*, Forecasting Zoonotic Infectious Disease Response to Climate Change: Mosquito Vectors and a Changing Environment. *Vet Sci* **6**, (2019).
5. M. U. Kraemer *et al.*, The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. *Elife* **4**, e08347 (2015).
6. L. Villar *et al.*, Efficacy of a tetravalent dengue vaccine in children in Latin America. *N Engl J Med* **372**, 113-123 (2015).
7. S. R. Hadinegoro *et al.*, Efficacy and Long-Term Safety of a Dengue Vaccine in Regions of Endemic Disease. *N Engl J Med* **373**, 1195-1206 (2015).
8. O. Dyer, Philippines halts dengue immunisation campaign owing to safety risk. *BMJ* **359**, j5759 (2017).
9. S. C. Weaver, Emergence of Epidemic Zika Virus Transmission and Congenital Zika Syndrome: Are Recently Evolved Traits to Blame? *mBio* **8**, (2017).
10. G. L. Ming, H. Tang, H. Song, Advances in Zika Virus Research: Stem Cell Models, Challenges, and Opportunities. *Cell Stem Cell* **19**, 690-702 (2016).
11. M. J. Counotte *et al.*, Sexual transmission of Zika virus and other flaviviruses: A living systematic review. *PLoS Med* **15**, e1002611 (2018).
12. A. Pattnaik, B. R. Sahoo, A. K. Pattnaik, Current Status of Zika Virus Vaccines: Successes and Challenges. *Vaccines (Basel)* **8**, (2020).
13. S. HOTTA, Experimental studies on dengue. I. Isolation, identification and modification of the virus. *J Infect Dis* **90**, 1-9 (1952).
14. A. B. Sabin, R. W. Schlesinger, PRODUCTION OF IMMUNITY TO DENGUE WITH VIRUS MODIFIED BY PROPAGATION IN MICE. *Science* **101**, 640-642 (1945).
15. E. C. Holmes, S. S. Twiddy, The origin, emergence and evolutionary genetics of dengue virus. *Infect Genet Evol* **3**, 19-28 (2003).
16. M. G. Guzman *et al.*, Dengue: a continuing global threat. *Nat Rev Microbiol* **8**, S7-16 (2010).
17. C. f. D. C. a. Prevention. (https://www.cdc.gov/dengue/about/index.html, 2021).
18. M. R. Capeding *et al.*, Clinical efficacy and safety of a novel tetravalent dengue vaccine in healthy children in Asia: a phase 3, randomised, observer-masked, placebo-controlled trial. *Lancet* **384**, 1358-1365 (2014).
19. R. Shukla, V. Ramasamy, R. K. Shanmugam, R. Ahuja, N. Khanna, Antibody-Dependent Enhancement: A Challenge for Developing a Safe Dengue Vaccine. *Front Cell Infect Microbiol* **10**, 572681 (2020).

20. S. S. Whitehead, J. E. Blaney, A. P. Durbin, B. R. Murphy, Prospects for a dengue virus vaccine. *Nat Rev Microbiol* **5**, 518-528 (2007).
21. T. C. Pierson, M. S. Diamond, The continued threat of emerging flaviviruses. *Nat Microbiol* **5**, 796-812 (2020).
22. A. L. Rothman, Immunity to dengue virus: a tale of original antigenic sin and tropical cytokine storms. *Nat Rev Immunol* **11**, 532-543 (2011).
23. N. T. Ngo *et al.*, Acute management of dengue shock syndrome: a randomized double-blind comparison of 4 intravenous fluid regimens in the first hour. *Clin Infect Dis* **32**, 204-213 (2001).
24. H. Puerta-Guardo, D. R. Glasner, E. Harris, Dengue Virus NS1 Disrupts the Endothelial Glycocalyx, Leading to Hyperpermeability. *PLoS Pathog* **12**, e1005738 (2016).
25. D. R. Glasner *et al.*, Dengue virus NS1 cytokine-independent vascular leak is dependent on endothelial glycocalyx components. *PLoS Pathog* **13**, e1006673 (2017).
26. P. Scaturro, M. Cortese, L. Chatel-Chaix, W. Fischl, R. Bartenschlager, Dengue Virus Non-structural Protein 1 Modulates Infectious Particle Production via Interaction with the Structural Proteins. *PLoS Pathog* **11**, e1005277 (2015).
27. J. J. Miner, M. S. Diamond, Zika Virus Pathogenesis and Tissue Tropism. *Cell Host Microbe* **21**, 134-142 (2017).
28. J. M. Mansuy *et al.*, Zika virus in semen and spermatozoa. *Lancet Infect Dis* **16**, 1106-1107 (2016).
29. G. Joguet *et al.*, Effect of acute Zika virus infection on sperm and virus clearance in body fluids: a prospective observational study. *Lancet Infect Dis* **17**, 1200-1208 (2017).
30. Q. Shao *et al.*, Zika virus infection disrupts neurovascular development and results in postnatal microcephaly with brain damage. *Development* **143**, 4127-4136 (2016).
31. Y. Pan *et al.*, Flaviviruses: Innate Immunity, Inflammasome Activation, Inflammatory Cell Death, and Cytokines. *Front Immunol* **13**, 829433 (2022).
32. M. I. Faizan *et al.*, Zika Virus-Induced Microcephaly and Its Possible Molecular Mechanism. *Intervirology* **59**, 152-158 (2016).
33. K. Rabelo *et al.*, Zika Induces Human Placental Damage and Inflammation. *Front Immunol* **11**, 2146 (2020).
34. N. Pardigon, Pathophysiological mechanisms of Flavivirus infection of the central nervous system. *Transfus Clin Biol* **24**, 96-100 (2017).
35. Q. Chen *et al.*, Metabolic reprogramming by Zika virus provokes inflammation in human placenta. *Nat Commun* **11**, 2967 (2020).
36. T. E. Morrison, M. S. Diamond, Animal Models of Zika Virus Infection, Pathogenesis, and Immunity. *J Virol* **91**, (2017).
37. L. Degenhardt *et al.*, The global epidemiology and burden of opioid dependence: results from the global burden of disease 2010 study. *Addiction* **109**, 1320-1333 (2014).
38. R. A. Wise, The role of reward pathways in the development of drug dependence. *Pharmacol Ther* **35**, 227-263 (1987).
39. J. Gholami *et al.*, Mortality and negative outcomes of opioid use and opioid use disorder: a six-year follow-up study. *Addiction*, (2022).
40. S. Walker *et al.*, More than saving lives: Qualitative findings of the UNODC/WHO Stop Overdose Safely (S-O-S) project. *Int J Drug Policy* **100**, 103482 (2022).
41. C. N. C. f. H. Statistics. (National Center for Health Statistics, 2021), vol. 2022.

42. A. Peterkin, J. Laks, Z. M. Weinstein, Current Best Practices for Acute and Chronic Management of Patients with Opioid Use Disorder. *Med Clin North Am* **106**, 61-80 (2022).
43. A. Bisaga *et al.*, Antagonists in the medical management of opioid use disorders: Historical and existing treatment strategies. *Am J Addict* **27**, 177-187 (2018).
44. C. Ezeomah *et al.*, Fentanyl self-administration impacts brain immune responses in male Sprague-Dawley rats. *Brain Behav Immun* **87**, 725-738 (2020).
45. L. L. García, L. Padilla, J. C. Castaño, Inhibitors compounds of the flavivirus replication process. *Virol J* **14**, 95 (2017).
46. N. A. Prow, D. N. Irani, The opioid receptor antagonist, naloxone, protects spinal motor neurons in a murine model of alphavirus encephalomyelitis. *Exp Neurol* **205**, 461-470 (2007).
47. M. A. Garcia-Blanco, S. G. Vasudevan, S. S. Bradrick, C. Nicchitta, Flavivirus RNA transactions from viral entry to genome replication. *Antiviral Res* **134**, 244-249 (2016).
48. C. J. Neufeldt, M. Cortese, E. G. Acosta, R. Bartenschlager, Rewiring cellular networks by members of the Flaviviridae family. *Nat Rev Microbiol* **16**, 125-142 (2018).
49. M. Cortese *et al.*, Ultrastructural Characterization of Zika Virus Replication Factories. *Cell reports* **18**, 2113-2123 (2017).
50. C. C. Wang, Z. S. Huang, P. L. Chiang, C. T. Chen, H. N. Wu, Analysis of the nucleoside triphosphatase, RNA triphosphatase, and unwinding activities of the helicase domain of dengue virus NS3 protein. *FEBS Lett* **583**, 691-696 (2009).
51. C. V. Filomatori *et al.*, A 5 ' RNA element promotes dengue virus RNA synthesis on a circular genome. *Gene Dev* **20**, 2238-2249 (2006).
52. W. L. Pong, Z. S. Huang, P. G. Teoh, C. C. Wang, H. N. Wu, RNA binding property and RNA chaperone activity of dengue virus core protein and other viral RNA-interacting proteins. *FEBS Lett* **585**, 2575-2581 (2011).
53. Y. Amador-Cañizares *et al.*, miR-122, small RNA annealing and sequence mutations alter the predicted structure of the Hepatitis C virus 5' UTR RNA to stabilize and promote viral RNA accumulation. *Nucleic Acids Res* **46**, 9776-9792 (2018).
54. P. Schult *et al.*, microRNA-122 amplifies hepatitis C virus translation by shaping the structure of the internal ribosomal entry site. *Nat Commun* **9**, 2613 (2018).
55. J. Blazevic, H. Rouha, V. Bradt, F. X. Heinz, K. Stiasny, Membrane Anchors of the Structural Flavivirus Proteins and Their Role in Virus Assembly. *J Virol* **90**, 6365-6378 (2016).
56. G. J. Chang, B. S. Davis, A. R. Hunt, D. A. Holmes, G. Kuno, Flavivirus DNA vaccines: current status and potential. *Ann N Y Acad Sci* **951**, 272-285 (2001).
57. B. M. Kümmerer, C. M. Rice, Mutations in the yellow fever virus nonstructural protein NS2A selectively block production of infectious particles. *J Virol* **76**, 4773-4784 (2002).
58. W. J. Liu, H. B. Chen, A. A. Khromykh, Molecular and functional analyses of Kunjin virus infectious cDNA clones demonstrate the essential roles for NS2A in virus assembly and for a nonconservative residue in NS3 in RNA replication. *J Virol* **77**, 7804-7813 (2003).
59. X. Zhang *et al.*, Zika Virus NS2A-Mediated Virion Assembly. *mBio* **10**, (2019).
60. X. Xie *et al.*, Dengue NS2A Protein Orchestrates Virus Assembly. *Cell Host Microbe*, (2019).

61. E. Jaworski, A. Routh, Parallel ClickSeq and Nanopore sequencing elucidates the rapid evolution of defective-interfering RNAs in Flock House virus. *PLoS Pathog* **13**, e1006365 (2017).
62. N. D. Collins *et al.*, Using Next Generation Sequencing to Study the Genetic Diversity of Candidate Live Attenuated Zika Vaccines. *Vaccines (Basel)* **8**, (2020).
63. E. Jaworski *et al.*, Tiled-ClickSeq for targeted sequencing of complete coronavirus genomes with simultaneous capture of RNA recombination and minority variants. *Elife* **10**, (2021).
64. C. Charre *et al.*, Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *Virus Evol* **6**, veaa075 (2020).
65. J. Li *et al.*, Rapid genomic characterization of SARS-CoV-2 viruses from clinical specimens using nanopore sequencing. *Sci Rep* **10**, 17492 (2020).
66. A. E. Mongan, J. S. B. Tuda, L. R. Runtuwene, Portable sequencer in the fight against infectious disease. *J Hum Genet* **65**, 35-40 (2020).
67. T. N. Adikari *et al.*, Single molecule, near full-length genome sequencing of dengue virus. *Sci Rep* **10**, 18196 (2020).
68. J. Batovska, S. E. Lynch, B. C. Rodoni, T. I. Sawbridge, N. O. Cogan, Metagenomic arbovirus detection using MinION nanopore sequencing. *J Virol Methods* **249**, 79-84 (2017).
69. P. Reteng *et al.*, A targeted approach with nanopore sequencing for the universal detection and identification of flaviviruses. *Sci Rep* **11**, 19031 (2021).
70. G. Beauclair *et al.*, : defective interfering viral genomes' detector for next-generation sequencing data. *RNA* **24**, 1285-1296 (2018).
71. Q. Wang, P. Jia, Z. Zhao, VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. *Genome Med* **7**, 2 (2015).
72. Y. Sun *et al.*, A specific sequence in the genome of respiratory syncytial virus regulates the generation of copy-back defective viral genomes. *PLoS Pathog* **15**, e1007707 (2019).
73. A. Routh, J. E. Johnson, Discovery of functional genomic motifs in viruses with ViReMa-a Virus Recombination Mapper-for analysis of next-generation sequencing data. *Nucleic Acids Res* **42**, e11 (2014).
74. A. Routh, M. W. Chang, J. F. Okulicz, J. E. Johnson, B. E. Torbett, CoVaMa: Co-Variation Mapper for disequilibrium analysis of mutant loci in viral populations using next-generation sequence data. *Methods* **91**, 40-47 (2015).
75. S. Wang *et al.*, Covariation of viral recombination with single nucleotide variants during virus evolution revealed by CoVaMa. *Nucleic Acids Res*, (2022).
76. M. S. Amaral *et al.*, Differential gene expression elicited by ZIKV infection in trophoblasts from congenital Zika syndrome discordant twins. *PLoS Negl Trop Dis* **14**, e0008424 (2020).
77. G. Bonenfant *et al.*, Asian Zika Virus Isolate Significantly Changes the Transcriptional Profile and Alternative RNA Splicing Events in a Neuroblastoma Cell Line. *Viruses* **12**, (2020).
78. K. Etebari *et al.*, Global Transcriptome Analysis of. *mSphere* **2**, (2017).
79. B. Hu *et al.*, ZIKV infection effects changes in gene splicing, isoform composition and lncRNA expression in human neural progenitor cells. *Virol J* **14**, 217 (2017).

80. S. Khaiboullina *et al.*, Transcriptome Profiling Reveals Pro-Inflammatory Cytokines and Matrix Metalloproteinase Activation in Zika Virus Infected Human Umbilical Vein Endothelial Cells. *Front Pharmacol* **10**, 642 (2019).
81. M. C. Lima *et al.*, The Transcriptional and Protein Profile From Human Infected Neuroprogenitor Cells Is Strongly Correlated to Zika Virus Microcephaly Cytokines Phenotype Evidencing a Persistent Inflammation in the CNS. *Front Immunol* **10**, 1928 (2019).
82. P. K. Singh *et al.*, Determination of system level alterations in host transcriptome due to Zika virus (ZIKV) Infection in retinal pigment epithelium. *Sci Rep* **8**, 11209 (2018).
83. D. Michalski *et al.*, Zika virus noncoding sfRNAs sequester multiple host-derived RNA-binding proteins and modulate mRNA decay and splicing during infection. *J Biol Chem* **294**, 16282-16296 (2019).
84. D. Kovanich *et al.*, Analysis of the Zika and Japanese Encephalitis Virus NS5 Interactomes. *J Proteome Res* **18**, 3203-3218 (2019).
85. P. Loke *et al.*, Gene expression patterns of dengue virus-infected children from nicaragua reveal a distinct signature of increased metabolism. *PLoS Negl Trop Dis* **4**, e710 (2010).
86. M. Li *et al.*, Transcriptome Analysis of Responses to Dengue Virus 2 Infection in. *Viruses* **13**, (2021).
87. H. Luo *et al.*, Zika, dengue and yellow fever viruses induce differential anti-viral immune responses in human monocytic and first trimester trophoblast cells. *Antiviral Res* **151**, 55-62 (2018).
88. B. Pozzi *et al.*, Dengue virus targets RBM10 deregulating host cell splicing and innate immune response. *Nucleic Acids Res* **48**, 6824-6838 (2020).
89. F. A. De Maio *et al.*, The Dengue Virus NS5 Protein Intrudes in the Cellular Spliceosome and Modulates Splicing. *PLoS Pathog* **12**, e1005841 (2016).
90. J. Ule *et al.*, CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302**, 1212-1215 (2003).
91. J. Ule, K. Jensen, A. Mele, R. B. Darnell, CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods* **37**, 376-386 (2005).
92. C. Harold, D. Cox, K. J. Riley, Epstein-Barr viral microRNAs target caspase 3. *Virol J* **13**, 145 (2016).
93. I. Haecker, R. Renne, HITS-CLIP and PAR-CLIP advance viral miRNA targetome analysis. *Crit Rev Eukaryot Gene Expr* **24**, 101-116 (2014).
94. I. Haecker *et al.*, Ago HITS-CLIP expands understanding of Kaposi's sarcoma-associated herpesvirus miRNA function in primary effusion lymphomas. *PLoS Pathog* **8**, e1002884 (2012).
95. J. M. Luna *et al.*, Hepatitis C virus RNA functionally sequesters miR-122. *Cell* **160**, 1099-1110 (2015).
96. R. Craigie, The molecular biology of HIV integrase. *Future Virol* **7**, 679-686 (2012).
97. O. Delelis, K. Carayon, A. Saib, E. Deprez, J. F. Mouscadet, Integrase and integration: biochemical activities of HIV-1 integrase. *Retrovirology* **5**, 114 (2008).
98. T. K. Chiu, D. R. Davies, Structure and function of HIV-1 integrase. *Curr Top Med Chem* **4**, 965-977 (2004).
99. A. Engelman, In vivo analysis of retroviral integrase structure and function. *Adv Virus Res* **52**, 411-426 (1999).

100. B. C. Johnson, M. Metifiot, A. Ferris, Y. Pommier, S. H. Hughes, A homology model of HIV-1 integrase and analysis of mutations designed to test the model. *Journal of molecular biology* **425**, 2133-2146 (2013).
101. K. D. Mohammed, M. B. Topper, M. A. Muesing, Sequential deletion of the integrase (Gag-Pol) carboxyl terminus reveals distinct phenotypic classes of defective HIV-1. *Journal of virology* **85**, 4654-4666 (2011).
102. R. Lu *et al.*, Class II integrase mutants with changes in putative nuclear localization signals are primarily blocked at a postnuclear entry step of human immunodeficiency virus type 1 replication. *Journal of virology* **78**, 12735-12746 (2004).
103. A. Engelman, G. Englund, J. M. Orenstein, M. A. Martin, R. Craigie, Multiple effects of mutations in human immunodeficiency virus type 1 integrase on viral replication. *Journal of virology* **69**, 2729-2736 (1995).
104. J. J. Kessl *et al.*, HIV-1 Integrase Binds the Viral RNA Genome and Is Essential during Virion Morphogenesis. *Cell* **166**, 1257-1268 e1212 (2016).
105. S. Bannwarth, A. Gatignol, HIV-1 TAR RNA: the target of molecular interactions between the virus and its host. *Curr HIV Res* **3**, 61-71 (2005).
106. N. Lee *et al.*, Genome-wide analysis of influenza viral RNA and nucleoprotein association. *Nucleic Acids Res* **45**, 8968-8977 (2017).
107. Y. Zhou, A. Routh, Mapping RNA-capsid interactions and RNA secondary structure within virus particles using next-generation sequencing. *Nucleic Acids Res* **48**, e12 (2020).
108. S. B. Kutluay, P. D. Bieniasz, Analysis of HIV-1 Gag-RNA Interactions in Cells and Virions by CLIP-seq. *Methods in molecular biology* **1354**, 119-131 (2016).
109. S. B. Kutluay *et al.*, Global changes in the RNA binding specificity of HIV-1 gag regulate virion genesis. *Cell* **159**, 1096-1109 (2014).
110. M. Stefanik *et al.*, FDA-Approved Drugs Efavirenz, Tipranavir, and Dasabuvir Inhibit Replication of Multiple Flaviviruses in Vero Cells. *Microorganisms* **8**, (2020).
111. S. Sotcheff, A. Routh, Understanding Flavivirus Capsid Protein Functions: The Tip of the Iceberg. *Pathogens* **9**, (2020).
112. A. Sampath, R. Padmanabhan, Molecular targets for flavivirus drug discovery. *Antiviral Res* **81**, 6-15 (2009).
113. Y. Zhou *et al.*, Structure and function of flavivirus NS5 methyltransferase. *J Virol* **81**, 3891-3903 (2007).
114. A. Gharbi-Ayachi, A. El Sahili, J. Lescar, Purification of Dengue and Zika Virus Non-structural Protein 5 for Crystallization and Screening of Antivirals. *Methods Mol Biol* **2409**, 47-61 (2022).
115. Y. Shi, G. F. Gao, Structural Biology of the Zika Virus. *Trends Biochem Sci* **42**, 443-456 (2017).
116. W. M. Kok, New developments in flavivirus drug discovery. *Expert Opin Drug Discov* **11**, 433-445 (2016).
117. B. Wang *et al.*, Structural basis for STAT2 suppression by flavivirus NS5. *Nat Struct Mol Biol* **27**, 875-885 (2020).
118. S. M. Best, The Many Faces of the Flavivirus NS5 Protein in Antagonism of Type I Interferon Signaling. *J Virol* **91**, (2017).
119. N. J. Barrows *et al.*, Biochemistry and Molecular Biology of Flaviviruses. *Chem Rev* **118**, 4448-4482 (2018).

120. W. W. Phoo *et al.*, Crystal structures of full length DENV4 NS2B-NS3 reveal the dynamic interaction between NS2B and NS3. *Antiviral Res* **182**, 104900 (2020).
121. N. J. Braun *et al.*, Structure-Based Macrocyclization of Substrate Analogue NS2B-NS3 Protease Inhibitors of Zika, West Nile and Dengue viruses. *ChemMedChem* **15**, 1439-1452 (2020).
122. A. A. Rabaan *et al.*, Overview of hepatitis C infection, molecular biology, and new treatment. *J Infect Public Health* **13**, 773-783 (2020).
123. K. A. Salam, N. Akimitsu, Hepatitis C virus NS3 inhibitors: current and future perspectives. *Biomed Res Int* **2013**, 467869 (2013).
124. W. Carter, S. Connelly, K. Struble, Reinventing HCV Treatment: Past and Future Perspectives. *J Clin Pharmacol* **57**, 287-296 (2017).
125. T. Dokland *et al.*, West Nile virus core protein; tetramer structure and ribbon formation. *Structure* **12**, 1157-1163 (2004).
126. Z. Shang, H. Song, Y. Shi, J. Qi, G. F. Gao, Crystal Structure of the Capsid Protein from Zika Virus. *J Mol Biol* **430**, 948-962 (2018).
127. T. Poonsiri, G. S. A. Wright, T. Solomon, S. V. Antonyuk, Crystal Structure of the Japanese Encephalitis Virus Capsid Protein. *Viruses* **11**, (2019).
128. C. T. Jones *et al.*, Flavivirus capsid is a dimeric alpha-helical protein. *J Virol* **77**, 7143-7149 (2003).
129. M. M. Samsa, J. A. Mondotte, J. J. Caramelo, A. V. Gamarnik, Uncoupling cis-Acting RNA elements from coding sequences revealed a requirement of the N-terminal region of dengue virus capsid protein in virus particle formation. *J Virol* **86**, 1046-1058 (2012).
130. L. Ma, C. T. Jones, T. D. Groesch, R. J. Kuhn, C. B. Post, Solution structure of dengue virus capsid protein reveals another fold. *Proc Natl Acad Sci U S A* **101**, 3414-3419 (2004).
131. M. M. Samsa *et al.*, Dengue virus capsid protein usurps lipid droplets for viral particle formation. *PLoS pathogens* **5**, e1000632 (2009).
132. G. H. Samuel, M. R. Wiley, A. Badawi, Z. N. Adelman, K. M. Myles, Yellow fever virus capsid protein is a potent suppressor of RNA silencing that binds double-stranded RNA. *Proc Natl Acad Sci U S A* **113**, 13863-13868 (2016).
133. C. V. Filomatori *et al.*, A 5' RNA element promotes dengue virus RNA synthesis on a circular genome. *Genes Dev* **20**, 2238-2249 (2006).
134. L. A. Byk, A. V. Gamarnik, Properties and Functions of the Dengue Virus Capsid Protein. *Annu Rev Virol* **3**, 263-281 (2016).
135. M. M. Samsa *et al.*, Dengue virus capsid protein usurps lipid droplets for viral particle formation. *PLoS Pathog* **5**, e1000632 (2009).
136. R. Bhuvanakantham, M. K. Chong, M. L. Ng, Specific interaction of capsid protein and importin-alpha/beta influences West Nile virus production. *Biochem Biophys Res Commun* **389**, 63-69 (2009).
137. I. C. Martins *et al.*, The disordered N-terminal region of dengue virus capsid protein contains a lipid-droplet-binding motif. *Biochem J* **444**, 405-415 (2012).
138. N. G. Iglesias *et al.*, Dengue Virus Uses a Non-Canonical Function of the Host GBF1-Arf-COPI System for Capsid Protein Accumulation on Lipid Droplets. *Traffic* **16**, 962-977 (2015).
139. K. Ishida *et al.*, Functional Correlation between Subcellular Localizations of Japanese Encephalitis Virus Capsid Protein and Virus Production. *J Virol* **93**, (2019).

140. W. Oh *et al.*, Jab1 mediates cytoplasmic localization and degradation of West Nile virus capsid protein. *J Biol Chem* **281**, 30166-30174 (2006).
141. L. P. Slomnicki *et al.*, Ribosomal stress and Tp53-mediated neuronal apoptosis in response to capsid protein of the Zika virus. *Scientific reports* **7**, 16652 (2017).
142. Y. Tsuda *et al.*, Nucleolar protein B23 interacts with Japanese encephalitis virus core protein and participates in viral replication. *Microbiol Immunol* **50**, 225-234 (2006).
143. Z. Xu, R. Anderson, T. C. Hobman, The capsid-binding nucleolar helicase DDX56 is important for infectivity of West Nile virus. *J Virol* **85**, 5571-5580 (2011).
144. M. R. Yang *et al.*, West Nile virus capsid protein induces p53-mediated apoptosis via the sequestration of HDM2 to the nucleolus. *Cell Microbiol* **10**, 165-176 (2008).
145. N. D. Elrod, E. A. Jaworski, P. Ji, E. J. Wagner, A. Routh, Development of Poly(A)-ClickSeq as a tool enabling simultaneous genome-wide poly(A)-site identification and differential expression analysis. *Methods* **155**, 20-29 (2019).
146. A. Routh *et al.*, Poly(A)-ClickSeq: click-chemistry for next-generation 3′-end sequencing without RNA enrichment or fragmentation. *Nucleic Acids Res* **45**, e112 (2017).
147. A. D. Tang *et al.*, Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun* **11**, 1438 (2020).
148. S. Ikemoto, Brain reward circuitry beyond the mesolimbic dopamine system: a neurobiological theory. *Neurosci Biobehav Rev* **35**, 129-150 (2010).
149. I. Sziráki, H. Sershen, A. Hashim, A. Lajtha, Receptors in the ventral tegmental area mediating nicotine-induced dopamine release in the nucleus accumbens. *Neurochem Res* **27**, 253-261 (2002).
150. L. Xu, J. Nan, Y. Lan, The Nucleus Accumbens: A Common Target in the Comorbidity of Depression and Addiction. *Front Neural Circuits* **14**, 37 (2020).
151. K. E. Savell *et al.*, A dopamine-induced gene expression signature regulates neuronal function and cocaine response. *Sci Adv* **6**, eaba4221 (2020).
152. M. D. Scofield *et al.*, The Nucleus Accumbens: Mechanisms of Addiction across Drug Classes Reflect the Importance of Glutamate Homeostasis. *Pharmacol Rev* **68**, 816-871 (2016).
153. Z. B. You, Y. Q. Chen, R. A. Wise, Dopamine and glutamate release in the nucleus accumbens and ventral tegmental area of rat following lateral hypothalamic self-stimulation. *Neuroscience* **107**, 629-639 (2001).
154. T. A. Zhang, R. E. Maldve, R. A. Morrisett, Coincident signaling in mesolimbic structures underlying alcohol reinforcement. *Biochem Pharmacol* **72**, 919-927 (2006).
155. E. Jaworski, A. Routh, ClickSeq: Replacing Fragmentation and Enzymatic Ligation with Click-Chemistry to Prevent Sequence Chimeras. *Methods Mol Biol* **1712**, 71-85 (2018).
156. A. Routh, S. R. Head, P. Ordoukhanian, J. E. Johnson, ClickSeq: Fragmentation-Free Next-Generation Sequencing via Click Ligation of Adaptors to Stochastically Terminated 3'-Azido cDNAs. *J Mol Biol* **427**, 2610-2616 (2015).
157. N. D. Grubaugh *et al.*, An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol* **20**, 8 (2019).
158. A. Routh, DPAC: A Tool for Differential Poly(A)-Cluster Usage from Poly(A)-Targeted RNAseq Data. *G3 (Bethesda)* **9**, 1825-1830 (2019).
159. T. L. Bailey, DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653-1659 (2011).

160. K. B. Cook, H. Kazan, K. Zuberi, Q. Morris, T. R. Hughes, RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res* **39**, D301-308 (2011).
161. W. J. Kent *et al.*, The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).
162. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).
163. P. Germain, A. Lun, W. Macnair, M. Robinson, Doublet identification in single-cell sequencing data using scDblFinder. *f1000research*, (2021).
164. M. U. G. Kraemer *et al.*, Past and future spread of the arbovirus vectors Aedes aegypti and Aedes albopictus. *Nat Microbiol* **4**, 854-863 (2019).
165. S. Grayo, Is the ZIKV Congenital Syndrome and Microcephaly Due to Syndemism with Latent Virus Coinfection? *Viruses* **13**, (2021).
166. J. Mlakar *et al.*, Zika Virus Associated with Microcephaly. *N Engl J Med* **374**, 951-958 (2016).
167. E. Y. Chen *et al.*, Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
168. M. V. Kuleshov *et al.*, Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90-97 (2016).
169. Z. Xie *et al.*, Gene Set Knowledge Discovery with Enrichr. *Curr Protoc* **1**, e90 (2021).
170. T. L. Bailey, P. Machanick, Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res* **40**, e128 (2012).
171. E. L. Van Nostrand *et al.*, A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711-719 (2020).
172. T. Hall-Pogar, S. Liang, L. K. Hague, C. S. Lutz, Specific trans-acting proteins interact with auxiliary RNA polyadenylation elements in the COX-2 3'-UTR. *RNA* **13**, 1103-1115 (2007).
173. S. Kaneko, O. Rozenblatt-Rosen, M. Meyerson, J. L. Manley, The multifunctional protein p54nrb/PSF recruits the exonuclease XRN2 to facilitate pre-mRNA 3' processing and transcription termination. *Genes Dev* **21**, 1779-1789 (2007).
174. I. Slišković, H. Eich, M. Müller-McNicoll, Exploring the multifunctionality of SR proteins. *Biochem Soc Trans*, (2021).
175. S. Liang, C. S. Lutz, p54nrb is a component of the snRNP-free U1A (SF-A) complex that promotes pre-mRNA cleavage during polyadenylation. *RNA* **12**, 111-121 (2006).
176. P. D. Scotti, S. Dearing, D. W. Mossop, Flock House virus: a nodavirus isolated from Costelytra zealandica (White) (Coleoptera: Scarabaeidae). *Arch Virol* **75**, 181-189 (1983).
177. D. H. Mathews *et al.*, Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci U S A* **101**, 7287-7292 (2004).
178. Z. Lu, J. Gong, Q. C. Zhang, PARIS: Psoralen Analysis of RNA Interactions and Structures with High Throughput and Resolution. *Methods Mol Biol* **1649**, 59-84 (2018).
179. O. Ziv *et al.*, COMRADES determines in vivo RNA structures and interactions. *Nat Methods* **15**, 785-788 (2018).

180. Y. F. Chang, S. M. Wang, K. J. Huang, C. T. Wang, Mutations in capsid major homology region affect assembly and membrane affinity of HIV-1 Gag. *J Mol Biol* **370**, 585-597 (2007).
181. C. Shan *et al.*, Using a Virion Assembly-Defective Dengue Virus as a Vaccine Approach. *J Virol* **92**, (2018).
182. R. Bocanegra, A. Rodríguez-Huete, M. Fuertes, M. Del Álamo, M. G. Mateu, Molecular recognition in the human immunodeficiency virus capsid and antiviral design. *Virus Res* **169**, 388-410 (2012).
183. J. Dietz *et al.*, Inhibition of HIV-1 by a peptide ligand of the genomic RNA packaging signal Psi. *ChemMedChem* **3**, 749-755 (2008).
184. S. J. Kim, M. Y. Kim, J. H. Lee, J. C. You, S. Jeong, Selection and stabilization of the RNA aptamers against the human immunodeficiency virus type-1 nucleocapsid protein. *Biochem Biophys Res Commun* **291**, 925-931 (2002).
185. A. Machara *et al.*, Specific Inhibitors of HIV Capsid Assembly Binding to the C-Terminal Domain of the Capsid Protein: Evaluation of 2-Arylquinazolines as Potential Antiviral Compounds. *J Med Chem* **59**, 545-558 (2016).
186. C. Tang *et al.*, Antiviral inhibition of the HIV-1 capsid protein. *J Mol Biol* **327**, 1013-1020 (2003).
187. S. Thenin-Houssier, S. T. Valente, HIV-1 Capsid Inhibitors as Antiretroviral Agents. *Curr HIV Res* **14**, 270-282 (2016).
188. S. Wu *et al.*, Discovery and Mechanistic Study of Benzamide Derivatives That Modulate Hepatitis B Virus Capsid Assembly. *J Virol* **91**, (2017).
189. T. Atieh, C. Baronti, X. de Lamballerie, A. Nougairède, Simple reverse genetics systems for Asian and African Zika viruses. *Sci Rep* **6**, 39384 (2016).
190. T. L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**, 28-36 (1994).
191. R. G. Huber *et al.*, Structure mapping of dengue and Zika viruses reveals functional long-range interactions. *Nat Commun* **10**, 1408 (2019).
192. S. Bhola *et al.*, Neurological toll of COVID-19. *Neurol Sci*, (2022).
193. M. A. Haidar *et al.*, SARS-CoV-2 involvement in central nervous system tissue damage. *Neural Regen Res* **17**, 1228-1239 (2022).
194. M. A. Salman *et al.*, Characteristics of Patients with SARS-CoV-2 Positive Cerebrospinal Fluid: A Systematic Review. *Int J Gen Med* **14**, 10385-10395 (2021).
195. S. Sharma, H. Jagadeesh, A. Saxena, H. Chakravarthy, V. Devanathan, Central nervous system as a target of novel coronavirus infections: Potential routes of entry and pathogenic mechanisms. *J Biosci* **46**, (2021).
196. M. Izrael *et al.*, Astrocytes Downregulate Inflammation in Lipopolysaccharide-Induced Acute Respiratory Distress Syndrome: Applicability to COVID-19. *Front Med (Lausanne)* **8**, 740071 (2021).
197. R. H. Cales, S. C. Cales, J. Shreffler, M. R. Huecker, The COVID-19 pandemic and opioid use disorder: Expanding treatment with buprenorphine, and combining safety precautions with telehealth. *J Subst Abuse Treat* **133**, 108543 (2022).
198. K. Alexander *et al.*, The impact of COVID-19 on healthcare delivery for people who use opioids: a scoping review. *Subst Abuse Treat Prev Policy* **16**, 60 (2021).
199. A. Treviño, V. Soriano, The opioid epidemic during the COVID-19 pandemic: Impact on HIV and HCV control. *AIDS Rev* **23**, 227 (2021).

200. F. Ahmad, L. Rossen, P. Sutton. (National Center for Health Statistics, 2022).
201. H. Mizher *et al.*, Plasma Concentrations of Pro-inflammatory Cytokine IL-6 and Antiinflammatory Cytokine IL-10 in Short- and Long-term Opioid Users with Noncancer Pain. *J Pharm Bioallied Sci* **12**, S663-S666 (2020).
202. Y. H. Chen *et al.*, Methadone enhances human influenza A virus replication. *Addict Biol* **22**, 257-271 (2017).
203. N. Schmidt *et al.*, The SARS-CoV-2 RNA-protein interactome in infected human cells. *Nat Microbiol* **6**, 339-353 (2021).
204. T. T. Karagiannis *et al.*, Single cell transcriptomics reveals opioid usage evokes widespread suppression of antiviral gene program. *Nat Commun* **11**, 2611 (2020).
205. Y. Lee *et al.*, Comparing mortality from covid-19 to mortality due to overdose: A micromort analysis. *J Affect Disord* **296**, 514-521 (2022).
206. K. Cameron, L. Rozano, M. Falasca, R. L. Mancera, Does the SARS-CoV-2 Spike Protein Receptor Binding Domain Interact Effectively with the DPP4 (CD26) Receptor? A Molecular Docking Study. *Int J Mol Sci* **22**, (2021).
207. E. F. Healy, M. Lilic, A model for COVID-19-induced dysregulation of ACE2 shedding by ADAM17. *Biochem Biophys Res Commun* **573**, 158-163 (2021).
208. N. L. Lartey *et al.*, ADAM17/MMP inhibition prevents neutrophilia and lung injury in a mouse model of COVID-19. *J Leukoc Biol*, (2021).
209. E. Azizan, M. Brown, ACE2 role in SARS-CoV-2 infectivity and Covid-19 severity. *Malays J Pathol* **42**, 363-367 (2020).
210. S. Lukassen *et al.*, SARS-CoV-2 receptor ACE2 and TMPRSS2 are primarily expressed in bronchial transient secretory cells. *EMBO J* **39**, e105114 (2020).
211. D. Zipeto, J. D. F. Palmeira, G. A. Argañaraz, E. R. Argañaraz, ACE2/ADAM17/TMPRSS2 Interplay May Be the Main Risk Factor for COVID-19. *Front Immunol* **11**, 576745 (2020).
212. M. Perera-Lecoin, L. Meertens, X. Carnec, A. Amara, Flavivirus entry receptors: an update. *Viruses* **6**, 69-88 (2013).
213. C. R. Ojha *et al.*, Toll-like receptor 3 regulates Zika virus infection and associated host inflammatory response in primary human astrocytes. *PLoS One* **14**, e0208543 (2019).
214. H. P. Jia *et al.*, Ectodomain shedding of angiotensin converting enzyme 2 in human airway epithelia. *Am J Physiol Lung Cell Mol Physiol* **297**, L84-96 (2009).
215. I. C. Huang *et al.*, SARS coronavirus, but not human coronavirus NL63, utilizes cathepsin L to infect ACE2-expressing cells. *J Biol Chem* **281**, 3198-3203 (2006).
216. R. Saji *et al.*, Combining IL-6 and SARS-CoV-2 RNAaemia-based risk stratification for fatal outcomes of COVID-19. *PLoS One* **16**, e0256022 (2021).
217. A. Santa Cruz *et al.*, Interleukin-6 Is a Biomarker for the Development of Fatal Severe Acute Respiratory Syndrome Coronavirus 2 Pneumonia. *Front Immunol* **12**, 613422 (2021).
218. C. f. D. Control. (2022), vol. 2022.
219. X. G. Yan *et al.*, [Isolation and identification of SARS virus in Guangdong province]. *Zhonghua Shi Yan He Lin Chuang Bing Du Xue Za Zhi* **17**, 213-216 (2003).
220. A. Bermingham *et al.*, Severe respiratory illness caused by a novel coronavirus, in a patient transferred to the United Kingdom from the Middle East, September 2012. *Euro Surveill* **17**, 20290 (2012).

221. G. Lu, D. Liu, SARS-like virus in the Middle East: a truly bat-related coronavirus causing human diseases. *Protein Cell* **3**, 803-805 (2012).
222. J. J. Zhang, X. Dong, G. H. Liu, Y. D. Gao, Risk and Protective Factors for COVID-19 Morbidity, Severity, and Mortality. *Clin Rev Allergy Immunol*, (2022).
223. J. C. Arévalo-Lorido *et al.*, The importance of association of comorbidities on COVID-19 outcomes: a machine learning approach. *Curr Med Res Opin*, 1-32 (2022).
224. E. Simon-Loriere, E. C. Holmes, Why do RNA viruses recombine? *Nat Rev Microbiol* **9**, 617-626 (2011).
225. C. B. López, Unexpected lessons from the neglected: How defective viral genomes became important again. *PLoS Pathog* **15**, e1007450 (2019).
226. E. Genoyer, C. B. López, The Impact of Defective Viruses on Infection and Immunity. *Annu Rev Virol* **6**, 547-566 (2019).
227. S. K. Lau *et al.*, Severe Acute Respiratory Syndrome (SARS) Coronavirus ORF8 Protein Is Acquired from SARS-Related Coronavirus from Greater Horseshoe Bats through Recombination. *J Virol* **89**, 10532-10547 (2015).
228. M. N. Vu *et al.* (BioRxiv, 2021).
229. B. A. Johnson *et al.*, Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis. *Nature* **591**, 293-299 (2021).
230. X. Li *et al.*, Emergence of SARS-CoV-2 through Recombination and Strong Purifying Selection. *bioRxiv*, (2020).
231. C. Genomes Project *et al.*, A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
232. S. Sotcheff *et al.* (BioRxiv, BioRxiv, 2022).
233. F. G. Alnaji *et al.*, (2018).
234. C. K. Pfaller *et al.*, Measles Virus Defective Interfering RNAs Are Generated Frequently and Early in the Absence of C Protein and Can Be Destabilized by Adenosine Deaminase Acting on RNA-1-Like Hypermutations. *J Virol* **89**, 7735-7747 (2015).
235. P. Calain, J. Curran, D. Kolakofsky, L. Roux, Molecular cloning of natural paramyxovirus copy-back defective interfering RNAs and their expression from DNA. *Virology* **191**, 62-71 (1992).
236. O. G. Andzhaparidze, I. S. Boriskin, N. N. Bogomolova, [Defective interfering mumps virus produced by chronically infected cell cultures]. *Vopr Virusol* **27**, 405-408 (1982).
237. S. R. Welch *et al.*, Inhibition of Nipah Virus by Defective Interfering Particles. *J Infect Dis* **221**, S460-S470 (2020).
238. G. G. Re, K. C. Gupta, D. W. Kingsbury, Sequence of the 5' end of the Sendai virus genome and its variable representation in complementary form at the 3' ends of copy-back defective interfering RNA species: identification of the L gene terminus. *Virology* **130**, 390-396 (1983).
239. G. G. Re, K. C. Gupta, D. W. Kingsbury, Genomic and copy-back 3' termini in Sendai virus defective interfering RNA species. *J Virol* **45**, 659-664 (1983).
240. E. Genoyer, C. B. López, Defective Viral Genomes Alter How Sendai Virus Interacts with Cellular Trafficking Machinery, Leading to Heterogeneity in the Production of Viral Particles among Infected Cells. *J Virol* **93**, (2019).
241. J. Xu *et al.*, Replication defective viral genomes exploit a cellular pro-survival mechanism to establish paramyxovirus persistence. *Nat Commun* **8**, 799 (2017).

242. X. Mercado-López *et al.*, Highly immunostimulatory RNA derived from a Sendai virus defective viral genome. *Vaccine* **31**, 5713-5721 (2013).
243. R. D. Bradley, D. M. Hillis, Recombinant DNA sequences generated by PCR amplification. *Mol Biol Evol* **14**, 592-593 (1997).
244. C. Wychowski, S. U. Emerson, J. Silver, S. M. Feinstone, Construction of recombinant DNA molecules by the use of a single stranded DNA generated by the polymerase chain reaction: its application to chimeric hepatitis A virus/poliovirus subgenomic cDNA. *Nucleic Acids Res* **18**, 913-918 (1990).
245. D. Veesler *et al.*, Atomic structure of the 75 MDa extremophile Sulfolobus turreted icosahedral virus determined by CryoEM and X-ray crystallography. *Proc Natl Acad Sci U S A* **110**, 5504-5509 (2013).
246. D. Mao, D. W. Grogan, How a Genetically Stable Extremophile Evolves: Modes of Genome Diversification in the Archaeon Sulfolobus acidocaldarius. *J Bacteriol* **199**, (2017).
247. B. Coutard *et al.*, The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res* **176**, 104742 (2020).
248. B. A. Johnson *et al.*, Furin Cleavage Site Is Key to SARS-CoV-2 Pathogenesis. *bioRxiv*, (2020).
249. Z. Liu *et al.*, Identification of Common Deletions in the Spike Protein of Severe Acute Respiratory Syndrome Coronavirus 2. *J Virol* **94**, (2020).
250. D. Kim, B. Langmead, S. L. Salzberg, HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360 (2015).
251. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
252. Y. Sun *et al.*, Immunostimulatory Defective Viral Genomes from Respiratory Syncytial Virus Promote a Strong Innate Antiviral Response during Infection in Mice and Humans. *PLoS Pathog* **11**, e1005122 (2015).

VITA

Stephanea Sotcheff was born March 28, 1991 in San Antonio, TX. She graduated from New Braunfels High School in 2009 and received Bachelor of Science degrees from the University of Texas at Austin (UT) in Biochemistry (2013), Chemistry (2013), and later Molecular Biology (2017). From February 2014 through August 2017, Stephanea worked as a Technical Support Engineer for Forcepoint, a software company in Austin, TX. In 2015, while maintaining her position at Forcepoint, returned to UT to complete an additional bachelor's degree while gaining laboratory experience working as an undergraduate research assistant for Dr. Mike Rose in the Inorganic Chemistry department at UT. She began graduate school at the University of Texas Medical Branch at Galveston (UTMB) in August 2017, where she began work under Dr. Andrew Routh in the department of Biochemisty and Molecular Biology – hoping to combine her experience in software with her interest in viral evolution. Her dissertation work centered on using and developing novel techniques to understand flavivirus pathogenesis, primarily focused on the increased incidence of microcephaly in infants born to infected mothers during the 2015-2016 ZIKV outbreak in S. America.

EDUCATION

B.S., December 2013, The University of Texas at Austin, Austin, TX

B.S., December 2013, The University of Texas at Austin, Austin, TX

B.S., May 2017, The University of Texas at Austin, Austin, TX

PUBLICATIONS

IN PREPARATION

1. **Sotcheff, S**.; Elrod, N.; Chen, J.; Cao, J.; Kuymuycu-Martinez, M.; Shi, P-Y.; Routh, A. Zika virus infection alters gene expression and poly-adenylation patterns in human placental cells. (*in preparation for submission to Viruses by MDPI*)
2. **Sotcheff, S.**; Zheng, J.; Stafford, S.; Routh, A.; Anastasio, N.; Cunningham, K. Acute withdrawal from fentanyl may be a co-morbidity for SARS-CoV-2 infection. (*in Preparation for submission to Nature Comunications*)
3. Swetnam, D.; Alvarado, R.E.; **Sotcheff, S.;** Mitchell, B.; Haseltine, F.; Maknojia, S.; Smith, A.; Ren, P.; Keiser, P.; Weaver, S.; Routh, A. Investigation of a SARS-CoV-2 outbreak in a Texas summer camp from a single introduction. *(in Preparation for EID)*

IN PRESS

1. **Sotcheff, S**.; Zhoi, Y.; Sun, Y.; Johnson, J.E.; Torbett, B.D.; Routh, A. ViReMa – A Virus Recombination Mapper of Next Gneration Sequencing data characterizes diverse recombinant viral nucleic acids. *BioRxiv*. https://doi.org/10.1101/2022.03.12.484090
2. Vu, M.; Lokugamage, K.; Plante, J.; Scharton, D.; Johnson, B.; **Sotcheff, S**.; Swetnam, D.; Schindewolf, C.; Alvarado, R.; Crocquet-Valdes, P.; Debbink, K.; Weaver, S.; Walker, D.; Routh, A.; Plante, K.; Menachery, V. QTQTN motif upstream of the furin cleavage site plays key role in SARS-CoV2 infection and pathogenesis. BioRxiv. 2021 Dec. https://doi.org/10.1101/2021.12.15.472450 (*in Review at Cell Host Microbe*)

3. Wang, S.; **Sotcheff, S.**; Gallardo, C.; Jaworski, E.; Torbett, B.; Routh, A. Co-variation of viral recombination with single nucleotide variants during virus evolution revealed by CoVaMa. *NAR.* 2022 Jan. https://doi.org/10.1093/nar/gkab1259
4. Jaworski, E.; Langsjoen, R.; Mitchell, B.; Barbara, J.; Newman, P.; Plante, J.; Plante, K.; Miller, A.; Zhou, Y.; Swetnam, D.; **Sotcheff, S.**; Morris, V.; Saada, N.; Muchado, R.; McConnell, A.; Widen, S.; Thompson, J.; Dong, J.; Ping, R.; Pyles, R.; Ksaizek, T.; Menachery, V.; Weaver, S.; Routh, A. Tiled-ClickSeq for targeted sequencing of complete coronavirus genomes with simultaneous capture of RNA recombination and minority variants. *eLife*. 2021 Sept. http://dx.doi.org/10.7554/eLife.68479

REVIEWS

1. **Sotcheff, S**.; Routh, A. Understanding Flavivirus Capsid Protein Functions: The Tip of the Iceberg. *Pathogens.* 2020 Jan, 9, 42. https://doi.org/10.3390/pathogens9010042
2. Zhou, Y.; **Sotcheff, S**.; Routh, A. Next Generation Sequencing: a new approach to understanding viral RNA-protein interactions. JBC. 2022 April. https://doi.org/10.1016/j.jbc.2022.101924

ABSTRACTS

1. Swetnam, D.; Alvarado, R.E.; **Sotcheff, S.;** Mitchell, B.; Haseltine, F.; Maknojia, S.; Smith, A.; Ren, P.; Keiser, P.; Weaver, S.; Routh, A. Molecular epidemiological investigation of COVID-19 outbreaks among Galveston County summer camp resulting from a single introduction, June 2021. *UTMB Public Health Symposium*. April 6, 2022. Galveston, TX, USA
2. Tat, V.; Morris, D.; Trevino, M.; Bohn, K.; **Sotcheff, S.;** Hansen, C.; Tat, N.; Rao, A.; Hanley, E.; Tat, C.; Croisant, S.P.; Hallberg, L.; Weaver, S.; Pennel, C. “Taking Our Best

Shot" Against COVID-19 Misinformation in Galveston County. *UTMB Public Health Symposium*. April 6, 2022. Galveston, TX, USA

3. Schindewolf, C.**;** Vu, M.; Johnson, B.; **Sotcheff, S.**; Routh, A.; Menachery, V. Mutation of SARS-CoV-2 nonstructural protein 16 sensitizes the virus to type I interferon and attenuates pathogenesis *in vivo*. American Society of Virology 2022. July 16-20, 2022. Madison, WI, USA.
4. Vu, M.; Lokugamage, K.; Plante, J.; Scharton, D.; Johnson, B.; **Sotcheff, S.**; Swetnam, D.; Schindewolf, C.; Alvarado, R.; Crocquet-Valdes, P.; Debbink, K.; Weaver, S.; Walker, D.; Routh, A.; Plante, K.; Menachery, V. Transcriptomic QTQTN motif upstream of SARS-CoV-2 furin cleavage stie contributes to pathogenesis. American Society of Virology 2022. July 16-20, 2022. Madison, WI, USA.
5. Silva, J.**;** **Sotcheff, S.**; Merritt, C. Routh, A.; Anastasio, N.; Cunningham, K. Transcriptomic Profiling Reveals Mesolimbic Gene Targets Associated with Oxycodone-Seeking During Abstinence. Experimental Biology 2022. April 2-5, 2022. Philadelphia, PA, USA.
6. **Sotcheff, S.**, Chen, J., Shi, P-Y, Routh, A. Zika virus alters many facets of the host transcriptome in human placental cells. *BCMB Student Research Expo*. October 28, 2021. Galveston, TX, USA. (virtual)
7. **Sotcheff, S.** Routh, A. Investigating alternative roles of capsid in flavivirus replication cycle. *American Society of Virology*. July 19-23, 2021. (virtual)
8. **Sotcheff, S.** Suggestions for Graduate Education at UTMB: Preparing Student for a Broad Range of Career Paths. *Scholars in Education Symposium*. April 20, 2021. Galveston, TX, USA.

9. **Sotcheff, S.** Suggestions for Graduate Education at UTMB: Preparing Student for a Broad Range of Career Paths. *Scholars in Education Symposium*. April 20, 2021. Galveston, TX, USA.
10. Rodriguez, R., Solomon, S., **Sotcheff, S.**, Orellana, C., Hooks, S., Hoang, T., Swetnam, D., Dann, S., Pennel, C. Students Educating Students: K-12 Activities for Public Health and Pandemic Preparedness. *UTMB Public Health Symposium*. April 7, 2021. Galveston, TX, USA (virtual – ORAL ABSTRACT)
11. **Sotcheff, S.,** Zhou, Y., Routh, A. Understanding ZIKV RNA-capsid interactions using viral photo-activatable ribonucleoside crosslinking (vPAR-CL). *American Society of Virology*. June 15, 2020. Fort Collins, CO, USA. (Accepted, conference cancelled due to COVID-19)
12. **Sotcheff, S.,** Zhou, Y., Bradrick, S., Routh, A. Understanding ZIKV RNA-capsid interactions using crosslinking and NGS technologies. *UTMB BMB Oktoberfest.* October 29, 2019. Galveston, TX, USA.
13. **Sotcheff, S.,** Zhou, Y., Bradrick, S., Routh, A. Understanding ZIKV RNA-capsid interactions using crosslinking and NGS technologies. *Keystone Symposia: Positive-Strand RNA Viruses (E2)*. June 9-13, 2019. Killarney, Co. Kerry, Ireland.
14. **Sotcheff, S.**, Zhou, Y., Bradrick, S., Routh, A. Crosslinking and NGS technologies to shed light on packaging of ZIKV RNA. *2019 IHII/McLaughlin Colloquium.* March 29, 2019. Galveston, TX, USA.

SUMMARY OF DISSERTATION

With new next generation sequencing technologies and methodologies being published often, our ability to sequence viruses and host transcriptomes (in response to viral infection) has expanded – driving virology forward. Here we show the development and use of novel methodologies to study RNA viruses from their impact on the host, to variations in viral genomes, to how context of an infection may alter disease outcomes. We have published methods such as Tiled Click-Seq (TCS) and poly-A Click-Seq (PAC-Seq) with corresponding pipelines (Virus Recombination Mapper and Differential Poly-A Cluster [*DPAC*]) to study variation in viral genomes and transcriptomic changes respectively. We have also used single nuclei RNA sequencing (snRNA-Seq) for a more granular look at transcriptomic changes using the package Seurat in R.

As the 2015-2016 Zika virus (ZIKV) outbreak in S. America was associated with development of microcephaly in infants born to expectant mothers infected early in pregnancy we wanted to study the transfer of ZIKV from mother to fetus. As this involves placental infection, we extracted total cellular RNA from ZIKV infected (or mock-infected) human placental (JEG3) cells and used it to construct PAC-Seq libraries. Subsequent *DPAC* analysis provided data on differential gene expression, alternative poly-adenylation (APA), and use of alternative terminal exons. We found that up-regulated poly-A sites (PASs) lacked the sequences for canonical poly-adenylation (AAUAAA ~20 nts upstream of the PAS or a GU region just downstream) that were found in down-regulated PASs. Here we present a potential mechanism for the large-scale APA occurring in response to ZIKV infection in JEG3 cells.

Microcephaly, and other CNS issues, can be symptoms of ZIKV infection we wanted to look at the brain as well. As opioid overdose deaths have increased in recently in the U.S. (coinciding with the COVID-19 pandemic) we investigated the potential impact of opioid use on

severity of neurological disease from RNA virus infection, looking at expression of genes involved in SARS-CoV-2 or ZIKV infection in the mesolimbic pathway, using both snRNA-Seq and PAC-Seq. Our results suggest that opioid use may exacerbate symptoms of these infections by up-regulating inflammation and down-regulating anti-viral pathways in this brain region.

CURRICULUM VITAE

NAME: Stephanea Sotcheff Date:04/19/2022

PRESENT POSITION: Graduate Assistant

The University of Texas Medical Branch

Department of Biochemistry and Molecular Biology

6.142 T.G. Blocker Medical Research Building

301 University Boulevard

Galveston, TX 77555-1061

BIOGRAPHICAL: Born: San Antonio, TX

Home Address: 12206 Beamer Rd

Houston, TX 77089

Phone: 210-854-8032 (cell)

E-mail: slsotche@utmb.edu

EDUCATION:

Aug 2017 – Present *Ph.D. Graduate Student (Candidate as of 8/2019)*

Department of Biochemistry and Molecular Biology

The University of Texas Medical Branch

Mentor: Dr. Andrew Routh

GPA: 4.0

Aug 2015 – May 2017 *Bachelor of Science in Cell and Molecular Biology*

The University of Texas at Austin

Mentor: Dr. Mike Rose

GPA: 3.78

Aug 2009 – Dec 2013 *Bachelor of Science in Chemistry and Biochemistry*

The University of Texas at Austin

GPA: 3.15

Aug 2005 – May 2009 *High School*

New Braunfels High, New Braunfels, TX

GPA: 3.96

PROFESSIONAL AND TEACHING EXPERIENCE:

TEACHING EXPERIENCE:

Jan 2020 – Present *Teaching Assistant in Biostatistics*

The University of Texas Medical Branch

Contact: Dr. Heidi Spratt

Jan 2020 – Present *Teaching Assistant in Biostatistics*

The University of Houston Clear Lake

Contact: Dr. Julianna Dean

Sept 2019 – Present *Scholars in Education – Scholar*

The University of Texas Medical Branch

Office of Educational Development

Contact: Dr. Era Buck

Sept 2018 – present *Bench Mentor*

The University of Texas Medical Branch

Community Outreach and Engagement Core

Contact: Chantele Singleton

Jan 2019 – Apr 2019 *Teaching Assistant in Molecular Biology*

The University of Texas Medical Branch

Director: Dr. Yogesh Waikar

Aug 2018 – Dec 2018 *Teaching Assistant in Biochemistry*

The University of Texas Medical Branch

Directors: Dr. Thomas Smith and Dr. Bernard Pettit

Aug 2014 – May 2015 *Head Women's Rowing Coach*

The University of Texas Crew in Austin, TX

PROFESSIONAL EXPERIENCE:

Aug 2017 – Present *Graduate Research Assistant (Candidate as of 8/19)*

Department of Biochemistry and Molecular Biology

The University of Texas Medical Branch

Dr. Andrew Routh

Jan 2018 – Feb 2018 *Graduate Research Assistant*

Department of Biochemistry and Molecular Biology

The University of Texas Medical Branch

Dr. Pei-Yong Shi

Feb 2014 – Aug 2017 *Technical Support Engineer*

Cloud Web Services

Forcepoint LLC in Austin, TX

Jul 2015 – May 2017 *Undergraduate Research Assistant*

Department of Chemistry

The University of Texas at Austin

Dr. Mike Rose

RESEARCH ACTIVITIES:

RESEARCH INTERESTS: I am interested in using and developing novel next generation sequencing methods to elucidate patterns in viral evolution and nucleic acid structure.

DOCTORAL RESEARCH: **Department of Biochemistry and Molecular Biology**

The University of Texas Medical Branch

Mentor: Dr. Andrew Routh

Field of Study: Viral Nucleic Acids

Currently my studies are focused on determining the nucleic acid sequences and capsid residues involved in flavivirus RNA packaging. One study includes the use of photo-activateable ribonucleoside -enhanced crosslinking (PAR-CL) coupled with next generation sequencing to denote capsid binding sites in Zika virus. I have investigated what transcriptomic changes are triggered by ZIKV and DENV infection. Future studies will include expressing ZIKV capsid mutants and determining the transcriptomic changes from capsid and mutants alone, then determining viral fitness of specific capsid mutations.

UNDERGRADUATE RESEARCH: **Department of Chemistry**

The University of Texas at Austin

Mentor: Dr. Mike Rose

Field of Study: Organometallic Chemistry

During my time with the Rose lab I synthesized ligands which could be metallated with manganese or iron. These ligands bind the metal center with two nitrogens and a sulfur and are termed NNS ligands. The complex is intended to mimic the active site of mono-iron dehydrogenase. The overall goal of the lab was to split water and produce energy by breaking hydrogen molecules with the hydrogenase active site complexes.

EXTRACURRICULAR ACTIVITIES:

Aug 2017 – Present	Graduate Student Organization (UTMB)	
Jun 2019 – Jun 2021	Elected Position:	President
Jun 2018 – May 2019		Vice President
Aug 2017 – Present	Biological Chemistry Student Organization	
July 2018 – Jul 2019	Elected Position:	Co-Chair
Dec 2013 – Present	University of Texas Crew Alumni Foundation	
Feb 2017 - Jul 2019	Elected Position:	President
Jul 2019 – Present		Marketing Chair
Jan 2010 – Dec 2013	University of Texas Crew Rowing Club	
May 2012 – Dec 2013	Elected Position:	Treasurer
Aug 2011 – Dec 2013		Women's Captain
May 2011 – May 2012		Travel Coordinator

PROFESSIONAL MEMBERSHIPS:

Apr 2018 – Present	**American Society of Virology,** *Student Member*

HONORS AND AWARDS:

Apr 2022	**Pathogens 2022 Best Paper Award** MDPI Pathogens
Jan 2022	**Edith and Robert Zinn Presidential Scholarship** The University of Texas Medical Branch at Galveston Graduate School of Biomedical Sciences
Dec 2021	**Barbara Bowman Scholarship** University of Texas Medical Branch at Galveston Department of Biochemistry and Molecular Biology
Sept 2021- Aug 2023	**McLaughlin Pre-Doctoral Fellowship** Institute for Human Infections & Immunity University of Texas Medical Branch at Galveston
Sept 2021- Aug 2022	**Jeane B. Kempner Pre-Doctoral Fellowship** Graduate School of Biomedical Sciences University of Texas Medical Branch at Galveston
May 2021	**Best Presentation in Microbiology** National Student Research Forum University of Texas Medical Branch at Galveston
May 2021	**Bench Mentors: Mentor of the Year** University of Texas Medical Branch Sealy Center for Environmental Health & Medicine

December 2020	**Arthur and Dorothy Barrett Scholarship** University of Texas Medical Branch at Galveston Graduate School of Biomedical Sciences
November 2020	**Irma Mendoza Award** University of Texas Medical Branch at Galveston Department of Biochemistry and Molecular Biology
October 2019	**BCSO Award** University of Texas Medical Branch at Galveston Department of Biochemistry and Molecular Biology
June 2019	**Nowinski Travel Award** University of Texas Medical Branch at Galveston Department of Biochemistry and Molecular Biology
May 2019	**Bench Mentors: Team Science Award** University of Texas Medical Branch Sealy Center for Environmental Health and Medicine

COMMUNITY SERVICE:

Dec 2020	Helped with shopping for and wrapping gifts for the Student Government Association's Holiday Toy Drive with BCSO. Also participated December 2019, 2018, and plan to participate 2021.
Mar 2019	Organized BCSO involvement with United to Serve, the university-wide service event benefiting the Galveston community.

BIBLIOGRAPHY:

Oral Abstracts:

1. **Sotcheff, S**. Chen, J. Shi, P.Y., Routh, A. Zika virus infection alters gene expression, splicing, and poly-adenylation patterns in placental cells. *National Student Research Forum*. May 14-15, 2021. Galveston, TX, USA.
2. Rodriguez, R., Solomon, S., **Sotcheff, S.**, Orellana, C., Hooks, S., Hoang, T., Swetnam, D., Dann, S., Pennel, C. Students Educating Students: K-12 Activities for Public Health and Pandemic Preparedness. *UTMB Public Health Symposium*. April 7, 2021. Galveston, TX, USA (virtual)

Poster Abstracts:

1. Schindewolf, C.**;** Vu, M.; Johnson, B.; **Sotcheff, S.**; Routh, A.; Menachery, V. Mutation of SARS-CoV-2 nonstructural protein 16 sensitizes the virus to type I interferon and attenuates pathogenesis *in vivo*. American Society of Virology 2022. July 16-20, 2022. Madison, WI, USA.

2. Vu, M.; Lokugamage, K.; Plante, J.; Scharton, D.; Johnson, B.; **Sotcheff, S.**; Swetnam, D.; Schindewolf, C.; Alvarado, R.; Crocquet-Valdes, P.; Debbink, K.; Weaver, S.; Walker, D.; Routh, A.; Plante, K.; Menachery, V. Transcriptomic QTQTN motif upstream of SARS-CoV-2 furin cleavage stie contributes to pathogensis. American Society of Virology 2022. July 16-20, 2022. Madison, WI, USA.

3. Swetnam, D.; Alvarado, R.E.; **Sotcheff, S.;** Mitchell, B.; Haseltine, F.; Maknojia, S.; Smith, A.; Ren, P.; Keiser, P.; Weaver, S.; Routh, A. Molecular epidemiological

investigation of COVID-19 outbreaks among Galveston County summer camp resulting from a single introduction, June 2021. *UTMB Public Health Symposium*. April 6, 2022. Galveston, TX, USA

4. Tat, V.; Morris, D.; Trevino, M.; Bohn, K.; **Sotcheff, S.;** Hansen, C.; Tat, N.; Rao, A.; Hanley, E.; Tat, C.; Croisant, S.P.; Hallberg, L.; Weaver, S.; Pennel, C. "Taking Our Best Shot" Against COVID-19 Misinformation in Galveston County. *UTMB Public Health Symposium*. April 6, 2022. Galveston, TX, USA
5. Silva, J.**; Sotcheff, S.**; Merritt, C. Routh, A.; Anastasio, N.; Cunningham, K. Transcriptomic Profiling Reveals Mesolimbic Gene Targets Associated with Oxycodone-Seeking During Abstinence. Experimental Biology 2022. April 2-5, 2022. Philadelphia, PA, USA.
6. **Sotcheff, S.**, Chen, J., Shi, P-Y, Routh, A. Zika virus alters many facets of the host transcriptome in human placental cells. *BCMB Student Research Expo*. October 28, 2021. Galveston, TX, USA. (virtual)
7. **Sotcheff, S.** Routh, A. Investigating alternative roles of capsid in flavivirus replication cycle. *American Society of Virology*. July 19-23, 2021. (virtual)
8. **Sotcheff, S.** Suggestions for Graduate Education at UTMB: Preparing Student for a Broad Range of Career Paths. *Scholars in Education Symposium*. April 20, 2021. Galveston, TX, USA.
9. **Sotcheff, S.,** Zhou, Y., Routh, A. Understanding ZIKV RNA-capsid interactions using viral photo-activatable ribonucleoside crosslinking (vPAR-CL). *American Society of Virology*. June 15, 2020. Fort Collins, CO, USA. (Accepted, conference cancelled due to COVID-19)

10. **Sotcheff, S.,** Zhou, Y., Bradrick, S., Routh, A. Understanding ZIKV RNA-capsid interactions using crosslinking and NGS technologies. *UTMB BMB Oktoberfest.* October 29,2019. Galveston, TX, USA.
11. **Sotcheff, S.,** Zhou, Y., Bradrick, S., Routh, A. Understanding ZIKV RNA-capsid interactions using crosslinking and NGS technologies. *Keystone Symposia: Positive-Strand RNA Viruses (E2)*. June 9-13, 2019. Killarney, Co. Kerry, Ireland.
12. **Sotcheff, S.**, Zhou, Y., Bradrick, S., Routh, A. Crosslinking and NGS technologies to shed light on packaging of ZIKV RNA. *2019 IHII/McLaughlin Colloquium.* March 29, 2019. Galveston, TX, USA.

Publications:

1. Swetnam, D.; Alvarado, R.E.; **Sotcheff, S.;** Mitchell, B.; Haseltine, F.; Maknojia, S.; Smith, A.; Ren, P.; Keiser, P.; Weaver, S.; Routh, A. Investigation of a SARS-CoV-2 outbreak in a Texas summer camp from a single introduction. *(in Preparation for EID)*

2. **Sotcheff, S.**; Torbett, B.; Routh, A. ViReMa – detection of virus recombination events using a moving seed. *BioRxiv.* https://doi.org/10.1101/2022.03.12.484090

3. **Sotcheff, S.**; Chen, J.; Shi, P-Y.; Routh, A. Zika virus infection alters gene expression, splicing, and poly-adenylation patterns in placental cells. *(in Preparation for submission to Viruses by MDPI)*

4. **Sotcheff, S.**; Zheng, J.; Stafford, S.; Routh, A.; Anastasio, N.; Cunningham, K. Acute withdrawal from fentanyl may be a co-morbidity for SARS-CoV-2 infection. (*in Preparation for submission to Nature Comunications*)

5. Vu, M.; Lokugamage, K.; Plante, J.; Scharton, D.; Johnson, B.; **Sotcheff, S.**; Swetnam, D.; Schindewolf, C.; Alvarado, R.; Crocquet-Valdes, P.; Debbink, K.; Weaver, S.; Walker, D.;

Routh, A.; Plante, K.; Menachery, V. QTQTN motif upstream of the furin cleavage site plays key role in SARS-CoV2 infection and pathogenesis. *BioRxiv*. 2021 Dec. https://doi.org/10.1101/2021.12.15.472450 *(in Review at PNAS)*

6. Zhou, Y.; **Sotcheff, S**.; Routh, A. Next Generation Sequencing: a new approach to understanding viral RNA-protein interactions. *Journal of Biological Chemistry.* 2022 April. https://doi.org/10.1016/j.jbc.2022.101924
7. Wang, S.; **Sotcheff, S.**; Gallardo, C.; Jaworski, E.; Torbett, B.; Routh, A. Co-variation of viral recombination with single nucleotide variants during virus evolution revealed by CoVaMa. *NAR*. 2022 January. https://doi.org/10.1093/nar/gkab1259
8. Jaworski, E.; Langsjoen, R.; Mitchell, B.; Barbara, J.; Newman, P.; Plante, J.; Plante, K.; Miller, A.; Zhou, Y.; Swetnam, D.; **Sotcheff, S.**; Morris, V.; Saada, N.; Muchado, R.; McConnell, A.; Widen, S.; Thompson, J.; Dong, J.; Ping, R.; Pyles, R.; Ksaizek, T.; Menachery, V.; Weaver, S.; Routh, A. Tiled-ClickSeq for targeted sequencing of complete coronavirus genomes with simultaneous capture of RNA recombination and minority variants. *eLife*. 2021 Sept. http://dx.doi.org/10.7554/eLife.68479
9. **Sotcheff, S**.; Routh, A. Understanding Flavivirus Capsid Protein Functions: The Tip of the Iceberg. *Pathogens.* 2020 Jan, 9, 42. https://doi.org/10.3390/pathogens9010042

Permanent address: 12206 Beamer Road, Houston, TX 77089

This dissertation was typed by Stephanea Sotcheff