Copyright

by:

Elizabeth Jaworski

2018

The Dissertation Committee for Elizabeth Jaworski certifies that this is the approved version of the following dissertation:

The broken genome: Merging cutting edge technologies for a molecular understanding of the Flock House virus defective interfering particle.

Committee:

Andrew L. Routh, PhD., Mentor

Eric J. Wagner, PhD., Chair

Shelton S. Bradrick, PhD.

Naomi Forrester, PhD.

John E. Johnson, PhD.

David Niesel, PhD., Dean, Graduate School

The broken genome: Merging cutting edge technologies for a molecular understanding of the Flock House virus defective interfering particle.

By:

Elizabeth Jaworski, B.S.

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

August, 2018

Dedication

This dissertation is dedicated to my family: My mom, my rock and unconditional fan, who has always supported me and shown me that with hard work anything can be accomplished. My dad, who always believed in me and pushed me to be the best that I could be. And my brother, who helped me stay sane and level headed though the hard times.

Acknowledgements

Many people have contributed to this dissertation. First and foremost, I would like to thank my mentor, Andrew Routh, who was integral in not only the science of this document but also my personal development. I could not be more thankful for his trust in me, his patience in explaining the same thing many times over, and for always keeping me grounded during the unexpected.

I would also like to thank the members of my lab both past and present; Rose Langsjoen, Tommy Zhou, Stephanea Sotcheff, Stephen Kunkel, Laura Mascibroda, and all the others that have passed through; who have not only helped me physically complete some experiments but have also provided me with a mentally supportive environment.

I am extremely grateful to the Ball High School Bench Mentors Program which has allowed me to mentor three fantastic high school students; Gabby Martinez, Jason Ju, and Yareli Perez, who have not only contributed to a significant amount of data but have also taught me slow down and appreciate the wonders of discovery.

Thank you to the Wagner lab; Eric Wagner, Nathan Elrod, and Ping Ji. They have been valuable collaborators in the development, optimization, and application of the Poly(A)-ClickSeq protocol and have graciously allowed me to use their Illumina MiSeq which was fundamental to the NGS methodology development.

I would like to thank the UTMB sequencing core; Tom Wood, Steve Widen, and Jill Thompson, who provided countless sequencing support and use of their facilities.

I greatly appreciate the Bradrick/Garcia-Blanco lab members, who have graciously shared their space, resources, and support with the Routh Lab when our group was starting out and need a physical space.

I would also like to acknowledge our collaborators; Albert Heck and Tobias Wörner, from Utrecht University, who completed the native MS experiments; as well as Peter Stockley

and Daniel Maskell, from the University of Leeds, for their work on the cryo-EM models. A big thank you to Misha Sherman; while none of our work came to fruition, spent many hours at the microscope teaching me the fundamentals of electron microscopy.

Thank you to my committee members; Eric Wagner, Shelton Bradrick, Naomi Forrester, and Jack Johnson for their guidance, enthusiasm, and insightful discussions.

Lastly, I would like to thank my family, both near (Iwanowski and Lewandowski) and far (Brocki and Jaworski/Zaluski), who have always believed in me and have surrounded me with their love. Furthermore, thank you to my friends who have been my family away from home. Particularly, Geraldine, Tera, Dan, and Sammy, who throughout my time at UTMB provided unconditional friendship and support.

The broken genome: Merging cutting edge technologies for a molecular understanding of the Flock House virus defective interfering particle.

Publication No._____

Elizabeth Jaworski, PhD.

The University of Texas Medical Branch, 2018

Supervisor: Andrew L. Routh

Defective RNAs are natural versions of a viral genome that have been truncated or rearranged by non-homologous recombination. While not encoding for functional viruses, they can be amplified and co-passaged with the wild-type virus, effectively parasitizing the normal viral machinery. Some defective RNAs can replicate so successfully so as to subdue the replication of the wild-type virus, forming 'Defective-Interfering RNAs' (DI-RNAs). As a result, DI-RNAs may promote the establishment of chronic infections, may prolong the host's infectious period, and may even be exploited as antiviral therapies or vaccines. Therefore, understanding the mechanisms of how DI-RNAs are formed and what roles they play in infections is important.

I have characterized the process of defective-RNA emergence and evolution of Flock House virus in cell culture. Using a combination of short and long read sequencing technologies, 'ClickSeq' (to resolve recombination events with nucleotide resolution) and

Oxford Nanopore Technologies' MinION (to characterize full-length and defective genomes) I have characterized the step wise progression of DI-RNA formation and the species distribution of these genomes. I observed a rapid accumulation of mature DI-RNAs suggesting that intermediate DI-RNA species are not competitive and that multiple recombination events interact epistatically to confer 'mature' DI-RNAs.

These sequencing approaches have allowed me to characterize in detail the genetic makeup of a viral population, identifying samples that are predominantly defective or predominantly wild type. Therefore, I sought to understand how defective genomes affect virus particle structure, whether defective particles display any morphological defects, and if structure can impose selective pressures to the accumulation of defective genomes. Applying cryo-electron microscopy paired with native ultra-high mass spectrometry has shown that there are no structural differences of defective virus particles compared to wild-type particles suggesting that packaging mechanisms play an important role in the selection of defective genomes. Overall, these insights have important consequences for our understanding of viral RNA packaging and assembly, and of the mechanisms, determinants and limitations in the emergence and evolution of DI-RNAs in RNA viruses.

TABLE OF CONTENTS

LIST OF FIGURES ..... **XIII**

LIST OF TABLES ..... **XVII**

LIST OF ABBREVIATIONS ..... **XVIII**

CHAPTER 1: AN INTRODUCTION TO THE DEFECTIVE (INTERFERING) VIRUS **20**

Historical Perspective 21
List of Defective Interfering Particles 23
Characteristics of Defective Interfering Particles 23
Physical Properties of Defective (Interfering) Particles 24
Genome Properties 27
Mechanisms of Defective Particle Formation 29
Replication Dependent Recombination 29
Non-Replicative DI Formation 33
Other Factors and the Involvement of the Host in DI Formation 33
Interesting Observations and Mechanistic Conclusions 34
Mechanism of Interference 35
Competition for Resources 35
Induction of Host Immunity 36
Defective Interfering Particles *in vivo* 38
Role of DI-RNAs in Persistence 38
Cyclical Production of Particles 39
Natural Infections 41
Defective Particles as Therapeutics and Antivirals 42
Conclusions 42

CHAPTER 2: PARALLEL CLICKSEQ AND NANOPORE SEQUENCING ELUCIDATES THE RAPID EVOLUTION OF DEFECTIVE-INTERFERING RNAS IN FLOCK HOUSE VIRUS **44**

Introduction 44
Methods 49
Cell Culture and Virus Passaging 49
Virus Isolation and Purification 49

Short-read HiSeq Sequencing of Viral RNA 50
HiSeq Analysis and Processing 51
Long-read Oxford Nanopore Technologies' MinION Sequencing 51
ONT Nanopore Data Processing and Alignment 52
Annotation of Full-Length Defective RNAs and Recombination Events 53
Shannon Entropy Index 53
Effective MOI and Specific Infectivity 54
Results 55
Serial Passaging of Flock House virus 55
Characterization of FHV Genomic RNA with Short-read ClickSeq Sequencing 56
Recombination Profiling Reveals Emergence and Accumulation of DI-RNAs 61
Open Reading Frames are Maintained in Most DI-RNAs 65
Conservation Mapping Illustrated Emergence and Accumulation of DI-RNAs 68
MinION Nanopore Long-Read Sequencing of Flock House virus 68
Long-read nanopore data characterize defective RNAs and the correlation of deletions 72
Complex Rearrangements are Observed by MinION and Confirmed by ClickSeq 76
MinION Nanopore Sequencing Reveals the Emergence, Diversity, and Evolution of DI-RNAs 77
Specific infectivity correlates with abundance of defective RNAs 80
Conclusions 83

CHAPTER 3: STRUCTURAL ANALYSIS OF THE DEFECTIVE FHV PARTICLE ELUCIDATES SELECTIVE PRESSURES IMPOSED UPON VIRAL GENOMES 89

Introduction 89
Methods 92
Virus Culturing, Isolation, and Purification 92
Northern Blotting 92
Viral RNA Sequencing 93
Mass Spectrometry 94
Cryo Electron Microscopy 94
Transfections for Replication Kinetics 95
Results 97
Characterizations of Viral Populations by Long- and Short-Read Sequencing 97

Nucleotide Length of Defective Genomes is Preserved in Most DI-RNAs 102
Hypothesized Packaging of Defective Viral Genomes 103
Native Mass Spectrometry Indicates DI- Particles Package RNA to 'Capacity' 106
Cryo-Electron Microscopy Elucidates the Structure of the FHV Defective Particle 107
Length Dependence on the Replication of Defective Viral Genomes 108
Conclusions 114

CHAPTER 4: METHODS DEVELOPMENT: NEXT GENERATION SEQUENCING 115

Introduction 115
ClickSeq 117
Overview 117
Protocol 121
Materials 121
Methods 123
Notes. 130
Poly(A)-ClickSeq 140
Overview 140
Protocol 144
Materials 144
Methods 144
Notes. 147
Method Application 148
Polymerase Profiling with ClickSeq 149
Overview 149
Protocol 153
Materials 153
Methods 154
Method Application 160

CHAPTER 5: DISCUSSION **164**

APPENDIX A: TABLE OF VIRUSES WITH DEFECTIVE (INTERFERING) PARTICLES **174**

APPENDIX B: CHAPTER 2 SUPPLEMENTAL DATA **176**

APPENDIX C: COPYRIGHT PERMISSIONS **180**

REFERENCES **182**

VITA **200**

LIST OF FIGURES

Figure 1.1: Electron micrograph of Vesicular Stomatitis virus.....25

Figure 1.2: Types of defective viral RNA genome arrangements 27

Figure 1.3: Proposed models for DI-RNA formation.....31

Figure 1.4: Pathway of defective viral genome interference by interferon (IFN) activation. 37

Figure 1.5: Oscillations of DIP particles relative to standard virus particles 40

Figure 2.1: Serial passaging of Flock House virus in *D. melanogaster*.....55

Figure 2.2: Percentage of ClickSeq reads mapping to host RNA through each passage and replicate 58

Figure 2.3: RNA recombination is characterized using RNAseq during serial passaging of FHV in cell-culture. 59

Figure 2.4: Scatter plots comparing frequency of unique recombination events found in replicate ClickSeq libraries of P7R2.....61

Figure 2.5: Nucleotide preference at the 5’ and 3’ sites of recombination junctions. 64

Figure 2.6: Open reading frame conservation during serial passaging. 65

Figure 2.7: Recombination profiling of conserved regions in the DI-RNAs. 66

Figure 2.8: Comparison of Illumina HiSeq to Oxford Nanopore MinION read data 69

Figure 2.9: The frequency of deletions in the FHV genomic RNAs found by MinION nanopore sequencing. 71

Figure 2.10: Error rate of the Oxford Nanopore MinION. 73

Figure 2.11: Graphical model of a complex rearrangement of a DI-RNA1. 76

Figure 2.12: Evolutionary pathways of full-length RNA genomes. 79

Figure 2.13: TCID50, specific infectivity and Cytopathic effect (CPE) of each passage for replicate 2 81

Figure 3.1: Characterizations of DVGs in various FHV populations 98

Figure 3.2: Conservation HeatMaps and most common DI- Genomes 101

Figure 3.3: Histogram of defective genomes nucleotide lengths. 102

Figure 3.4: FHV particle packaging models and expected masses. 104

Figure 3.5: Ultra-high mass native MS of defective particles. 106

Figure 3.6: Symmetrical reconstruction of the defective FHV particle.....107

Figure 3.7: Radial density of defective FHV particles.....108

Figure 3.8: Experimental design to determine how genome length influences replication 109

Figure 3.9: Defective genomes of varying lengths are replicated by FHV RdRps 111

Figure 4.1: Prototypical click-chemistry reactions.....118

Figure 4.2: Examples of triazole-linked nucleic acids 119

Figure 4.3: ClickSeq Protocol Schematic.....120

Figure 4.4: Gel electrophoresis of a final cDNA library.....129

Figure 4.5: Optimizations of the ClickSeq protocol: DMSO concentrations.....132

Figure 4.6: Optimizations of the ClickSeq protocol: Cu(II) Additions 133

Figure 4.7: Optimizations of the ClickSeq protocol 135

Figure 4.8: SPRI bead fragment size selection.....137

Figure 4.9: Reproducibility of ClickSeq to identifying FHV recombination events within the same sample. 139

Figure 4.10: Schematic overview of Poly(A)ClickSeq (PAC-seq) 143

Figure 4.11: PAC-seq Optimizations.....147

Figure 4.12: Schematic overview of 3’ end sequencing.....152

Figure 4.13: Mapped polymerase pausing sites in FHV infections. 163

Appendix B.1: ClickSeq conservation maps of all passages and replicates 177

Appendix B.2: Stacked-area plot of showing the pathways of FHV DI-RNA evolution.....178

Appendix B.3: Scatterplots showing live cell gating. 178

Appendix B.4: Raw virus recombination events data from *ViReMa* analysis of ClickSeq data. 178

Appendix B.5: Genomes characterized by MinION nanopore sequencing. 179

Appendix B.6: Accession numbers for all raw data files.....179

LIST OF TABLES

Table 2.1: Mapping of RNAseq reads.....56

Table 2.2: Five most common events in each genomic RNA in the final passage of each replicate 63

Table 2.3: Mapping of nanopore data to the FHV genome using BBMap.....75

Table 3.1: Read mapping with ONT's MinION sequencer. 97

Table 3.2: Read mapping with ClickSeq 100

Table 4.1: Sensitivity of ClickSeq in identifing nucleotide errors.....138

Table 4.2: Mapping of 2PC-seq reads. 162

Appendix A: RNA viruses with known defective (interfering) particles.....175

LIST OF ABBREVIATIONS

2PC-Seq	Polymerase Profiling with ClickSeq
3'NT-seq	3’ ends of native transcripts sequencing
3'RE	3' Response Element
4sU	4-thiouridine
5'SL	5' Stem Loop
APA	Alternative polyadenylation
BMV	Brome mosaic virus
bp	Base pair
BRIC-seq	BrU-immunoprecipitation chase-deep sequencing
BrU	5'-bromo-uridine
BVDV	Bovine viral diarrhea virus
CPE	Cytopathic effect
CuAAC	Copper-catalyzed alkyne-azide cycloaddition
DI	Defective interfering
DIP	Defective interfering particle
DI-RNA	Defective interfering RNA
DRB	5,6-dichlorobenzimidazole 1-beta-D-ribofuranoside
DSCE	Distal-Subgenomic Control Element
DSE	Downstream sequence element
DVG	Defective viral genome
EM	Electron microscopy
FHV	Flock House virus
GRO-seq	Global run-on-sequencing
HCV	Hepatitis C virus
HDV	Hepatitis D virus
HIV	Human immunodeficiency virus
hpi	Hours post induction
IFN	Interferon
intRE	Internal response elements
LDL	Low-density lipoproteins
MOI	Multiplicity of infection
MS	Mass spectrometry
NB	Northern blot
NET-seq	Native elongating transcript sequencing
NGS	Next Generation Sequencing
nt	Nucleotide
ONT	Oxford Nanopore Technologies
PAC-seq	Poly(A)-ClickSeq

PAS	Polyadenylation signal
PRO-seq	Precision nuclear run-on-sequencing
PSCE	Proximal-Subgenomic Control Element
PTGS	Posttranscriptional gene silencing
QE-UHMR	Q Exactive Plus ultra-high mass range spectrometry
RdRp	RNA-dependent RNA polymerase
RISC	RNA-inducing silencing complex
RLR	RIG I-like receptors
RNAi	RNA interference
RNAP	RNA polymerase
RSV	Respiratory syncytial virus
SPAAC	Strain-promoted Azide-Alkyne Cycloaddition
TBSV	Tomato bushy stunt virus
UTR	Untranslated region
ViReMa	Viral Recombination Mapper
VLDL	Very low-density lipoproteins
VSV	Vesicular Stomatitis virus

CHAPTER 1: AN INTRODUCTION TO THE DEFECTIVE (INTERFERING) VIRUS

"Big fleas have little fleas upon their backs to bite 'em,

And little fleas have lesser fleas, and so, ad infinitum.

And the great fleas, themselves, in turn, have greater fleas to go on;

While these again have greater still, and greater still, and so on."

-Augustus De Morgan, in 'A Budget of Paradoxes'

The concept that viruses themselves, often thought of as parasites, have parasites of their own is an idea that has been explored almost as long as humans have been studying viruses. As early as the late 1940s Preben von Magnus had first described the curious phenomenon of the paradoxical effect where applying high concentrations of viral stock produced lower yields of virus during influenza infections[1]. For many years von Magnus studied this phenomenon, but it wasn't until 1970 that these virus types were named as defective interfering particles (DIPs) by Huang and Baltimore[2, 3].

Defective viral particles are viruses that do not contain the full length or wild type genome and were originally thought to interfere with the viral infection. They are termed as such because they do not contain fully functioning genomes, most commonly created by deletions, and therefore are defective and without the capacity to code for all viral proteins. They also rely on the full length (also called helper) virus to help them replicate. Defective particles can be interfering when they attenuate the viral infection caused by the parent genome – giving rise to the specific term: "Defective-Interfering Particles" (DIPs). While broadly accepted as such, recent

studies are showing that defective genomes are not necessarily interfering as some studies have speculated that they can actually promote specific stages within the viral lifecycle[4].

Since their description over 60 years ago, the field investigating the function, generation, and use of DI-particles has moved very slowly compared to the field of virology. For a long period of time, defective particles were thought to only be an artifact of high-titer cell culturing. With advancements in molecular techniques it is becoming apparent that DIPs are more common than previously acknowledged. Particularly, improvements in high throughput sequencing technologies have allowed us to observe and characterize defective genomes, not only in the laboratory but also in natural settings. As the number of studies of defective interfering viruses has been increasing, we have learned that there are many different kinds of defective genomes, but the mechanism of their generation and their role in nature still remains elusive. Understanding how and why these genomes are formed gives us important insights into many aspects of a virus's life-cycle and can lead us to develop better treatments against virus borne diseases.

HISTORICAL PERSPECTIVE

The first account of defective viruses was most likely in 1943 by Friedewald and Pickels while they were studying high speed centrifugation sedimentation of influenza virus[5]. There, they discovered that virus from infected allantoic fluids of a chick embryo sedimented at two different constants. Upon further investigation it appeared that the slower-sedmenting population was non-infectious when applied to red blood cells[5, 6]. At the time Friedewald and Pickels believed this was due to the virus undergoing disintegration. Over the next decade a handful of other groups had also provided evidence of defective virus particles[7, 8]. Most notably was the work done by Prebus von Magnus, where in a four part series called "Propagation of the PR8 strain on

influenza A virus in chick embryos," he examined the effect of growth conditions on viral replication, the presence of incomplete virus, and the properties of this noninfectious virus in terms of Influenza PR8[9-12]. In these first papers, von Magnus found that applying high inoculation doses of influenza into chicken embryonic eggs yielded high levels of non-infectious particles, deduced by the level of infectivity to hemagglutinin (a surface protein of influenza). He also found that extended growth and serial passaging resulted in the decrease of the infectious particles to hemagglutinin ratio. This resulted in an infectious curve similar to that of bacteria. With these findings he suggested that the viral population was becoming non-infectious or, degrading[9]. It wasn't until 1954 when von Magnus finally gave this phenomenon a name by termed these particles 'incomplete' when he wrote a review encompassing not only his papers, but evidence he found in the existing literature[2]. At this time, von Magnus had already believed that defective viruses could have large implications in virological studies and could potentially be used as therapeutics or in vaccines.

In the following years, von Magnus' work led other scientists to look for defective particles or noninfectious particles following 'the Von Magnus phenomenon' in their virus preparations. Studies began to elucidate that 'von Magnus particle' production was dependent on the ratio of infectious dose to noninfectious particles and their production was a process of viral reproduction[13, 14]. During the 1950s and 60s, incomplete particles had been identified in a variety of RNA viruses including Rift Valley fever virus[15], Vesicular Stomatitis virus[16-18], Sendai virus[19], and lymphocytic choriomeningitis virus[20], to name just a few. Up until this time the origins of these particles was not very clear but by the late 1960s, reports were published that began to correlate the ratio of noninfectious particle to the presence of smaller genomic RNAs[21, 22]. Finally by 1970, enough evidence accumulated and these particles were officially coined 'defective interfering particles' or DIPs by Alice Huang and David Baltimore[3]. They went on to define the characteristics

of these particles and speculated their impact and significance during infections[3, 23]. For the next two decades, DIPs were identified in even more RNA viruses and it became evident that their presence was correlated with the establishment of persistent infections. By the 1990s, it was clear that almost all RNA viruses could produce defective particles and importantly, were present during natural infections of important human pathogens.

LIST OF DEFECTIVE INTERFERING PARTICLES

Defective interfering particles have been observed in almost all classes of virus. While these include both DNA and RNA viruses, the remainder of this review will mainly focus on defective RNA viruses, as defective genomes of DNA viruses are thought to play less of a role in their natural infection and are outside of the scope of this discussion. See **Appendix A** for a comprehensive table of defective RNA viruses. For reviews on defective DNA particles see works published by Rapp[24], Huang[23], or Patil *et al.*[25].

CHARACTERISTICS OF DEFECTIVE INTERFERING PARTICLES

The defect in 'defective particles' lies in the loss of part(s) of the genomic material in that particle. As originally defined by Alice Huang in her 1973 review[23], they are described to have the following characteristics: 1) their genomes are generated from the genome of the wild type virus; 2) they use the native structural proteins generated by the parental virus; 3) are replicatively incompetent unless in the presence of competent virions; and 4) reduce yields of the standard virus in co-infected cells. By 1986, two more properties were added by Barrett and Dimmock[26]: 5) during co-infection with standard virus, the relative amount of defective interfering virus is enhanced; and 6) they produce functional nucleic acids involved in interference. While broadly these properties hold true, the growing amount of research on this topic is changing and shaping our understanding of the characteristics of defective viruses. For example, studies have alluded

to the idea that the presence of defective particles could actually enhance the parental virus[4]. Furthermore, as we explore the function that Defective Interfering RNAs (DI-RNAs) play in natural infections we can see that they can play multiple, seemingly contradictory roles, such as immuno-stimulants and players in establishing persistence[27]. Therefore, these new findings are pushing the community to drop the word 'interfering' and instead term these viruses as containing 'defective viral genomes' or DVGs[28]. For the remainder of this document these terms will be used interchangeably.

It is important to note that due to these properties, DI- viruses are separate entities from other parasitic viruses such as satellite RNA viruses. Broadly, satellite viruses are viruses that are associated and dependent on another virus but contain extra RNA that is encapsidated by novel capsid proteins[29]. Interestingly, accounts have been observed that satellite viruses themselves can also be subject to defective RNAs[30].

PHYSICAL PROPERTIES OF DEFECTIVE (INTERFERING) PARTICLES

The second property of DIPs dictates that a defective particle must use the natural structural proteins provided by the wild type virus. Therefore, they will be antigenically identical and structurally be generally indistinguishable from the wild type virus[31]. In 1961, Eva Reczko[32] was able to show the earliest and most striking example of a visual difference in defective interfering virions of the significantly shorter versions of Vesicular Stomatitis virus (VSV) DIPs, as shown in **Figure 1.1**[33]. As studies have later shown that this type of particle shortening is expected due to the packaging of a shorter defective genome, as the amount of nucleocapsid corresponds directly to the amount of packaged RNA. Overall, the amount of variation between particle appearances mainly relies on the capsid classification of that virus. RNA viruses can be divided

into three categories based upon their capsid structures: naked icosahedral, icosahedral enveloped, and helical enveloped (as defined by Principles of Virology[34]).

Helical enveloped viruses include the families Rhabdoviridae (VSV) and Orthomyxoviridae (Influenza). Broadly, these viruses contain RNA that is tightly surrounded by nucleoproteins and further encapsidated into an envelope membrane. In this group of viruses, the largest structural variation between defective and wild type particles can be seen. This can be partially attributed to the mechanism of particle formation. As shown in **Figure 1.1,** with VSV, viral particles are assembled when capsid and the structural proteins progressively bind around the RNA genome as the virus buds out of the host lipid membrane[33, 35]. Electron microscopy images of infectious hematopoietic necrosis virus (*Rhabdoviridae*) also show severely truncated DI particles in persistently infected rainbow trout[36].

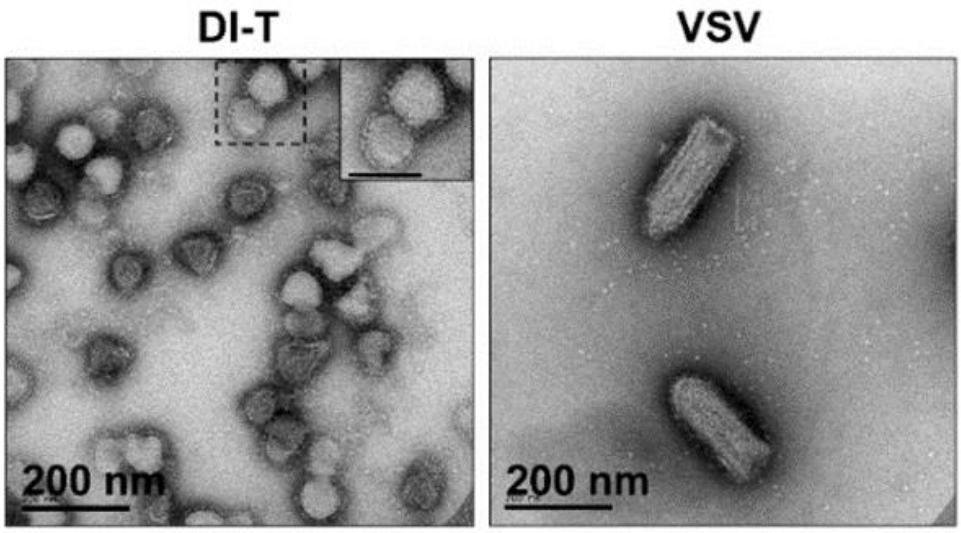


Figure 1.1: Electron micrograph of Vesicular Stomatitis virus

Electron micrograph of wild type VSV particle (right) and DI particle (left) negatively stained with PTA. Wild type VSV particles exhibit the prototypical bullet shape while DI virions are shorter in length. (Adapted from Cureton *et al.,* 2010)[33]

Conversely, for viruses such as influenza (Family: *Orthomyxoviridae*), where genomes bud out of the host membrane, negative stain electron microscopy has shown that influenza DIPs are actually larger in size than their wild type counterpart[37]. Upon analysis of the membrane components, Blough and Merline discovered that these DIPs shifted towards the incorporation of short-chain acids, long-chain polyunsaturated fatty acids, mono-glycerides, and di-glycerides which cause an increase in the internal fluids of the particles.

For icosahedral enveloped viruses, the viral genome is protected by an internal icosahedron protein capsid which is in turn enveloped by a lipid membrane. This includes virus families like Flaviviridae and Togaviridae. In a study conducted by Prince *et al.* on hepatitis C (HCV) defective virions, they alluded to the idea that DIPs select for low-density lipoproteins (LDL) conversely to wild type particles which were associated with very low-density lipoproteins (VLDL)[38]. Upon examination, they saw that DI- particles were 1.5-1.75x smaller than their standard counterparts (30-40nm and 60-70nm, respectively) while still presenting the same antigens on their surface. Interestingly, while there is a discrepancy in lipoprotein selection between influenza and HCV, the size differences seem to correlate to their respective preference (as VLDLs have high percentages of triglycerides). Furthermore, density studies conducted on Semliki Forrest virus and Rubella virus (Family: Togaviridae) have indicated that DIPs have different particle densities than their wild type counterparts[39, 40].

Non-enveloped icosahedral viruses comprise an encapsidated viral genome and an outer capsid protein(s). While structural characterizations of DIPs in this group of viruses is much more sparse, negative stain electron microscopy studies of DIPs in reovirus[41], tobacco ringspot virus[42], and Foot-and-mouth disease virus[43] have all indicated that there is no visible size differences amongst particles even though RNA analysis indicates nucleotide deficiencies.

GENOME PROPERTIES

One of the key features of RNA genomes in defective interfering particles is that they lack portions of their genetic material. While this normally is characterized as deletions of nucleic acid sequences, DI-RNAs can also arise as a product of insertions and various types of genomic rearrangements that disrupt the normal viral open reading frames[44]. Several types of genomic sequence arrangements have been identified in defective viruses as shown in **Figure 1.2**.

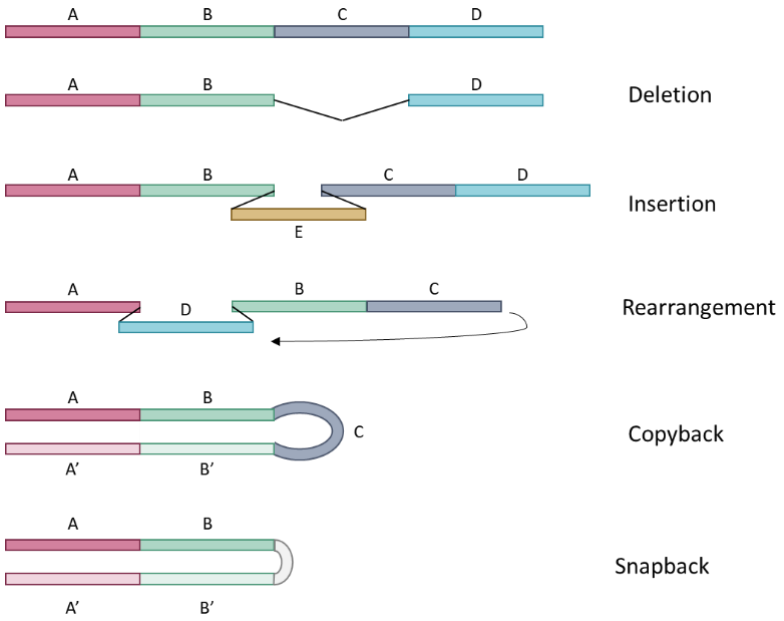


Figure 1.2: Types of defective viral RNA genome arrangements

Schematic diagrams of the most common viral genomes within defective interfering particles of ssRNA viruses. Genomes can be compounds of the depicted types.

Overall analysis of DI-RNA genomes in different viruses seem to indicate that the locations of deletions is not random. The fact that defective RNAs efficiently replicate in the presence of helper viruses indicates that these genomes must at least retain the basic essential characteristics required for replication. Indeed, the prototypical DI-RNAs generated during Flock House virus (FHV) or Tomato bushy stunt virus (TBSV) infections conserve important regulatory and

replication elements[45, 46]. For FHV, deep sequencing analysis has revealed that the exact boundaries of the DI-RNAs vary slightly while the regulatory elements are preserved[45]. It is clear that 5'UTR, 3'UTR, and internal control element deletions are not frequently observed. Studies conducted on alphaviruses [47, 48], VSV[49, 50], parainfluenza[51], and influenza[52] have also indicated that although a specific virus produces many different variations of their defective genomes, at the bare minimum they all conserve the replication elements. While these studies do not preclude the formation of defective genomic variants that excise or modify these regions, it is unlikely that these genomes would be able to be replicated or be packaged as this would disrupt the functional RNA motifs that lie in those regions[53, 54].

It is important to note that up until the recent advancements in sequencing technologies, the techniques by which defective viruses and their genomes have been characterized has generally been limited. Before the development of Next Generation Sequencing (NGS, also known as high throughput or deep sequencing), denaturing sequencing gels and Sanger sequencing were the preferred method for identifying structural genetic variations amongst defective genomes. While relatively accurate, these methods are unable to capture the full range of DI- diversity and have their own intrinsic limitations. Specifically, they only provide a restricted view of the DI-RNAs that can be easily separated and/or the most common defective virus genomes. Furthermore, classical analysis relied on ultracentrifugation steps to purify particles which could only be applied to viruses that contained DIPs of varying densities or by post gel electrophoresis nucleotide extractions. These methods provided only limited snapshot views of the potential DI-genomes that could be present in a population. Recently, a handful of groups have begun applying deep sequencing techniques to characterize DI-RNA genomes and are discovering new insights into the mechanisms of their formation, their roles in infections, and even identifying completely new species of defective genomes[50-52, 55].

MECHANISMS OF DEFECTIVE PARTICLE FORMATION

While it is clear that DI-RNAs are formed during viral infections, the exact mechanism behind their formation is still widely disputed. Viral RNA replication is carried out by the RNA-dependent RNA polymerase (RdRp) which also can replicate DI-RNAs. Currently, there are numerous models and hypotheses for how defective genomes are produced and interestingly, as early as 1959, Schafer[56] had already suggested the three main hypotheses for defective genome formation: 1) DIPs represent a step in the synthesis of new virus particles, 2) they are broken down infectious particles, and 3) they are side products of abnormal viral synthesis. Here I will explore the various theories.

REPLICATION DEPENDENT RECOMBINATION

The most broadly accepted model for DI-RNA formation is a template switching mechanism driven by the polymerase. The RNA dependent RNA polymerase (RdRp) is an inherently error-prone polymerase due to its lack of proofreading capabilities and, in terms of DI-formation, this is believed to be the likely property responsible in their generation. This can be driven through 'copy choice' recombination. In copy choice recombination, during replication the polymerase jumps from one template (donor) to another template (acceptor) while still attached to the elongating nascent chain (**Figure 1.3A**).

The formation of new viral species can be attributed to homologous recombination through template switching. While homologous recombination results in full length genomes, the same principles can be applied to non-homologous recombination (or micro-homologous recombination), which can result in complex genomic rearrangements including deletions and insertions. For template switching to occur the nascent RNA acts as a 'primer' to reinitiate replication on an acceptor template. A variety of factors can influence the rate at which template

switching can occur. These include the kinetics of replication, secondary structures (such as strong hairpin structures), and/or sequence patterns.

Evidence for replicase-driven template switching is abundant. Cell free assays of TBSV[57], Brome mosaic virus (BMV)[58], dengue virus[59], and Bovine viral diarrhea virus (BVDV)[60] have suggested that viral RNA can form recombinant species only in the presence of the RdRp. Further examination of these viral systems have indicated that short (2- to 5- nucleotide) complementary sequences between the elongating chain and acceptor template were able to re-prime and initiate replication[61].

Furthermore, a forced template-switching mechanism has also been proposed (**Figure 1.3B**). This model suggests that the polymerase jumps templates when it encounters the end of one template. For example, in TBSV, host endonucleases cleave full length viral genomes resulting in genomic fragments. The polymerase initiates replication on these fragments and when it reaches the premature 5' end of that temple, the polymerase can jump to a new template resulting in a head-to-tail DI-RNA dimer[62]. Similar studies in BVDV have shown that RNAs longer in length than the standard virus can be produced by the polymerase not properly terminating and jumping to an acceptor template[60]. An interesting study conducted by Monroe on DIPs of Sindbis virus indicated that some defective RNAs actually contained portions of a cellular $tRNA^{Asp}$ sequence their 5'-terminal end[63]. Although never empirically tested, they speculated that this event was a product of copy-choice synthesis.

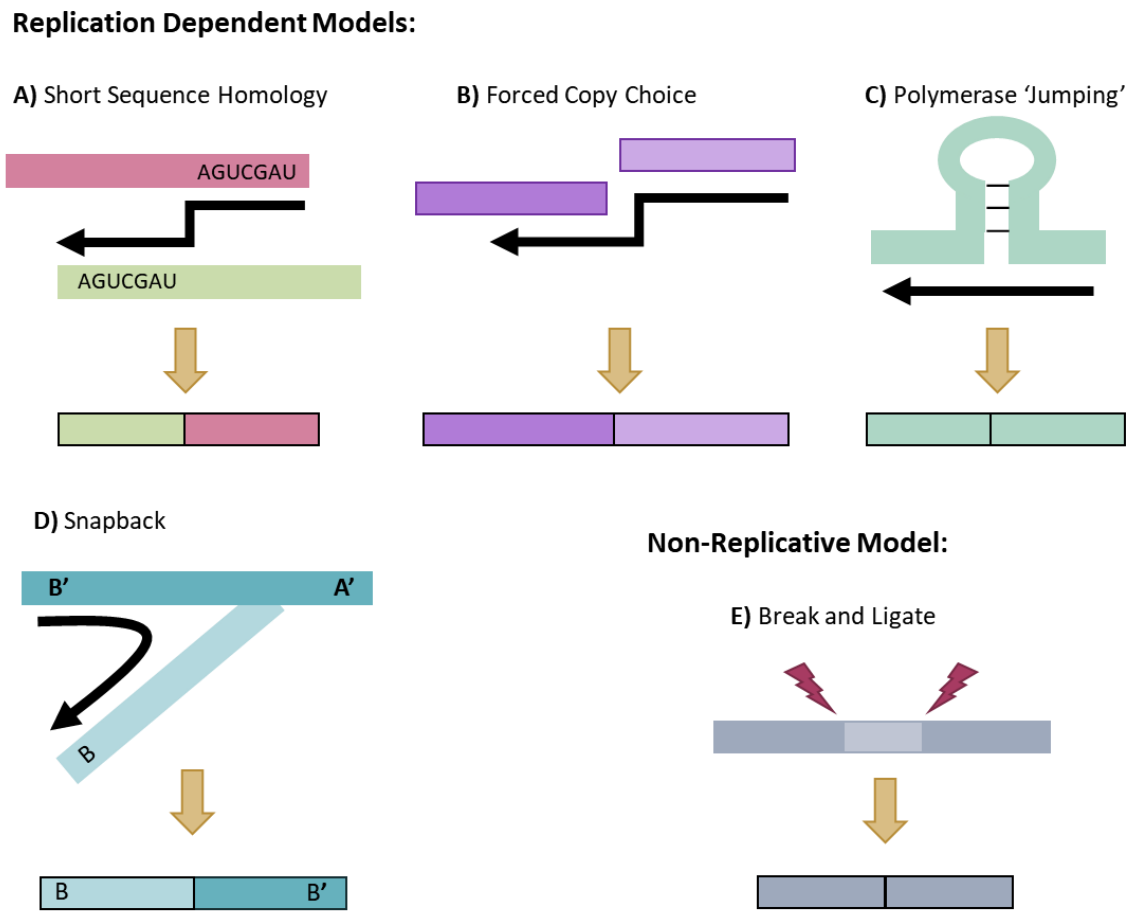


Figure 1.3: Proposed models for DI-RNA formation

Thick black line indicates pathway of the viral polymerase **(A)** In the 'copy choice' recombination model the RdRp jumps from one template (donor) strand to an acceptor strand. The jumping and re-priming is thought to be driven by short sequence homology of the elongating nascent chain and the acceptor strand. **(B)** Forced copy choice recombination occurs when the RdRp meets the end of a template and jumps to the 3' end of another template without terminating elongation. This can result in head-to-tail dimers that are normally longer than the standard virus. **(C)** Strong secondary structures have also been proposed to induce viral recombination. Here the polymerase can continue elongation bypassing tight RNA structures. **(D)** Snap back DI-RNAs are proposed to form when the polymerase jumps to another elongating stand during transcription. The polymerase then continues along on the newly synthesized chain which is of opposite strandedness than its original template resulting in a DVG of both positive and negative origin **(E)** In the non-replicative model of viral recombination, the viral genome is broken and then re-ligated with pieces of the original genome missing.

Other factors that have been proposed to influence polymerase dependent RNA recombination include the effects of viral elements such as RNA secondary structure. *Cis*-acting replication elements have been previously suggested to promote the formation of DI-RNAs[64]. In BMV, an active subgenomic promoter has been shown to support recombination[65]. Similarly, in TBSV and Turnip crinkle virus, recombination clustered around certain elements suggesting the idea of recombination 'hot spots'[57, 61, 66]. Furthermore, in certain viruses (BMV and HIV) these 'hot spots' seem to consist of AU-rich sequence stretches or unstable secondary structures of the template-elongation chain complex[67, 68]. In contrast to unstable secondary structures, extremely strong secondary structures can also induce viral recombination (**Figure 1.3C**)[69]. A very detailed study conducted on influenza virus indicated that there was a strong correlation between a deletion in the genome and the RNA secondary structure[70]. Electron microscopy images of the RNA genomes of influenza virus showed that the region of the removed RNA is present in a budding loop of the ribonucleoprotein (RNP) complex. These electron micrographs even allowed Jennings *et al.* to model how the RNA polymerase can jump across an RNA loop to generate both single and double deletion DI-RNAs (**Figure 1.3C**)[70].

For snapback and copyback DI-RNAs, it is proposed that this occurs when one polymerase catches up to another elongating polymerase and jumps across the replication fork onto that newly synthesized template (**Figure 1.3D**)[71]. The polymerase then continues transcribing the opposing stand resulting in an RNA of both positive and negative strandedness. These types of DI-RNAs more commonly occur in (-)ssRNA viruses such as VSV[72], Sendai [73], and respiratory syncytial virus (RSV)[74].

NON-REPLICATIVE DI FORMATION

Non-replicative RNA recombination is a model of break and ligate, a sort of '*virothripsis*', where viral RNAs are shattered and re-ligated to form viral chimeras (**Figure 1.3E**). The first piece of evidence for non-replicative transesterification was shown in a cell free model of the Qβ phage system[75]. Here Chetverin *et al.* were able to show that RNA fragments with overlapping 5' and 3' sequences recombined, which then could be replicated by the Qβ replicase[75]. They suggested that recombination was accomplished through a splicing-like reaction dictated by the RNA secondary structure. A similar study conducted with poliovirus was also reported[76]. In this study, multiple fragments were designed to be missing key portions of the polio genome and on their own, were non-infectious. One fragment contained the full 5'UTR which encodes for the translational control elements but lacked the polyprotein coding sequence and 3'UTR. The other fragment encodes the full polyprotein but contained mutations and deletions within its 5'UTR, therefore inactivating it. When these fragments were co-transfected, viral progeny was produced. Interestingly, not only full-length genomes were recovered but non-homologous (defective) genomes were also present[76, 77]. Similar experiments conducted in BVDV and HCV implied that homologous and non-homologous recombination can occur in the absence of the viral polymerase[78, 79]. The mechanisms behind this model have not been well explored but in both of these systems they identified viral fragments that contained 3'-phosphate and 5'-hydroxyl ends which supports the possibility of re-ligation, either through the use of cellular ligases or self-ligation[79].

OTHER FACTORS AND THE INVOLVEMENT OF THE HOST IN DI FORMATION

Currently, our knowledge of the role of host factors in viral recombination is the most limited. While DI-RNAs have been identified in a large range of different viruses it has been shown that certain cell types and hosts do not produce the same DIPs or even any at all[80]. For example,

even after 200 undiluted passages of Semliki Forest virus in a subline of HeLa cells no defective particles were produced. Conversely, the same strain of virus produced DIPs within 11 passage when infected in 5 different cell lines[81]. This, along with similar evidence in other viruses seems to indicate that the host can play a major role in DI- production[80].

The most progress on identifying host factors involved in DI- formation has been derived from the studies of TBSV and BMV infections in the model system yeast (*S. cerevisiae*). Using a variety of proteomic and genome wide screens, over 100 genes have been identified to be involved in TBSV or BMV replication, of which at least 16 play a role in either suppressing or accelerating viral RNA recombination[82, 83]. These genes include exoribonucleases (ex. XRN1), helicases (ex. DDX3 and eIF4AIII-like), and various transport proteins[62, 84]. Interestingly, different proteins within one host have antagonistic functions in the production of DI-RNAs. For example, the helicase DDX3 functions as a suppressor of viral recombination while another helicase, eIF4AIII-like, promotes it[62]. Furthermore, it was shown that these identified host factors could support both replicative and non-replicative models of recombination.

INTERESTING OBSERVATIONS AND MECHANISTIC CONCLUSIONS

RNA viruses such as FHV and TBSV contain two to three recombination events in their prototypical DI-RNA genomes. The progressive evolution of a mature DI-RNAs is also a widely debated topic. Studies in TBSV suggest that mature DI- formation occurs sequentially in a step-wise fashion[85]. Conversely, we have conducted a comprehensive experiment where FHV was passaged and sequenced at each step. Here the data failed to show the accumulation of defective genomes only containing one recombination event but instead fully formed and mature defective genomes (discussed in **Chapter 2**). This seems to suggest that mature DI-RNAs form

simultaneously where multiple reassembly and deletion events occur within one genome within a single step[86].

Overall, while many mechanisms have been proposed, it is not necessary that any of these mechanisms are mutually exclusive, whether is it template-switching, secondary structure jumping, forced copy choice, or non-replicative recombination. It is possible that DI- formation is a combination of one or more of these ideas. Furthermore, it is possible that the mechanism of formation is entirely virus (and/or host) specific and the observed defective genomes may be a result of a mixture of constraints imposed by the genomic composition, fluidity in particle structure, and/or the host's immune system.

MECHANISM OF INTERFERENCE

There is an overwhelming amount of evidence to suggest that purified DI- virus has the ability to attenuate viral infections thereby reducing disease symptoms. For example, early studies have indicated that purified DIPs have the ability to delay lethal encephalitis in mice following intracerebral inoculation of wild type VSV virus[87-89]. Similarly, intracerebral inoculation of defective influenza (H1N1) in adult mice showed a decrease in viral loads with protection from lethality[90]. DI- lymphocytic choriomeningitis virus has the same effect in young rats as well as in the lungs of mice infected with influenza[91]. While the roles that DVGs play in infections isn't very clear here I will explore some of the proposed mechanisms for interference.

COMPETITION FOR RESOURCES

During viral infections (ones lacking DI-RNAs), the wildtype virus parasitizes the necessary pools of proteins, nucleotides, and other host factors to aid in its own replication. Once in the presence of DI-RNAs, these factors have to be shared, and therefore under limiting conditions

both the wild-type and defective genomes can compete for the same resources. For DI-RNAs, this competition is in their favor. This is because DI-RNAs: 1) maintain important replication and regulatory elements allowing them to be efficiently replicated by the RdRp, 2) are smaller in size, and 3) are thought to lack translational competition[46]. For example, studies of VSV, indicate that the promoters of snapback DI-RNAs are more effective in binding the polymerase than their wild type counterparts. This is because the DI- antigenomic promoter can drive replication 20-30 times faster than its full length genomic counterpart[92]. Furthermore, their smaller size allows the replicase to transcribe their genomes quicker which results in their faster accumulation. A replication kinetic study of TBSV showed that equimolar amounts of DI- to normal RNAs suppressed the production of full length genomes by 65% in which suppression was a result of a decreased rate of wild type genome accumulation[93]. Consequently, the defective genomes outcompete the standard genomes becoming the predominant species and thereby attenuating viral symptoms.

INDUCTION OF HOST IMMUNITY

Post-transcriptional gene silencing (PTGS), also known as RNA interference (RNAi), is a biological process found in many eukaryotes that inhibits gene expression or translation. It is a response to the presence of double stranded RNA, which can be endogenous or foreign[94]. Both plants and animals can use this mechanism as an antiviral strategy[95]. Here, the RNAi machinery uses 21-25 nucleotide siRNAs to guide the PTGS complex to target and degrade viral genomes. Viruses can combat this by expressing suppressor proteins of this gene silencing system that bind to these guide siRNAs. TBSV is a plant viruses that expresses one of these gene silencing proteins, called p19[96]. Havelda *et al*. showed that the levels of virus specific siRNAs were dramatically elevated in the presence of TBSV DI-RNAs. The overproduction of siRNAs oversaturated p19 and

resulted in the accumulation of viral targeted siRNAs which in turn led to viral suppression[97]. In Flock House virus, DI-RNAs can be reverse transcribed (potentially by endogenous reverse transcriptases derived from remnants of host integrated long-terminal repeat (LTR) retrotransposons) into circular and linear viral DNAs, and it has been proposed that these DNAs serve as templates for viral siRNA production[98]. Interestingly, the production of these DI-RNA derived DNA products appears to be stimulated by the PTGS protein, Dicer-2. In RNAi, Dicer functions as the protein that cleaves dsRNA into siRNAs (through its RNase III domain) and facilitates in the activation of the RNA-inducing silencing complex (RISC). While the mechanism is still not very clear, Poirier *et al*. suggests that Dicer's helicase domain acts as an interaction point between retrotransposases and the DI-RNAs, effectively copying them into DNA[98].

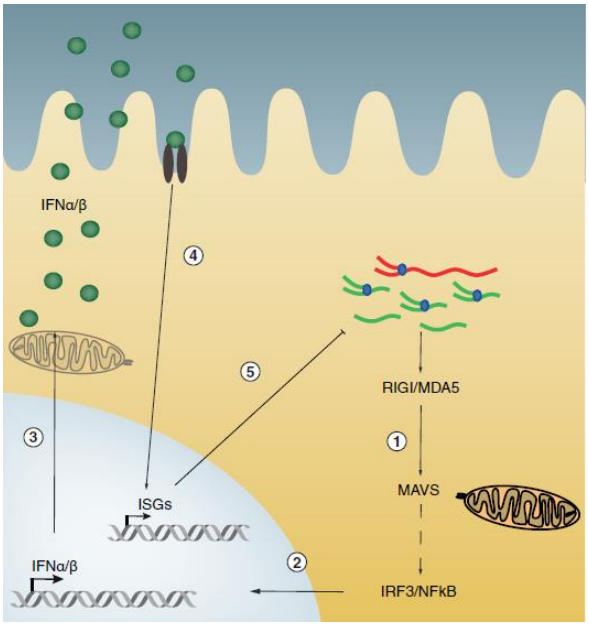


Figure 1.4: Pathway of defective viral genome interference by interferon (IFN) activation.

(1) Defective viral genomes stimulate RNA sensing receptors (i.e. RIG-I). The RIG-I receptors signal to MAVS adaptor proteins in the mitochondria that lead to the activation of the transcription factors IRF3 and NFκB. (2)The transcription factors translocate to the nucleus where they stimulate the production of IFNα/β. (3) IFN is secreted out of the cell and (4) signals in an autocrine or paracrine manner stimulating interferon stimulated genes (ISG). (5) ISGs inhibit viral replication. (Adapted from Manzoni and López, 2018)[27]

The interferon (IFN) signaling pathway is another way that the host protects itself from pathogens in vertebrates. During infections, pathogen molecules (viral glycoproteins, viral RNA, endotoxins, etc.) stimulate host recognition receptors (ie. Toll like receptors, RIG-I-like receptors (RLR), mitochondrial antiviral-signaling protein (MAVS)) which send signaling cascades that activate transcription and the production of cytokines (**Figure 1.4**)[27]. IFN is one of these proteins[99]. Studies have shown that DVGs can produce secondary structures that stimulate RLRs to activate antiviral responses[100]. In fact, for viruses like Sendai, Measles, Influenza, and Chikungunya virus, the RLR receptor preferentially binds to shorter (DI-) viral RNAs and produces a much more robust and quick response[74, 101, 102].

While DVGs are capable of stimulating immune pathways in a variety of hosts, either through IFN in mammals or the RNAi machinery of insects, the exact implications this has on viral infections is still unclear. It is speculated that these are strategies that viruses can employ to limit the extent of their own infections. This is to ensure that their host doesn't succumb to symptoms too quickly, giving the virus a chance to encounter another host. Furthermore, this could also be a way that the viruses establish persistence furthering their survival.

DEFECTIVE INTERFERING PARTICLES *IN VIVO*

ROLE OF DI-RNAS IN PERSISTENCE

Viral persistence is characterized by infections in which the virus does not clear but instead remains within the cells of their host. This can either involve stages of slow/silent replication or by latent infections[103]. While a multitude of viruses are known to establish persistent infections in human (HIV, hepatitis B virus, measles virus, etc.), an increasing number of RNA viruses, thought to only be acutely infecting, are being found to persist in humans. This

includes viruses like RSV[104, 105], Zika virus[106], chikungunya virus [107], and Ebolavirus [108], but the mechanisms behind their persistence is not well known.

Interest in defective interfering particles has stemmed not only from their very particular attenuating properties but also their role in establishing and maintaining persistent infections where as early as the 1970s this idea was postulated[3]. Overall, it has been well established that particles containing DI-RNAs can lead to persistence in cell culture[109-112], but there is not much evidence to support DVG induced persistence during *in vivo* infections. For example, Spandidos and Graham, were not only able to show that rat brains produced defective particles during the acute phase of reovirus infection but also during the chronic phase[113]. There, they were able to show that it was the presence of defective particles that helped push the rats into a persistence phase and reduced mortality[113]. Similarly, a study of Semliki Forrest virus infected in mice also showed that the establishment of persistence was also due to the presence of DVGs[114].

CYCLICAL PRODUCTION OF PARTICLES

A feature of persistent infections is a cyclical variation in viral titer during repeated passaging. Palma and Huang continuously passaged a reconstituted mixture of VSV and DI-VSV[115]. Upon examination of each of the particle concentrations they were able to show a continuous oscillating interaction between the relative proportions of the standard and DI- particles. As the relative levels of DIPs increased the levels of normal virus decreased until there was not enough 'helper' virus to perpetuate the production of DI- genomes, reducing their levels (**Figure 1.5**)[23]. This asynchronous cycling of standard virus and DIPs is thought to be the driver in the establishment of persistent infections and has been characterized in cell culture[115, 116], *in vivo*[15, 36, 116], and modeled using a predator-prey model[117].

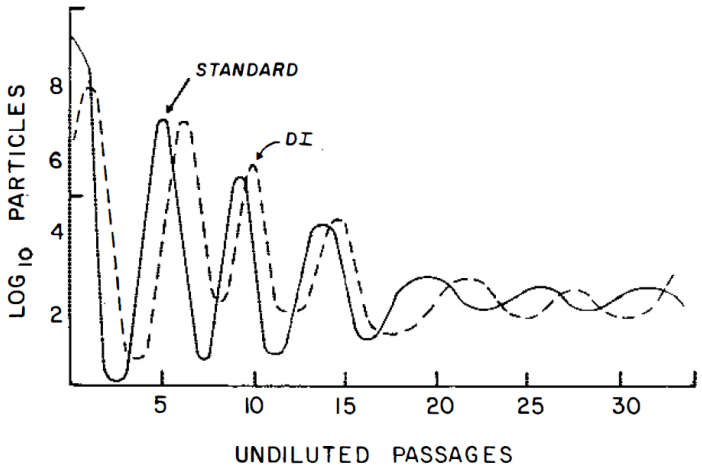


Figure 1.5: Oscillations of DIP particles relative to standard virus particles

Oscillation of the relative levels of DIPs to standard virus particles upon serial passaging. (Adapted from Huang, 1973)[23].

This apparent cycling is also perpetuated by a push-pull coevolution of sequence mutations between full length and defective genomes. In an extensive serial passaging study of VSV published by DePolo *et al.*, both DI- and standard virus accumulated mutations in their genomes[118]. Interestingly, mutations within the viral genome had effects on the ability of the DI-RNAs to attenuate infection. For example, passage 200 DI- particles had almost no attenuating effect on standard passage 287 virus. But, when applied to wild type passage 171 virus had extensive attenuating properties where even 20,000 fold more standard virus was not able to produce any plaques when in the presence of the later stage DI- population. Sequence analysis of the 5' termini of the genomic RNA from passage 287 identified at least 9 mutations and 5 bases exhibiting heterogeneity when compared to passage 171[118]. Similarly, in other studies conducted in mice persistently infected with West Nile virus and patients infected with Dengue virus showed that mutations between standard and defective virus make them less sensitive to the interference by DI-RNAs[119, 120].

NATURAL INFECTIONS

Historically, DIPs have been thought to be artifacts of laboratory manipulation. One of the major drawbacks for studying clinically relevant viruses in natural settings is the limitation in sample collection where virus isolations generally do not produce enough product to be sensitively and accurately examined. This has forced researchers into amplifying virus products in isolated mammalian systems or in unnatural animal model which can produce artifacts of laboratory manipulations. With improvement in detection technologies (such as deep sequencing) we are finding more and more evidence showing the presence of defective particles in natural infections.

It wasn't until 1989 that the first case of unequivocal isolation of defective virus from a clinical specimen was reported. Nüesch and colleages were able to collect fecal samples from three patients from independent outbreaks of Hepatitis A virus as well as viraemic blood from a transfustion associated infection, which interestingly, upon analysis, defective viral genomes were identified in all samples and were very similar to the deletions found in cell culture models[121]. Since then DVGs have been identified in a variety of virus families during *in vivo* infections, including: Measles[122], Rhabdovirus[36], Flaviviruses[120, 123, 124], influenza[52, 125], RSV[74], and Ebola[126]. Recently a study was not only able to identify Influenza DVGs in human patients but also, showed a strong correlation between the levels of defective genomes and clinical outcome (where fewer DVGs were observed in more severe/fatal patient outcomes)[125].

DIPs have more readily been identified in the study of plant and insect viruses whose culturing practices are often simpler on a large scale. In the plant species *Nicotiana edwardsonii,* Sonchus yellow net virus was isolated five months post infection and characterized. Of the collected virus, 73-86% were defective particles. Passaging of this population of virus into another

plant showed altered symptoms compared to what is seen with samples largely composed of infectious particles[127]. *De novo* analysis of DIPs with cucumber necrosis virus showed that generation of defective interfering RNAs occurred in late passages, resembling a chronic infection model[128]. There are no reports in the current literature of defective particles from wild caught insects however, DI-RNAs of insect viruses can be observed and maintained in live insects[98, 129].

DEFECTIVE PARTICLES AS THERAPEUTICS AND ANTIVIRALS

As early as 1950, Berkopf was able to show that mouse brains injected with DIPs resulted in fewer deaths during live Influenza infections[130]. Even as little as one defective interfering particle per cell has shown cytopathic protection[131]. Because of this many have proposed to use defective interfering particles as prophylaxis and antiviral treatments. Viral systems such as VSV[87, 88], Semliki Forest virus[132, 133], lymphocytic choriomeningitis virus[91], reovirus[113], among others have been explored[80]. The most progress in this field has been accomplished in influenza virus infections. For example, the defective RNA of influenza A/WSN (H1N1), called DI-244, was able to protect animals from a simultaneous dose of lethal virus when delivered to the respiratory tract. Additionally, a therapeutic effect was seen if DI-244 was administered 24-48 hours after challenge[80, 134].

CONCLUSIONS

Defective interfering RNAs are versions of a viral genome that arise naturally but have been rearranged or truncated through non-homologous recombination. They are a curious phenomenon of viral infections and have been observed in many different RNA virus infections. While it is believed that they don't code for functional proteins, they have the ability to be replicated, packaged, and passaged alongside standard viruses due to their retention of regulatory and replicative elements. Their overall characteristics result in them attenuating the

detrimental impact caused by the standard virus and have even been shown to aid in the establishment of persistent infections. Suppression of the standard virus seems counter intuitive. Why wouldn't a virus evolve mechanisms deterring the formation of attenuating species? There may be advantages in their formation as they might help limit the damage caused by the infection therefore prolonging infections and spreading the virus more. While we have learned much about DI-RNA genomes over the past 70 years it is clear that there is still so much more we need to learn.

CHAPTER 2: PARALLEL CLICKSEQ AND NANOPORE SEQUENCING ELUCIDATES THE RAPID EVOLUTION OF DEFECTIVE-INTERFERING RNAS IN FLOCK HOUSE VIRUS[†]

INTRODUCTION

RNA viruses are extremely diverse and rapidly evolving. Their RNA-dependent RNA polymerases (RdRps) readily generate single-nucleotide variants whilst lacking proof-reading capabilities[135]. RdRps are also highly prone to RNA recombination[136]; either through template-switching[137] or through non-replicative end-joining[78]. RNA recombination has been demonstrated to be responsible for the emergence of new strains or species of viruses such as rhinoviruses[138] and dengue virus[139], and the formation of vaccine-derived poliovirus[140]. Non-homologous RNA recombination is also responsible for the generation of defective interfering RNAs (DI-RNAs) as reviewed in **Chapter 1**.

DI-RNAs can attenuate viral infections via a variety of proposed mechanisms such as the saturation of the viral replicative machinery, sequestration of essential cellular cofactors, and/or induction of innate immune responses[3, 74, 141, 142]. DI-RNAs have been well characterized for a number of RNA viruses as they provide valuable tools to molecular virologists by revealing conserved regions and functional domains in the RNA genome such as binding sites for viral or host factors. Moreover, characterizing recombination loci reveal the mechanisms of recombination, impacting our understanding of viral evolution[143-145].

Until recently, due to difficulties in capturing and characterizing DI-RNAs *in vivo,* DI-RNAs were considered to be a curious epiphenomenon of cell-culturing practices[80]. As a result, our

[†]This chapter was adapted from Jaworski, E., and Routh, A. (2017). "Parallel ClickSeq and Nanopore sequencing elucidates the rapid evolution of defective-interfering RNAs in Flock House virus." *PLoS Pathogens* 13, e1006365. (PLoS Pathogens is an open access journal and applies the Creative Common Non-Commercial License)

appreciation of the diversity of DI-RNAs and the range of situations in which they could play a role was greatly limited. Increasingly, due to the use and sensitivity of Next-Generation Sequencing (NGS) technologies, DI-RNAs have been observed in a multitude of viral systems under laboratory conditions (e.g. SARS coronavirus[146], HIV[147]), in clinical settings (e.g. measles[148], dengue[123] and chronic hepatitis C[149]) and in metagenomic or 'wild' samples (e.g. West Nile virus[150], influenza virus[52]). Despite this burgeoning range of hosts for DI-RNAs, limitations in NGS technologies including high artifactual recombination rates, short reads and a limited range of bioinformatics tools tailored to viral RNA recombination discovery has hindered our ability to detect and characterize DI-RNAs in complex or clinical samples.

Flock House virus (FHV) is a positive-sense single-stranded RNA ((+)ssRNA) virus originally isolated from grass grubs in New Zealand[151] and is perhaps the best-studied *Alphanodavirus* from the *Nodaviridae* family. FHV infects *Drosophila* flies and cells in culture as well as medically important genera of insects including mosquitos, (*Anopheles gambiae)*, the tsetse fly (*Glossina morsitans morsitans Westwood*), and the Chagas vector (*Rhodnius prolixus Stal)*[152]. Infection of these organisms by FHV has been demonstrated to have similar characteristics in terms of viral titer, virus dissemination and mortality as has been shown for fruit fly infections. FHV provides an excellent model system to study (+)ssRNA virus evolution by virtue of having one of the smallest known eukaryotic virus genomes[153]. Moreover, the viral life-cycle and details of the molecular biology of virus particle assembly, cell entry, and particle disassembly are highly-characterized.

FHV contains two genomic RNAs. RNA1 (3107nts) encodes the viral RNA dependent RNA polymerase (RdRp) and RNA2 (1400nts) encodes the viral capsid protein. RNA1 also expresses a small sub-genomic region, called RNA3, that encodes the B2 protein responsible for inhibition of the anti-viral RNAi machinery[154]. FHV has been demonstrated to form DI-RNAs in multiple independent studies spanning three decades both in cell-culture[44, 45, 54, 143, 152, 155] and in *Drosophila*

melanogaster[129]. Many of these studies characterized individual DI-RNA genomes through sub-cloning and Sanger sequencing. Intriguingly, many of these DI-RNAs are highly similar. This indicates that either the DI-RNAs have emerged due to a common mechanism of formation or the presence of a common selectivity filter, or both. Our recent NGS studies of RNA recombination in FHV revealed a diverse array of RNA recombination events, suggesting that the genomic landscape of DI-RNAs is highly dynamic and likely contributes significantly to the diversity of viral genomes that form the viral quasi-species[156]. Despite these findings, studies to-date present only a single snap-shot of the DI-RNA genome landscape and do not capture the pathways of their emergence and evolution nor characterize any intermediate DI-RNA species that might arise during these processes.

In order to resolve the potential mechanisms of DI-RNA emergence and elucidate the evolutionary pathways that lead to the formation of 'mature' DI-RNAs, I performed high-titer serial passaging of FHV in cell culture and characterized the encapsidated RNA using RNAseq. Illumina HiSeq sequencing of ClickSeq generated libraries was then used to provide a high-resolution and high-confidence quantification of individual recombination events. Furthermore, I combined this information with long-read Oxford Nanopore Technologies' (ONT) MinION sequencing to resolve the topology of full-length and defective RNA genomes. By combining these data, I aimed to determine the correlation of these events within single RNA virus genomes, characterize the distribution of defective RNA genomes, and determine the exact make-up of DI-RNAs during serial passaging of FHV in cell culture.

Our lab recently developed the 'ClickSeq' method for RNAseq that uses copper-catalyzed alkyne-azide cycloaddition (CuAAC), a click-chemistry reaction, for RNAseq library synthesis. ClickSeq provides a robust platform on which to study RNA recombination in RNA viruses[157]. Artifactual recombination is a common contaminant in NGS library generation and can easily

obscure rare or non-canonical recombinant species. ClickSeq does not require template fragmentation and replaces enzymatic ligation steps commonly required in NGS library generation with click-chemistry. Therefore, this method reduces artifactual recombination to fewer than 3 events per million mapped reads[157]. As a result, ClickSeq provides a superior method for the detection of DI-RNAs and RNA recombination events. For an in-depth explanation and protocol of ClickSeq, see **Chapter 4**.

The Oxford Nanopore Technologies' (ONT) MinION is a small handheld sequencing device[158] poised to revolutionize the next-generation sequencing field by providing real-time, high-throughput and long-range (up to 2.3Mbp) sequences of DNA samples with minimal sample prep. ONT's nanopore sequencing has been used to rapidly characterize virus genomes from metagenomic samples[159], in the midst of Ebola virus outbreaks[160], and in targeted studies aimed at characterizing sequence variations within influenza virus samples[161]. Highly parallel direct RNA sequencing using Nanopore technology was also reported[162]. Due to the higher error-rate[163] of the nanopore sequencing technology compared to other RNAseq platforms, the exact identity of recombination events within single-molecule genomes may be inaccurate. However, long-read nanopore reads provide the distinct advantage of being able to sequence full-length cDNA copies of RNA virus genomes and thus can resolve multiple recombination events within a single RNA virus genome.

Here I wanted to provide a comprehensive analysis of the steps and pathways governing DI-RNA emergence and evolution starting from a plasmid-driven inoculum through to a highly-passaged sample. By combining short-read and long-read sequencing technologies, I determined both the exact identity of RNA recombination sites and their correlation within the viral quasispecies. My results show little evidence for the accumulation of intermediate defective RNA species that contain either only one, or smaller, deletions during the course of passaging. Rather,

fully formed ‘mature’ DI-RNAs that are characterized by two to three deletions between a limited number of positions in each of the FHV genomic RNAs appear after approximately 9 days of viral passaging and accumulate rapidly. The accumulation of DI-RNAs corresponded with a reduction in the specific infectivity of the viral samples in each passage. This implies that partially formed DI-RNA species are not competitive and cannot accumulate in the manner that mature DI-RNA species do, perhaps due to the epistatic interaction of multiple recombination events. Alternatively, the formation of mature DI-RNAs may occur in a single step involving multiple simultaneous genome rearrangements.

METHODS

CELL CULTURE AND VIRUS PASSAGING

D. melanogaster (S2) cells were grown at 28°C in Schneider's Drosophila Media supplemented with 10% fetal bovine serum and 1X Penicillin-Streptomycin using standard laboratory procedures. To generate initial Flock House virus inoculum, S2 cells were plated at 50-70% confluency in a six well plate and were transfected with 2.5µg of pMT plasmid containing FHV RNA1 (NC_004146) and 2.5µg of pMT plasmid containing FHV RNA2 (NC_004144) using Lipofectamine 3000 Transfection Reagent as per the manufacturer's protocol. Plasmid transcription was induced 24 hours post transfection with the addition of 50mM $CuSO_4$. Virus was then allowed to propagate for 3 days post induction. For successive passages (Passages 1-9), S2 cells were grown in T-25 flasks to 70-90% confluency (~1 x 10^7 cells), then infected with 1mL of viral inoculum from the previous passage. Virus was grown for 3 days, then fractions were harvested for viral purification or inoculation of the next passage.

VIRUS ISOLATION AND PURIFICATION

To purify virus from each consecutive serial passage, cells and supernatant were subjected to a freeze-thaw cycle in the presence of 1% Triton X-100 to release viral particles from infected cells. Virus particles were then purified on a 30% sucrose cushion by spinning the cell lysate at 40,000 RPM for 2.5 hours. Viral pellet was resuspended in 10mM Tris (pH 7.4). Virus was further purified by applying the resuspended virus atop a 10-40% sucrose gradient and spun at 40,000 RPM for 1.5 hours. The viral band was collected and subsequently treated with 1 Unit DNase and 1 Unit RNase and incubated at room temperature for at least one hour to remove any cellular nucleic acids not protected by the viral capsid. The virus sample was concentrated on a

100,000 NMWL centrifugal filter column and washed with at least 2 volumes of 10mM Tris (pH 7.4). Finally, encapsidated viral RNA was extracted using a QIAGEN RNeasy Mini Kit as per the manufacturer's protocol.

SHORT-READ HISEQ SEQUENCING OF VIRAL RNA

Next generation sequencing (NGS) libraries were generated using 100ng of RNA using the 'ClickSeq' protocol as previously described by Routh *et al.* and in detail in **Chapter 4**[157, 164]. Briefly, cDNA is synthesized through RT-PCR initiated from semi-random (6N) primers containing a partial Illumina p7 adapter (GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNN) and stochastically terminated by the addition of azido-NTPs (AzNTP) at a ratio of 1:35 AzNTP:dNTPs. Subsequently, the p5 Click-Adapter (5'-Hexynyl-NNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTC-GGTGGTCGCCGTATCATT, IDT) was click-ligated onto the azido-terminated cDNA fragment using copper-catalysed azide-alkyne cycloaddition (CuAAC) in the presence of TBTA ligand (Lumiprobe) and Vitamin C catalyst in 55% DMSO. After purifying the click-linked cDNA with a Zymo DNA clean column, 18 cycles of OneTaq (NEB) PCR amplification adds the remainder of the p7 adapter along with the desired TruSeq index sequence. PCR product was cleaned again with a Zymo DNA clean column to remove excess primers and then ran on a 1-2% precast agarose e-gel (Invitrogen, E-Gel Electrophoresis System). cDNA libraries between 400 to 700bp were excised corresponding to insert sizes of 250-550bp and cleaned using the Zymo Research Gel DNA Recovery Kit. Final cDNA libraries were quantified using a QuBit fluorimeter (Life Tech) and loaded on a HiSeq 1500 single read rapid run flowcell for 1x150 reads and 7 nucleotides of the index. FHV libraries used for the triplicate study shown in **Figure 2.4** were sequenced on a MiSeq platform with v3 chemistry for 600 cycles (2x300). Reads were trimmed to 150nts prior to analysis to emulate the libraries sequenced on the HiSeq.

HiSeq Analysis and Processing

Raw reads were processed by first removing the Illumina TruSeq adaptor using *Cutadapt*[165] with default parameters. Next, the first 6 nucleotides (corresponding to the random nucleotides and triazole-linkage included in the Click-Adaptor) were trimmed and any reads that contained nucleotides with a PHRED score <20 were removed using the *FASTX toolkit* (http://hannonlab.cshl.edu/fastx_toolkit/). The remaining reads were aligned end-to-end with *Bowtie* (v1.0.1)[166] (command line parameters: `-v 3 --best`) first to the FHV genome (NC_004414 and NC_004146) and next to the host *D. melanogaster* genome (fb5_22). The remaining unmapped reads were processed to identify recombination events using the python script '*ViReMa*' (Viral Recombination Mapper)[45] (command line parameters: `--N 1 --X 5 --Seed 25 --Host_Seed 30 --Defuzz 0 --MicroInDel 5`). The frequency of a specific recombination event is approximated by dividing the number of reads mapping to this recombination (N) by *N* plus the average of the number of reads mapping to the wild-type genome at each of the recombination coordinates.

Long-read Oxford Nanopore Technologies' MinION Sequencing

The Oxford Nanopore Technologies' (ONT) MinION and flowcells were acquired as part of the ONT early-access program. To prepare sequencing libraries for the MinION, 50ng of RNA was reverse transcribed using RNA specific primers that were complimentary to the 3' end of the respective genome (RNA1: ACCTCTGCCCTTTCGGGCTA or RNA2: ACCTTAGTCTGTTGACTTAA). cDNA was then amplified using the standard Phusion (NEB) PCR protocol using genome specific primers (RNA1_FP: GTTTTCGAAACAAATAAAAC or RNA2_FP: GTAAACAATTCCAAGTTCCA) for 19 cycles. Excess primers were removed from the PCR product using AMPure XP beads (Beckman Coulter) at a ratio of 1:1 AMPure bead:PCR product. Samples were then barcoded and prepared

following the manufacture's protocol (R9 Native Barcoding Kit I and Nanopore Sequencing Kit) with adjustments to tailor input cDNA quantities. A target of 1µg of fragmented DNA at approximately 8,000nts is considered optimal for library generation using this kit. The input amounts for RNA1 (3107bp) and RNA2 (1400bp) were thus adjusted to 192ng and 88ng respectively and combined in 46µL water to maintain optimal DNA end molarity. After ligation of barcodes, equal amount of each DNA library (9 samples in total) were pooled and loaded onto a MinION MkIB device equipped with an R9 flow cell. The MinKNOW control software was used to select a 48-hour sequencing protocol and was allowed to proceed for at least 36 hours, until high-quality data accumulation ceased. Raw data was uploaded automatically by Metrichor software for cloud-based base-calling using default settings and quality filtering for 2-dimensional reads. Reads in fastq format were extracted from HDF5 format files (fast5) using *poretools*[167].

ONT NANOPORE DATA PROCESSING AND ALIGNMENT

Full-length ONT reads were mapped to the Flock House virus genome using the pacbio wrapper from the *BBMAP* v36 suite (command line parameters: `fastareadlen=6000 vslow=t maxindel=3100 minid=0.5 local=f ignorebadquality=t usequality=f`). Alignment SAM files were visualized using Tablet sequence viewer[168]. SAM files were filtered to ensure that MinION reads mapped from the first 25nts to the final 25nts of the reference genome (accounting for deletions and insertions), due to the presence of truncated nanopore reads and mis-priming during the cDNA PCR amplification steps. Errors including substitutions, insertions and deletions were counted using the *samtools*[169] *mpileup* command and error rates at each position were calculated by dividing this value by the read depth at this position (**Figure 2.10**). Insertion and deletion events longer than 25nts were extracted using the CIGAR string of the SAM files using simple in-house scripts. For recombination sites containing 'fuzz',

where nucleotides surrounding the putative recombination events are the same for both the acceptor and donor sites, the recombination event was reported as occurring in the middle of the 'fuzzy' region, or at the 5' side of the middle two nucleotides in the orientation of the reference if the fuzzy site contained an even number of residues. This is the same methodology as employed in the *ViReMa* script[45] used to map recombination event in the ClickSeq data. Insertion events and soft-pads longer than 100nts were extracted and their nucleotide sequence was analyzed using an online BLASTn search to determine their identity.

ANNOTATION OF FULL-LENGTH DEFECTIVE RNAS AND RECOMBINATION EVENTS

To annotate defective genomes detected by MinION nanopore sequencing or recombination events detected by ClickSeq, I use underscores '_' to denote continued mapping, and carets '^' to denote a recombination events. For example, "1_317^1091_1242^2301_3107" indicates an authentic mapping from nt 1 to 317, then a deletion event removing nts 318 through 1090, then another authentic mapping from 1091 to 1242, followed by another deletion removing nts 1243 through 2300, and finally an authentic mapping from nt 2301 to 3107.

SHANNON ENTROPY INDEX

The Shannon Entropy Index is given by: $\boldsymbol{H(X) = -\sum_{i=0}^{N-1} p_i \ln p_i}$. For the ClickSeq data, each recombination event is treated as independent with its probability determined by dividing the number of reads mapping to the present recombination event divided by the average coverage over the whole viral RNA. For the MinION data, each individual read mapping is treated as an individual event with the frequency determined by dividing the number of identically mapping reads divided by total mapped reads.

EFFECTIVE MOI AND SPECIFIC INFECTIVITY

Tissue culture infective dose 50 (TCID50) analyses of the supernatants from each passage were performed using standard protocols[170]. For purified particles of each passage TCID50 was calculated with slight modifications. Specifically, 1 x 10^5 cells (S2) per well were plated in 96 well format. Virus samples were quantified by measuring the OD_{260nm}. An OD of 4.15 corresponds to 1mg virus[171], which in turn corresponds to 6.4 x 10^{13} virus particles assuming a virion mass of 9.4MDa[172]. Purified virus samples were diluted to a starting concentration of 47ng/µL, which corresponds to 3x10^{10} virus particles per 10µL. These quantities were chosen as particle-to-PFU ratios for rescued FHV has previously been reported to be 300-400 particles[152, 171]. Therefore 3 x 10^{10} particles per 10^5 cells corresponds to approximately 1000 PFUs. Eight serial 10-fold dilutions were subsequently made and added to each column of the 96-well plate (8 replicate wells per dilution), as per standard TCID50 protocols. Virus was allowed to grow for 4 days after which the number of positive wells exhibiting cytopathic effect (CPE) were counted. The TCID50 values and effective multiplicity of infection (MOI) were calculated using the Reed and Muench Calculator[170]. I further counted the total number of cells that were present in each well using a Guava easyCyte HT (Millipore) flow cytometer to provide us with quantifiable amount of cell death. 50µL from each well was diluted in 150µL PBS and injected using manufactures' protocols. The InCyt v3.1 software was used to collect data with the following parameters: collection time: 30sec; flow rate: 0.59µL/sec; FSC: 16; SSC: 25; threshold: 0. The region used to determine live cells was based on scattering features of the negative controls (wells with no virus infection) (representative gating is shown in **Appendix B.3**).

RESULTS

SERIAL PASSAGING OF FLOCK HOUSE VIRUS

D. melanogaster (S2) cells in culture were transfected with cDNA plasmids containing each of the Flock House virus genomic RNAs followed by a hepatitis D virus (HDV) ribozyme sequence. After induction, the HDV ribozyme regenerates the authentic 3' end of the positive sense viral RNA, which is thus successfully recognized by the FHV RdRp allowing the initiation of viral replication[173]. I chose to initiate replication with this method to ensure that the starting viral population would be homogeneous containing only the full-length RNAs derived from the plasmid

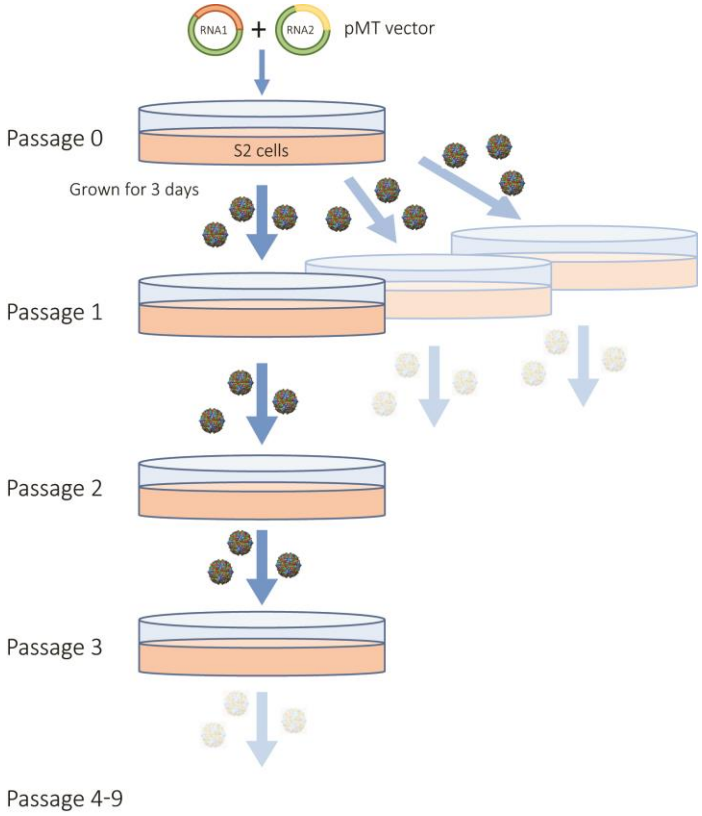


Figure 2.1: Serial passaging of Flock House virus in *D. melanogaster*

S2 cells were transfected with pMT vectors containing either FHV RNA1 or RNA2 and induced to generate genetically homogeneous viral particles. Virus was allowed to propagate for three days after which cells and the supernatant were collected. A fraction (1:10) was used to further infect a new population of S2 cells in a series of passages. The remaining fraction was collected for analysis. In total, three replicates of nine passages were collected.

cDNA. After transfection, the viral inoculum was allowed to amplify for 3 days (Passage number = P0), after which most cells exhibited cytopathic effect. Subsequently, the supernatant from infected cells was collected and a 1mL fraction (10% of the total volume) was used to infect 10mL of fresh S2 cells in triplicate (Replicates R1, R2, and R3). Again, after three days, the 1mL of the supernatant was harvested and used to infect fresh S2 cells in series for a total of nine 3-day passages (Passage numbers = P1 – P9). Therefore, one single inoculum was used to generate three distinct lineages as shown in **Figure 2.1**. For each passage and replicate, including the original inoculum, viral particles were purified over a sucrose cushion and non-encapsidated genetic material was degraded to ensure that the genetic material subsequently analyzed was packaged within the viral capsid. RNA was extracted from the purified viral particles using standard silica-based spin columns.

CHARACTERIZATION OF FHV GENOMIC RNA WITH SHORT-READ CLICKSEQ SEQUENCING

ClickSeq libraries[157] were synthesized from the purified viral genomic RNA and sequenced on an Illumina HiSeq 1500 for 1x150 single end reads. The inoculum sample was sequenced on a separate flowcell to all other samples to prevent any cross-contamination from incorrect demultiplexing. I obtained 1.2-30.6 million reads after trimming and quality filtering for each passaged sample, and 41.6 million reads for the original inoculum (**Table 2.1**). 1 million reads corresponds to an average coverage of greater than 33,000X across the FHV genome. Reads were aligned to the FHV genome and the host genome (*D. melanogaster*, fb5_22) using Bowtie end-to-

Table 2.1: Mapping of RNAseq reads

Quantity of reads generated from an Illumina HiSeq run for each passage are tabulated. Reads were mapped to either FHV or the host using *Bowtie*. Remaining reads were then processed using *ViReMa* which identifies recombination events. 'Inter-RNA' indicates recombination events between RNA1 to RNA2 or vice-versa. 'Other' indicates reads that contain unknown/ambiguous recombination events and unmapped read segments. (next page)

Replicate 1	Passage 0	Passage 1	Passage 2	Passage 3	Passage 4	Passage 5	Passage 6	Passage 7	Passage 8	Passage 9
Total Reads	41 578 802	8 425 984	18 582 914	3 030 368	6 558 293	7 333 790	7 529 709	4 919 929	8 017 787	9 084 665
FHV Genome	39 595 900	8 271 467	18 241 065	2 954 304	6 365 376	6 720 739	6 309 807	4 293 229	6 830 523	8 145 020
D. melanogaster	1 057 037	89 123	177 811	31 802	28 179	137 249	503 422	164 721	234 721	216 618
FHV Recombinations	161 205	7 578	34 573	24 517	123 637	356 277	500 760	372 253	810 863	587 347
RNA1-RNA1	2 902	3 903	13 586	3 384	55 454	269 016	444 624	330 290	788 324	569 366
RNA2-RNA2	81 765	3 558	20 603	21 097	67 686	86 553	56 085	41 575	21 602	17 236
Inter-RNA	76 538	117	384	36	497	708	51	388	937	745
Other	141 332	5 730	11 198	2 807	8 639	26 571	38 128	19 573	39 776	33 892
Unmapped	139 057	3 451	4 501	1 441	2 167	2 713	8 805	2 184	2 686	3 257

Replicate 2	Passage 1	Passage 2	Passage 3	Passage 4	Passage 5	Passage 6	Passage 7	Passage 8	Passage 9
Total Reads	11 303 746	30 630 462	15 720 784	9 438 862	9 526 140	7 118 697	4 316 239	9 249 953	8 108 519
FHV Genome	11 121 111	29 693 227	15 243 341	9 229 809	8 996 228	6 293 323	3 953 100	8 162 144	7 088 432
D. melanogaster	100 716	596 462	171 518	29 353	81 172	286 605	71 238	270 156	404 742
FHV Recombinations	9 259	46 962	192 737	126 061	377 703	409 316	246 959	679 389	437 566
RNA1-RNA1	5 103	25 052	24 809	46 225	241 048	333 125	204 609	566 092	376 893
RNA2-RNA2	3 845	20 919	167 464	79 550	136 141	75 424	42 076	111 895	59 650
Inter-RNA	311	991	464	286	514	767	274	1 402	1 023
Other	7 358	33 488	16 642	8 934	18 569	33 152	16 553	46 621	55 529
Unmapped	3 779	9 879	3 026	2 952	3 569	4 084	1 631	2 690	3 549

Replicate 3	Passage 1	Passage 2	Passage 3	Passage 4	Passage 5	Passage 6	Passage 7	Passage 8	Passage 9
Total Reads	16 420 686	17 358 462	15 283 445	7 908 511	6 784 205	6 438 793	9 369 727	1 189 217	9 080 091
FHV Genome	16 161 893	12 543 802	14 995 997	7 561 609	6 388 461	5 340 577	8 044 482	946 913	7 892 118
D. melanogaster	137 322	215 958	117 684	33 721	72 896	391 630	307 068	99 476	351 639
FHV Recombinations	13 847	25 826	70 443	238 482	244 913	507 397	767 795	82 977	485 925
RNA1-RNA1	8 022	8 212	19 213	78 822	133 208	266 278	716 381	77 638	475 383
RNA2-RNA2	5 517	17 288	50 877	159 107	111 256	240 588	50 729	5 332	10 123
Inter-RNA	308	326	353	553	449	531	685	7	419
Other	9 893	31 671	10 589	36 275	41 337	79 144	149 528	31 665	235 665
Unmapped	6 467	205 049	6 456	2 613	3 211	3 043	2 903	1 227	2 928

end mapping[166]. As expected, the majority of the reads aligned to FHV RNA1 (3107nts) and FHV RNA2 (1400nts) in a ratio reflecting the longer length of RNA1. As observed previously[156, 174], 0.3-8.4% of reads correspond to host RNAs that are encapsidated within the viral particles including mRNAs, ribosomal RNAs, and retrotransposons. Interestingly, the amount of encapsidated host RNA increases modestly through later passages (**Figure 2.2**).

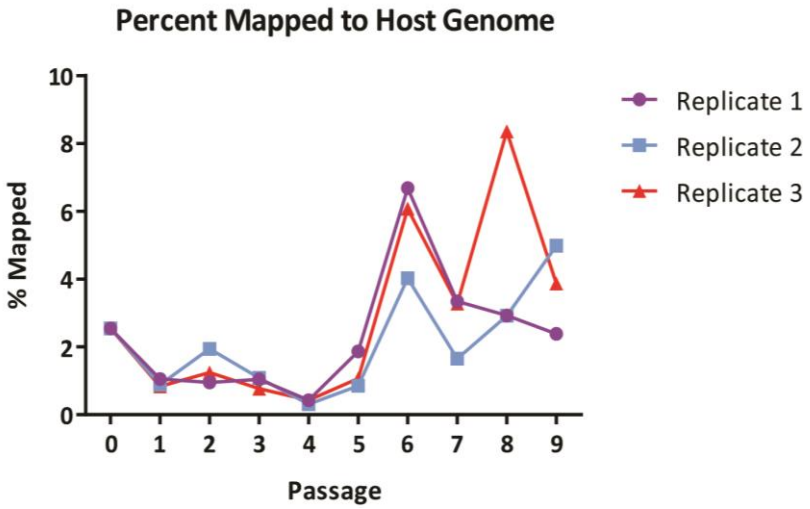


Figure 2.2: Percentage of ClickSeq reads mapping to host RNA through each passage and replicate

Percent mapping is calculated by the frequency of reads that mapped to the host genome compared to all processed reads.

Subsequently, unmapped reads were characterized with the Python script *ViReMa* ("Virus Recombination Mapper")[45]. *ViReMa* is a computational pipeline optimized for mapping virus recombination junctions in NGS data with nucleotide resolution by dynamically generating moving read segments. *ViReMa* is sensitive to many types of RNA recombination events. This includes micro-insertions and deletions (InDels comprising 5 or fewer nucleotides), duplications, deletions, inter-RNA recombination (denoting recombination between FHV RNA1 and RNA2) and virus-to-host recombination events, and reports both the identity and frequency of

recombination events. Recombination events that cannot be unambiguously identified due to unmapped read segments, mismatches occurring near to putative recombination events, or reads containing fragments of sequencing adaptors are flagged as '*other*' (**Table 2.1**)[45]. In each genomic RNA hundreds of unique recombination events were found, reflecting a diverse and complex mutational landscape (**Appendix B.4**). Broadly, there is an increase in the total number of recombination events during serial passaging of FHV and a corresponding increase in the Shannon diversity index (**Figure 2.3A** and **Figure 2.3B**).

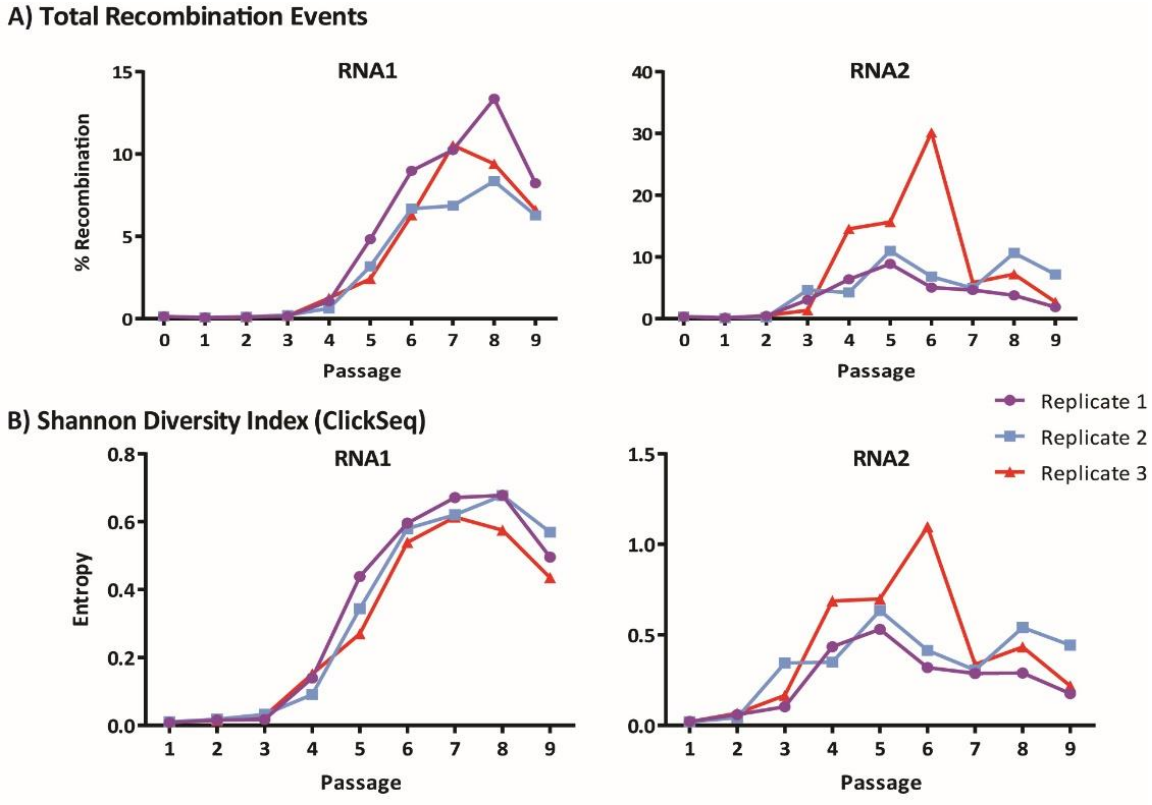


Figure 2.3: RNA recombination is characterized using RNAseq during serial passaging of FHV in cell-culture.

(A) Frequency of recombination events and **(B)** Shannon Diversity Index for all passages and replicates. Percent recombination is calculated from total number of FHV recombination events per total number of reads mapped to FHV.

Following these mapping procedures, few reads (0-1.2%, **Table 2.1**) remained uncharacterized. As in previous studies, these were found to be derived from incorrect demultiplexing of neighboring samples on the HiSeq flow-cell[175] or from contaminants in the RNAseq library generation[176]. Having accounted for almost all of the reads present in each dataset, I can be confident that this approach is capturing the full range of recombination events and/or other rearrangements present within each sample and thus are not missing important or significant events due to computational limitations.

To demonstrate the reproducibility of the ClickSeq approach and to assess the limit in terms of the ability to successfully detect rare recombination events, I generated three replicate ClickSeq libraries from the RNA sample P7R2, obtained 1x150bp reads and performed the same computational analyses as described above. I mapped 0.83M, 1.45M and 1.18M reads per library, giving an approximate coverage of ~42,000-139,000X coverage over FHV RNA1 and ~18,000-67,000X over FHV RNA2 (calculated from the average coverage over the conserved 5' and 3' ends). When comparing the frequency of unique recombination events in either RNA1 or RNA2 between any pair of the three replicates, there is an excellent correlation (Pearson > 0.99) even for very infrequent recombination events, as illustrated in the scatter plots in **Figure 2.4**. Events that were found in two replicates but not a third, never exceeded more than 20 mapped reads for RNA1 and 8 reads for RNA2. If these values are marked as a cut-offs, below which the technique fails to detect events, then a conservative estimate can be made of reproducibly sensitivity to recombinant species that are present at approximately 0.048% (20/42,000) of RNA1 population and 0.044% (8/18,000) of the total RNA2 population when obtaining ~1M sequence reads.

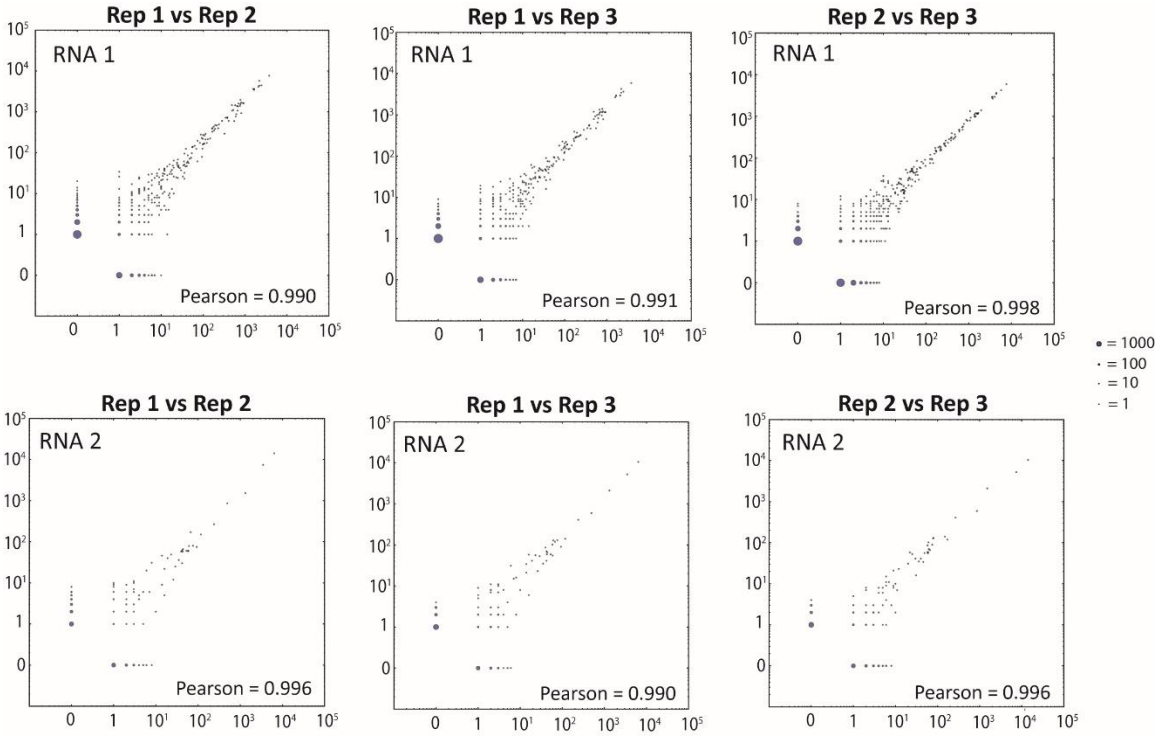


Figure 2.4: Scatter plots comparing frequency of unique recombination events found in replicate ClickSeq libraries of P7R2.

Three replicate ClickSeq libraries were generated from the same RNA sample to validate the reproducibility of ClickSeq and to determine the cut-off for sensitivity of discovery. Each point represents an individual recombination event and the x- and y- axes is the number of reads mapping to that specific event for each data set. The size of the point indicates the number of different events that share the same coordinates, as indicated by the key. These data illustrate the reproducibility that recombination events are found when multiple libraries are generated side-by-side. Pearson correlation coefficients exceed 0.99 when comparing RNA1 or RNA2 recombination between each pair of replicates.

RECOMBINATION PROFILING REVEALS EMERGENCE AND ACCUMULATION OF DI-RNAS

In the inoculum (P0), less than 0.2% of the all the reads mapped to recombination events (**Table 2.1, Figure 2.3A**). Inspection of these events reveals that they are dispersed throughout each of the genomic RNAs. RNA1 recombination events are the least frequent, with only 22 unique events detected represented by 2920 reads and without an apparent bias toward any specific location. The three most common RNA1 events in the inoculum are 441^541, 2681^2746,

and 1330^1702 with 544, 461, and 402 mapped reads respectively (**Appendix B.4**). Read depth for the wild-type genome at these loci ranges from 400K to 1.9M reads, therefore these recombination events make-up at less than 0.1% of the total viral population. These are not the events that have been previously reported as forming FHV DI-RNAs, moreover none of these events are observed again in subsequent passages, perhaps due to approaching the sensitivity of discovery limit as described above. However, due to the low-rate of artifactual recombination of the ClickSeq approach (>3 events per millions reads[157]), I can be confident that these are not sequencing artifacts[157]. Therefore, these events likely represent non-viable or transient recombination events that arose due to stochastic non-homologous recombination.

For RNA2 in the inoculum, the three most frequently observed events were 738^1219, 738^1222 and 1024^1190, with 9849, 5237, and 3673 mapped reads respectively (**Appendix B.4**). Coverage in the wild-type RNA2 mapping in these regions ranges between 2.6M and 4.8M mapped reads, therefore these recombinant species make-up approximately 0.2% of the viral RNA2 population. The majority of other recombination events are found to delete a similar region of RNA2. This region is important as it has previously been reported to be deleted in FHV DI-RNAs. However, in the inoculum I do not observe deletions upstream in the RNA2 gene (for example 250^513), also reported to be deleted in previously characterized FHV DI-RNAs. Therefore, this dataset suggests that intermediate DI-RNAs with only a single region deleted between ~740-1220 are formed very early during virus passaging (within 3 days).

In passages P1-P2, less than 1.3% of the all the reads mapped to recombination events (**Table 2.1, Figure 2.3A**). Again, these occur throughout each of the genomic RNAs. However, events that have previously been characterized as forming DI-RNAs, such as 311^1104 and 301^1100, are shown, even though these events are present at low levels (70 and 21 reads in Replicate 1 from a total of 8.2M reads mapped to the FHV genome) (**Appendix B.4**). In subsequent passages

there was a rapid increase in the total proportion of mapped recombination events (peaking at 11.9% in P8-R1) (**Table 2.1, Figure 2.3A**). In these latter passages for each replicate, it can be seen that the most common recombination events are deletions that span two regions in each genomic RNA including: nts 300-940 and nts 1240-2300 of RNA1; and nts 250-510 and nts 740-1220 of RNA2, consistent with previous observations of FHV DI-RNAs[44] (**Appendix B.4**). However, the exact sites of the recombination events, while repeatedly observed over time in each replicate, varied between replicates and each had distinct 'most popular' species in the final passages (**Table 2.2**). In some instances I was able to find that a specific event is predominant in one replicate while at low levels in another. Overall, these trends in the frequency of recombination events throughout passaging reveal that once defective RNA species emerge during viral passaging they rapidly accumulate.

Sample	RNA1 Events	Count	RNA2 Events	Count
Replicate 1	301^1100	181,226	249^517	4620
	2643^2700	150,961	1086^1175	4228
	1350^2191	132,053	247^257	2239
	1243^2309	26,928	734^1233	1152
	2545^2685	26,574	727^1229	1097
Replicate 2	313^941	36,749	250^513	22,164
	2629^2644	36,332	736^1219	21,178
	2545^2685	24,858	249^517	6342
	342^1083	24,754	223^521	1167
	1245^2514	18,430	734^1233	1012
Replicate 3	1241^2298	236,140	249^517	3236
	317^945	67,911	242^525	1865
	1241^2305	49,348	778^1219	1722
	2545^2685	28,833	738^1219	1716
	344^915	16,480	1086^1175	352

Table 2.2: Five most common events in each genomic RNA in the final passage of each replicate

The most common recombination events detected for each genomic RNA in passage 9 are indicated next to the number of reads that map over them. While similar regions are deleted, the exact recombination site varies slightly between each replicate.

Despite the predominance of certain recombination events in the later passages, there still remained a large number of infrequent events scattered throughout the viral genome. Many of these are observed only in one passage and not in subsequent passages. Again, these events likely correspond to stochastically generated RNA recombination events that form non-viable defective RNAs. Analyzing these events would more accurately reveal the nucleotide preference of the FHV RdRp for RNA recombination as they were not subject to replicative selection (although they must be packaged by FHV particles), unlike the DI-RNAs. Therefore I extracted all recombination events occurring with fewer than 10 mapped reads throughout all FHV passages and replicates (20'723 unique events from a total of 11.6M possible permutations [143]) and counted the frequency of nucleotides found both up and down-stream of 5' and 3' recombination sites in the reference genome. As shown in **Figure 2.5**, this revealed a preference for A's 1-3nts downstream of 5' sites, a preference for U's 1-3nts upstream of 5' sites, a weaker preference for a C 1nt up stream of 5' sites, and an aversion to G's 2nts both upstream and downstream of 5'

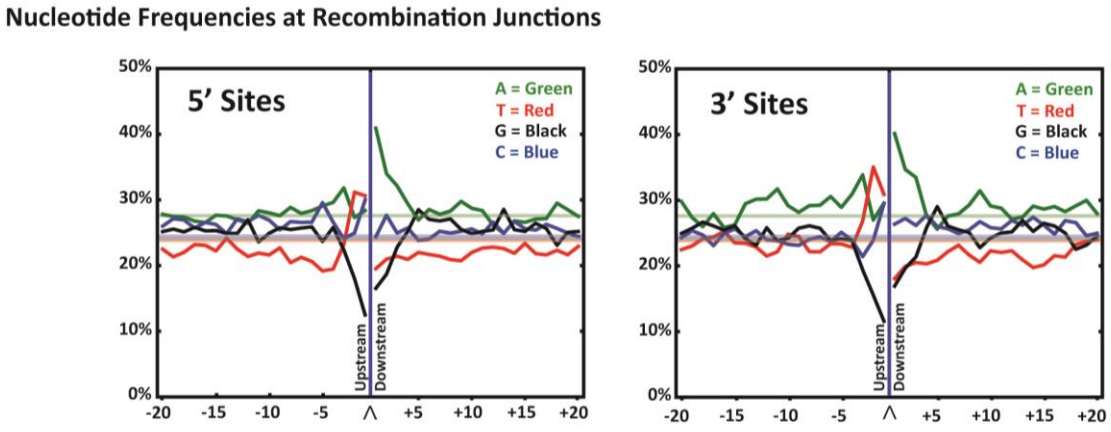


Figure 2.5: Nucleotide preference at the 5' and 3' sites of recombination junctions.

The frequency of each nucleotide found both upstream and downstream of the 5' and 3' sites of RNA recombination events are plotted. Only recombination events with fewer than 10 reads were included in this analysis to avoid over-sampling of highly replicated DI-RNAs. The composition (and therefore expected frequency) of nucleotides in the FHV genome is given by the colored horizontal lines in each plot.

sites. Interestingly, an almost identical trend was observed for the 3' sites. This trend was maintained also when analyzing only recombination events without ambiguity in the site of recombination (i.e. sites that lacked 'fuzziness' as reported by the *ViReMa* pipeline[45]). This result is similar to what has been previously reported[143]. However, here I provide a much larger dataset and analyze events that are not amplified in subsequent passages, providing greater confidence that these sites reflect the preference for RNA recombination at these sites rather than the selection of replicatively viable defective RNA species.

OPEN READING FRAMES ARE MAINTAINED IN MOST DI-RNAS

Since many recombination events resulted in deletions of the viral genome, I was curious to see if the open reading frame (ORF) was conserved, as conservation of an ORF has frequently been observed to be a property of defective and defective-interfering RNAs[177]. Moreover, it has previously been shown that cloned DI-RNAs vectors containing eGPF in their putative ORFs do indeed express fluorescent protein[178] although it is not clear whether a functional ORF is essential for DI-RNA formation or propagation. In the earliest passages, only ~33% of deletions removed a

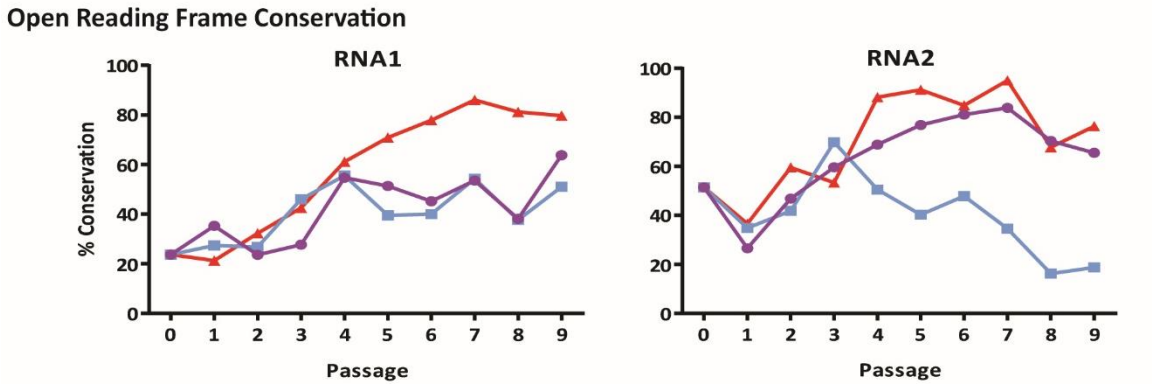


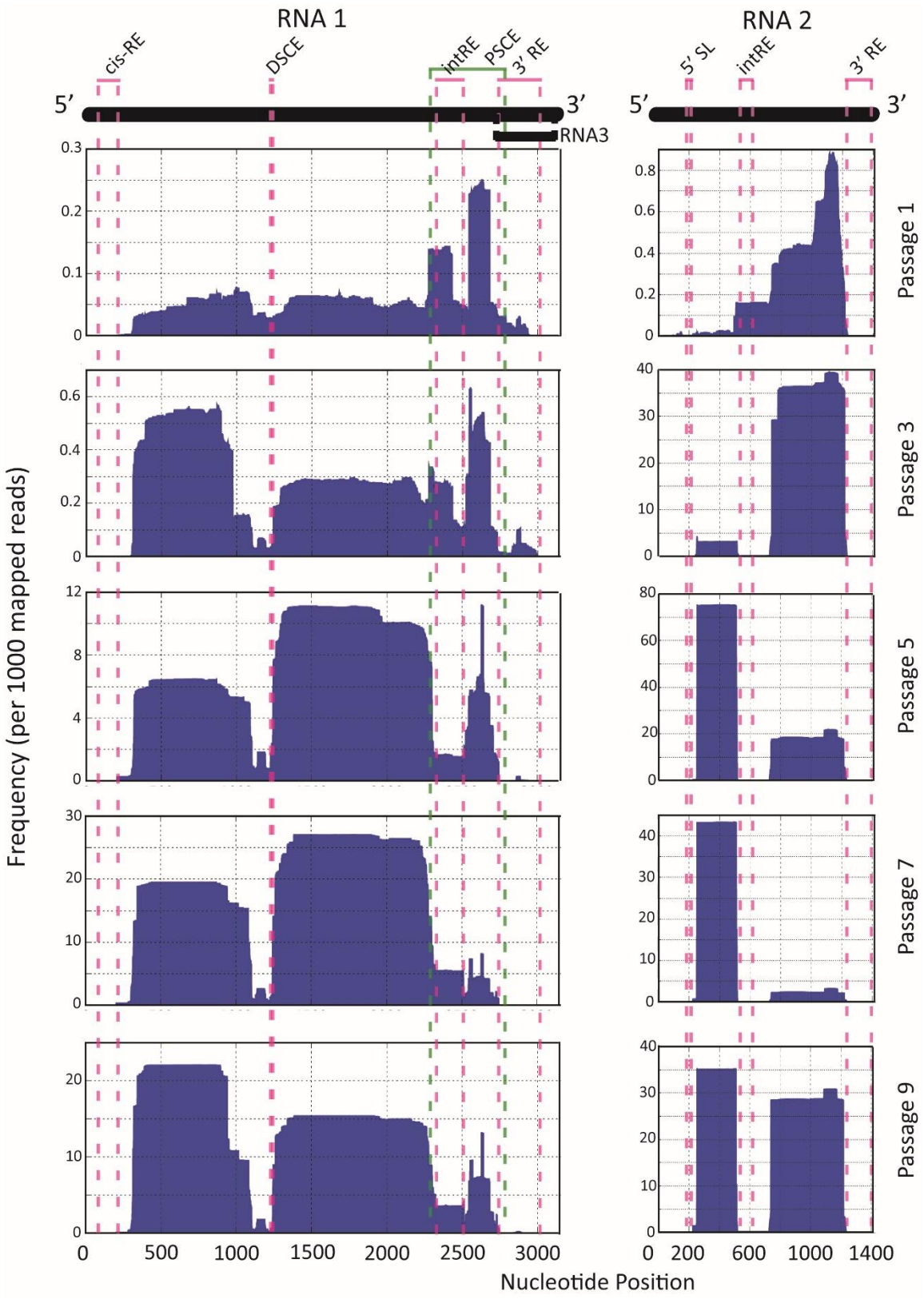
Figure 2.6: Open reading frame conservation during serial passaging.

A recombination event can retain a putative open reading frame by deleting 3n nucleotides. The frequency of conservation of the reading frame is calculated from the ratio of the number of recombination events that delete 3n nucleotide to the total number of recombination events mapped to that RNA.

multiple of 3 nucleotides (i.e. they thus conserved the ORF), as would be expected if deletion events occurred randomly throughout the genome. However, with continued passaging, there was a general trend toward conservation of the ORF for both RNA1 and RNA2 (**Figure 2.6**), although this was not the case for all replicates. Specifically, while initially showing an increase in ORF conservation, replicate 2 of RNA2 showed a decrease in the conservation of the ORF after passage 4, in contrast to the other two replicates, and in fact dips below 33%. Closer inspection of individual recombination events shows that this trend is driven by three of the four most common recombination events in replicate 2 passages 4 to 9: 249^517, **736^1219, 250^513** and **249^519** (the latter three events in bold do not maintain the ORF). These events are observed in the other replicates, but at a much lower frequency (no more than 1% of the total RNA2 recombination events for reps 1 and 3) (**Appendix B.4**). This indicates that DI-RNAs can indeed accumulate without a strict requirement for a functional ORF. However, this analysis only takes a single recombination event into consideration and the Illumina reads are not long enough (150bp) to resolve multiple deletions at the same time. Nevertheless, neither compensatory recombination events including small InDels that might restore the ORF, nor single nucleotide variants at putative stop-codons, were found.

Figure 2.7: Recombination profiling of conserved regions in the DI-RNAs.

Conservation map represents the frequency with which specific nucleotides in the FHV genome are deleted after recombination. Dashed lines indicate boundaries of functional RNA motifs. Odd numbered passages from replicate 2 are represented here. See **Appendix B.1** for all replicates and all passages. *Cis-RE:* cis-Regulatory Element; *DSCE/PSCE:* Distal/Proximal Subgenomic Control Elements; *intRE:* internal Response Element; *3' RE*: 3' Response Element; *5' SL*: 5' Stem Loop (next page)


RNA 1
RNA 2
cis-RE
DSCE
intRE
PSCE
3' RE
5' SL
5'
3'
RNA3
Passage 1
Passage 3
Passage 5
Passage 7
Passage 9
Frequency (per 1000 mapped reads)
Nucleotide Position
0
500
1000
1500
2000
2500
3000
200
600
1400

CONSERVATION MAPPING ILLUSTRATED EMERGENCE AND ACCUMULATION OF DI-RNAS

Using *ViReMa*, I was able to calculate the frequency with which each nucleotide was deleted, revealing areas of the viral genome that are conserved during serial passaging and required for DI-RNA replication. I plotted these data to generate recombination profiling maps for each RNA of FHV throughout passaging (**Figure 2.7** and **Appendix B.1**). In the first passage, there is a relatively even distribution of nucleotide deletions along the whole length of the genome with the exception of two frequently excised regions in the 3' end in RNA1 due to two common recombination events in each of the replicates: 2545^2685 and 2277^2435. By passage 3 the deletions along the genomic landscape begin to be '*sculpted*' whereby certain regions are deleted with a greater frequency than others. Passages 5, 7, and 9 were sculpted further revealing three major regions that were deleted in RNA1 and two in RNA2. Interestingly, for both RNA segments, while a range of deletions and rearrangements are generated during early passages, only the deletions that maintain regulatory and control elements are amplified during continued passaging, as previously observed in FHV[45]. These include the 5'/3' UTRs and internal response elements (intRE) of both genomes, as well as the Proximal- and Distal-Subgenomic Control Elements (PSCE and DSCE) in RNA1 (**Figure 2.7**), which correlates to the findings that these regions are important and required for RNA replication and encapsidation[45, 53, 54, 155, 179-182].

MINION NANOPORE LONG-READ SEQUENCING OF FLOCK HOUSE VIRUS

The short-read ClickSeq data provide in-depth and high resolution details of individual recombination events. However, in order to determine the correlation of these events over time, I used long-read ONT nanopore sequencing, which can characterize full-length wild-type and defective genomes (**Figure 2.8**). Both RNA genes were reverse transcribed and amplified using primers specific to the 5' and 3' UTRs of RNA1 and RNA2 from all passages of replicate 2 to obtain

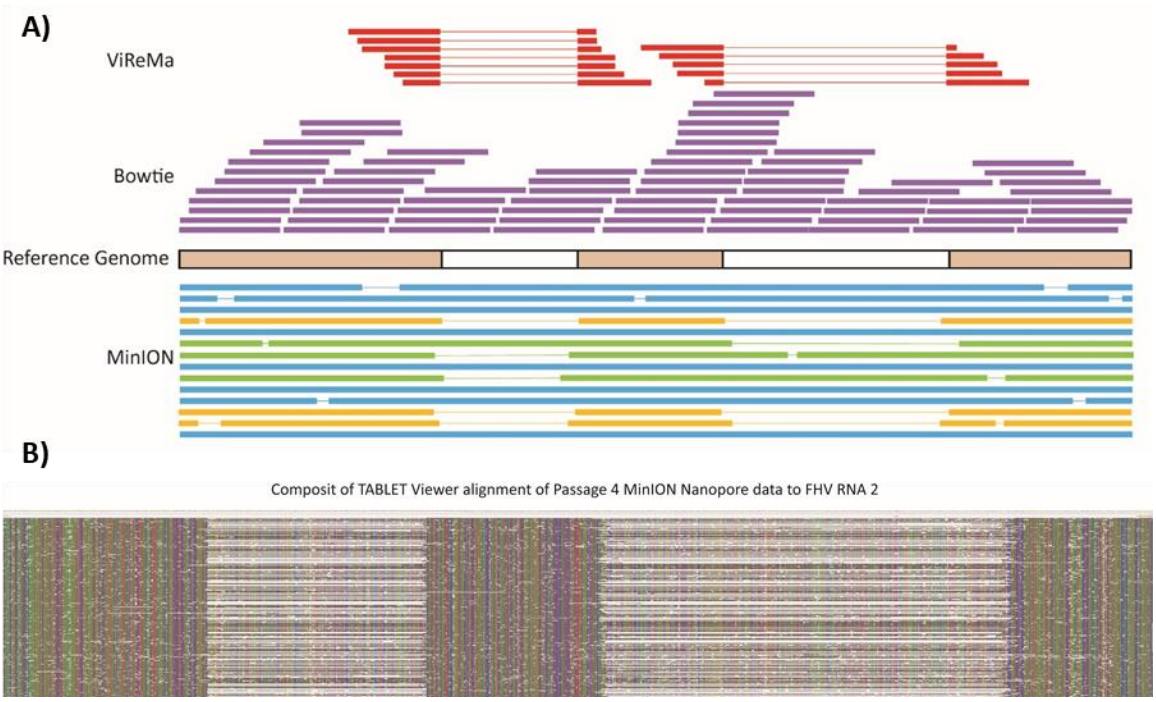


Figure 2.8: Comparison of Illumina HiSeq to Oxford Nanopore MinION read data

(A) Example of how reads generated from the Illumina HiSeq would map to a reference genome. The standard bowtie alignment is able to map 150bp reads along the reference with a relatively even coverage distribution. Unmapped reads are then aligned with the program ViReMa to accurately identify recombination events. The low error rate of high-throughput sequencing allows us to precisely define the boundaries of the junctions. Below the reference genome is an example of how reads generated from the Oxford Nanopore MinION would map to a reference genome. The MinION is able to generate full length reads at the expense of a high error rate which is ~7%. Full length analysis allows us to determine what recombination events a genome contains. Due to the error rate the exact boundaries of the recombination event are imprecise. **(B)** A composite snapshot of the TABLET sequence viewer alignment of RNA2 from the reads generated by the MinION (P4R2).

cDNA that could be barcoded and analyzed using protocols for 2D sequencing on the ONT MinION. The ClickSeq data shows that these regions were highly conserved during passaging (**Figure 2.7**), therefore I was confident using template-specific primers to these regions would capture both the full-length wild-type virus genomes as well as any defective RNAs. After PCR amplification, MinION cDNA libraries were analyzed using agarose gel electrophoresis to observe the distribution of cDNA fragments and ratios of full-length to defective RNA genomes (**Figure 2.9A**). While the cDNAs from the early passages are predominantly of the expected size for full-length RNA genomes, later passages contain an array of band sizes. This shows that in the early passages

the full-length genome is the predominant species while in later passages the truncated version becomes predominant. I also observe species appearing to be larger than RNA2. It is possible that some of these species correspond to RNA2 homodimers or other complex rearrangements that have previously been observed[183] and would result in an increased molecular weight. Evidence of RNA1 homodimers (3107^1), RNA2 homodimers (1400^1), and RNA2 to RNA3 heterodimers (1400^2720) can also be found in the ClickSeq data (**Appendix B.1**).

Amplified cDNAs from each sample was combined in equimolar ratios and the pooled, barcoded library was loaded onto a MinION MkIB device using an R9 flowcell as per the manufacture's protocol. I ran the standard protocol for obtaining 2-dimensional reads using the MinKNOW control software and collected nanopore reads for approximately 36 hours, upon which the quality and yield of reads dropped substantially. The sequencing run produced a total of 169,814 reads, of which 46,183 passed the default ONT filter and were successfully demultiplexed. This yielded between 2688 and 8815 full-length reads per passage and corresponds to approximately 0.1 Gigabases of sequence information. This would correspond to ~80,000 1x150bp Illumina reads per sample, assuming even coverage. With this depth, a comprehensive picture of the full-length genomic landscape of the viral samples can be built, allowing for the resolution of DI-RNA species even if they were present at less than 1% of the total viral genomic population.

The long-read nanopore sequencing data were aligned to the full-length FHV genome using the BBMAP suite (https://sourceforge.net/projects/bbmap/). This pipeline tolerates large insertions and deletions in the long-reads, thus allowing me to characterize the overall topology of the defective RNAs. Here, I mapped between 94 and 97.5% of the MinION reads from each passage to the FHV reference genome (**Table 2.3**). An example of reads aligned to FHV RNA2 is shown in **Figure 2.8B**. The error rate of aligned reads, including single nucleotide mismatches and

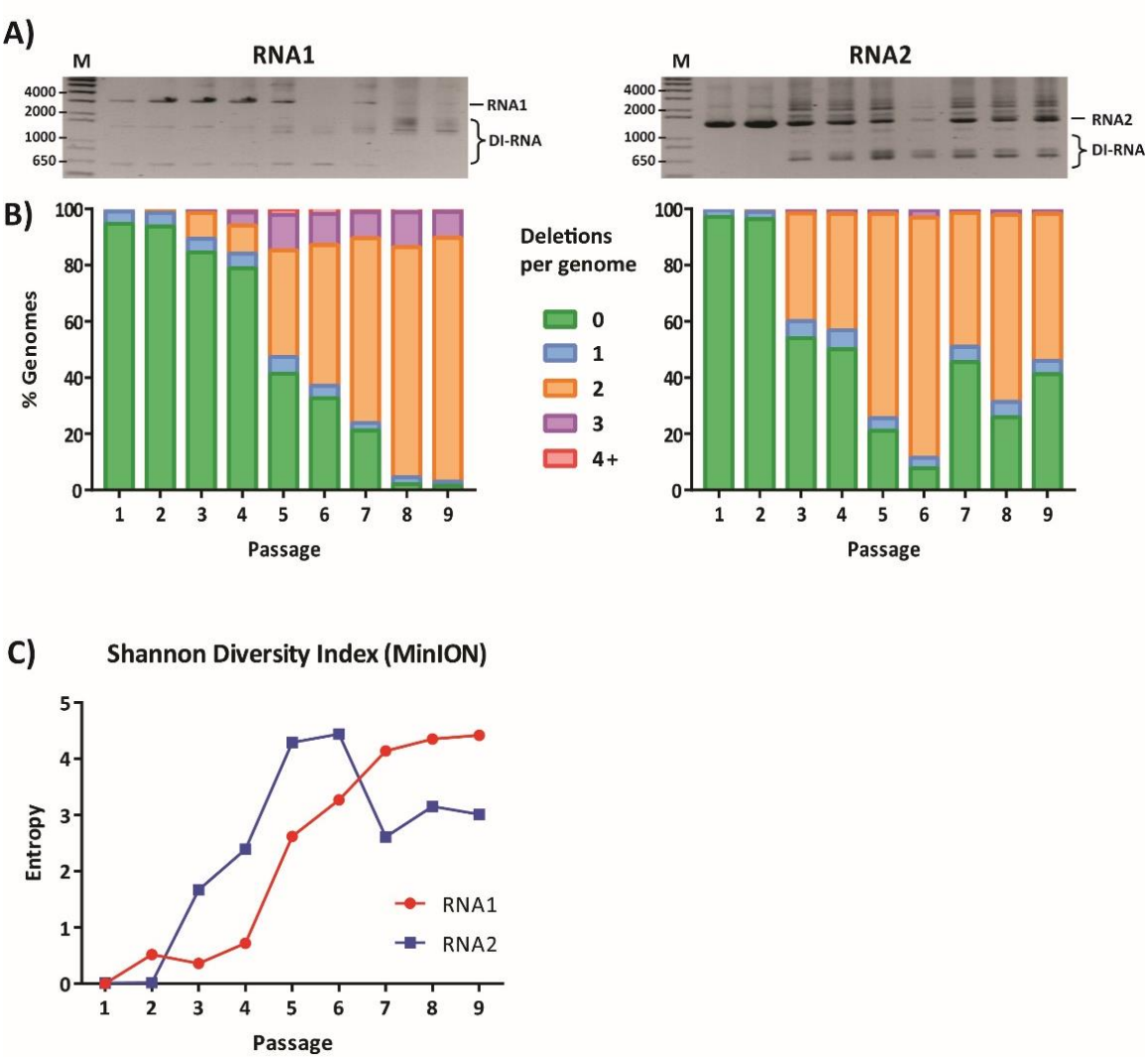


Figure 2.9: The frequency of deletions in the FHV genomic RNAs found by MinION nanopore sequencing.

(A) Gel electrophoresis analysis of cDNA copies for RNA1 and RNA2 for each passage in replicate 2 shows full-length viral genomic RNAs in early passages with increasing quantities of smaller defective RNAs bands in later passages. **(B)** Full length genomes were analyzed by the MinION and the number of deletions per genome were counted for each passage. Due to the higher error rate of the nanopore data only deletions ≥25nts were counted. **(C)** The Shannon Diversity Index of all RNA1 or RNA2 genes characterized with MinION nanopore sequencing is shown for each passage.

small InDels, was determined from alignment pileup files. Consistent with recent reports[184], I found the overall modal and mean error rates for all mapped position was 5.0% and 6.3% respectively, with 95% of the sites having an error rate better than 13.6% (N95 value = 0.864). A histogram of error rates for all mapped positions across all 9 datasets is shown in **Figure 2.10A**. Due to the large number of small InDels generated during nanopore sequencing[163], I also determined the frequency of deletions and insertions of different lengths (**Figure 2.10B** and **Figure 2.10C**). This shows that small InDels are frequent, but fall quickly in abundance with increasing length. 99.8% and 99.9% of all MinION deletions and insertions respectively were shorter than 25nts. Therefore, I considered only deletions and insertions of at least 25nts to be likely to be *bona fide* InDels present in the original viral RNA and corresponding to recombination events comprising defective RNA species rather than a sequencing error.

LONG-READ NANOPORE DATA CHARACTERIZE DEFECTIVE RNAS AND THE CORRELATION OF DELETIONS

The nanopore data reveals the presence and frequency of large deletions and insertions within defective RNA genomes. From these, I could reconstruct the population of either full-length or defective RNA genomes present in each of the viral passages (annotated as described in the *Methods* section). The full table of characterized defective RNAs and their frequencies in each passage can be found by following the link provided in **Appendix B.5**. In total, 6030 and 3639 unique defective RNAs of RNA1 and RNA2 respectively were found throughout all passages.

The frequency of individual recombination events found in both the ClickSeq data and the MinION data were compared and correlation coefficients calculated (**Table 2.3**). The correlation in the earliest passages was poor, due to the low abundance of events in both datasets. However, later passages correlate well with Pearson coefficients reaching 0.85. This is important as it

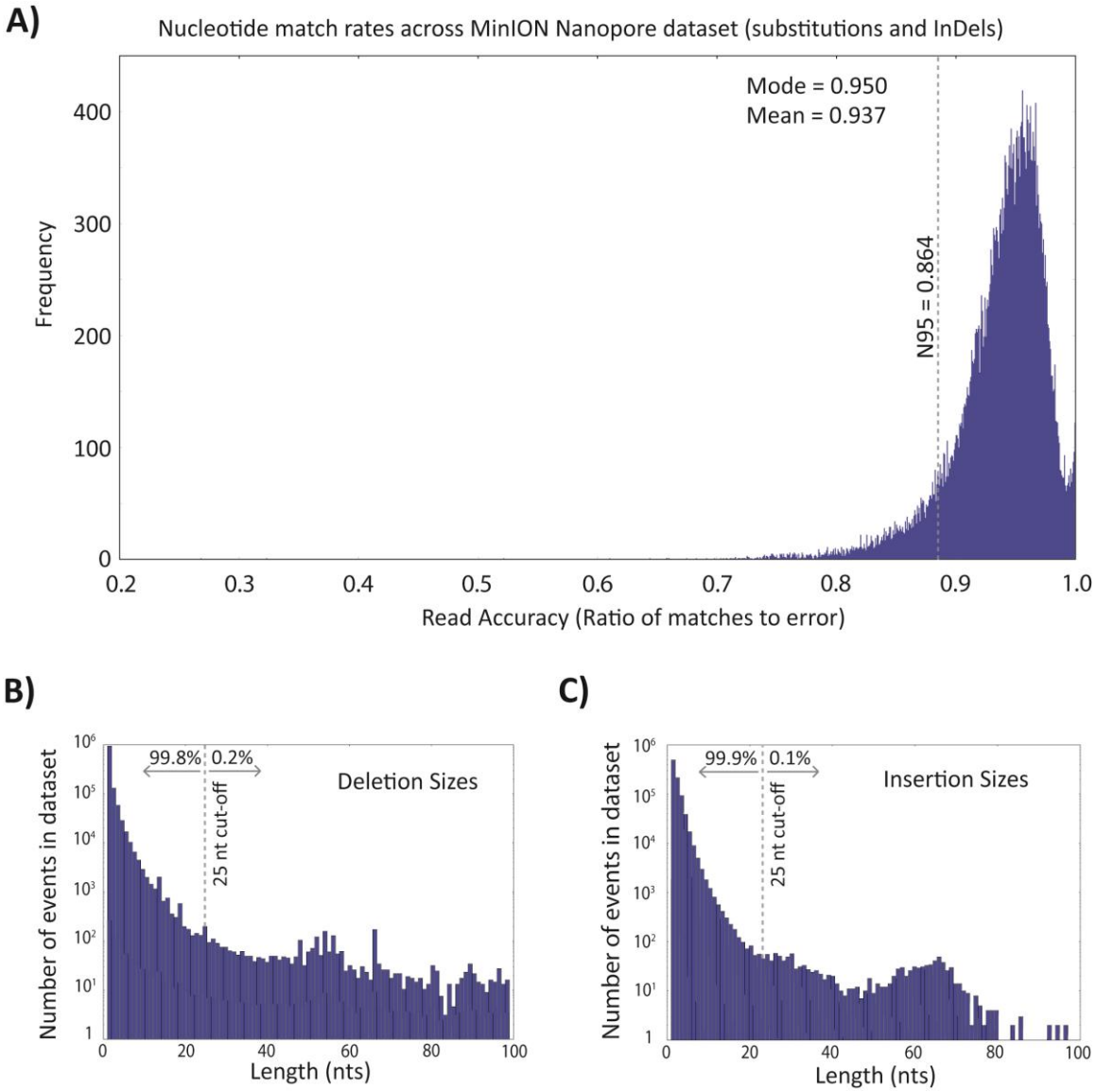


Figure 2.10: Error rate of the Oxford Nanopore MinION.

(A) A histogram of the frequency of error rates over every mapped position across all Nanopore datasets is shown. Errors include substitutions, insertions and deletions shorter than 25nts. The proportion of correct base matches to mismatches is shown on the x-axis. The y-axis indicates the number of nucleotide coordinates with the corresponding error rate. This reveals a mode and mean error rate of 5.0% and 6.3% respectively. Frequency of **(B)** deletions and **(C)** insertions within the MinION dataset.

demonstrates that the frequency of recombination events was not biased during cDNA amplification of the full-length or defective viral genomes. Similarly, the Shannon Entropy indices increase during passaging (**Figure 2.9C**), consistent with those from the ClickSeq data (**Figure 2.3B**).

The number of deletions in each passage in RNA1 and RNA2 are given in **Table 2.3** and illustrated in **Figure 2.9B**. The earliest passage contains very few deletions. In passage 1, 95.8% of the reads map to the full-length genome in its entirety. With subsequent passages, the number of reads containing deletions increases, reaching a plateau at passage 8 with 76.0% of the reads containing two deletions and 8.2% containing three deletions. DI-RNAs (e.g. 1_317^1091_1242^2301_3107 of RNA1) are easily identifiable as early as passage 2 and match well with the expected identities based on our ClickSeq results and previous studies[44, 143]. By the final passages these species predominate, leaving only a small percentage of full-length wild-type viral RNAs.

While the identity of mature DI-RNAs containing two or more deletions can readily be identified, few single-reads contain just one deletion (<6%) in all of the passages. Moreover, individual species are rarely observed again in subsequent passages (**Appendix B.2**). Importantly, most of these single events do not delete the expected regions common to FHV DI-RNAs. Therefore, they may either correspond to sequencing artifacts, or transient defective RNA species generated due to stochastic RNA recombination, similar to the low-frequency events observed in the ClickSeq datasets. In later passages (beginning at passage 3), the presumptive intermediates are seen (e.g. 1_317^1091_3107 or 1_1242^2301_3107 of RNA1) of mature DI-RNAs (e.g. 1_317^1091_1242^2301_3107). However, this is after the mature DI-RNAs were first observed, and after the point at which DI-RNAs have begun to accumulate. Indeed, in passages 2 and 3 respectively, mature DI-RNAs make up 1% and 30% of the viral population while the singly-deleted intermediates make up 0% (unobserved) and 3%. Together with the observation of rare DI-RNAs

in the inoculum with the ClickSeq recombination analysis, these data indicate that single-deletion species do occur early during passaging, but remain poorly abundant and do not accumulate. In contrast, mature DI-RNAs are observed to rapidly accumulate between passages, indicating that they possess a replicative advantage above both wild-type viral genomes and intermediate defective RNA species.

	Passage 1	Passage 2	Passage 3	Passage 4	Passage 5	Passage 6	Passage 7	Passage 8	Passage 9
Total Reads	**3842**	**5398**	**3055**	**5013**	**7675**	**3710**	**3413**	**2688**	**8815**
RNA1	**2118**	**2686**	**1426**	**2392**	**3703**	**2038**	**1991**	**1562**	**5387**
# of deletions:									
0	*2006*	*2519*	*1206*	*1886*	*1532*	*667*	*422*	*33*	*85*
1	*94*	*131*	*68*	*127*	*223*	*89*	*53*	*38*	*74*
2	*13*	*29*	*131*	*239*	*1405*	*1021*	*1310*	*1280*	*4674*
3	*5*	*6*	*19*	*111*	*464*	*225*	*184*	*193*	*499*
4	*0*	*1*	*1*	*26*	*66*	*30*	*19*	*18*	*48*
5	*0*	*0*	*1*	*3*	*10*	*6*	*3*	*0*	*7*
+	*0*	*0*	*0*	*0*	*3*	*0*	*0*	*0*	*0*
***R*, MinION vs ClickSeq**	0.08	0.02	0.38	0.30	0.63	0.68	0.64	0.27	0.52
RNA2	**1515**	**2388**	**1455**	**2336**	**3743**	**1595**	**1299**	**989**	**3207**
# of deletions:									
0	*1473*	*2301*	*786*	*1170*	*791*	*124*	*592*	*257*	*1325*
1	*36*	*64*	*89*	*158*	*164*	*59*	*71*	*53*	*150*
2	*6*	*22*	*558*	*968*	*2723*	*1364*	*617*	*658*	*1678*
3	*0*	*1*	*21*	*38*	*62*	*42*	*19*	*17*	*51*
4	*0*	*0*	*1*	*2*	*3*	*6*	*0*	*4*	*3*
+	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
R, MinION vs ClickSeq	0.09	0.01	0.55	0.85	0.73	0.59	0.63	0.38	0.56
Unmapped	**209**	**324**	**174**	**285**	**229**	**77**	**123**	**137**	**221**

Table 2.3: Mapping of nanopore data to the FHV genome using BBMap.

The number of demultiplexed nanopore reads passing quality filters are shown for each sample. These were mapped end-to-end to the FHV genome using the BBMap suite, allowing for large deletions and insertions, which were counted using the CIGAR string from the output alignment SAM file. The Pearson correlation coefficients of these events to those found using ClickSeq were also calculated. The number of reads mapping to FHV RNA1 and FHV RNA2 are indicated along with the number that contain 0-5 or more deletions of at least 25 nts. Only a small number (<5%) remained unmapped.

COMPLEX REARRANGEMENTS ARE OBSERVED BY MINION AND CONFIRMED BY CLICKSEQ

In addition to deletions, a small number defective RNAs first appearing at passage 5 contained insertions. The majority of these consisted of short insertions of ~200 nucleotides that were found in first 300nts of the MinION reads and were inserted between nts 19 and 20 of RNA1 (**Figure 2.11**). In each case, these inserts corresponded to nts 2300-2513 of RNA1. Interestingly, this region corresponds to an internal response element (intRE) of the Proximal Sub-genomic RNA Control Element (PSCE) previously identified as being essential for FHV RNA replication and conserved in DI-RNAs species[179]. The most common deletion in the DI-RNAs in this region of RNA1 for the final passages are from 1242 to 2301, which retains the intRE. However, there are also a large number of deletions ranging from 1245 to 2514, which would delete the essential intRE. Closer inspection of the MinION data reveals that the majority of these reads (>90%) that contain the 200nt intRE insertion concomitantly contained deletions from 1245-2514, indicating that these two events are correlated.

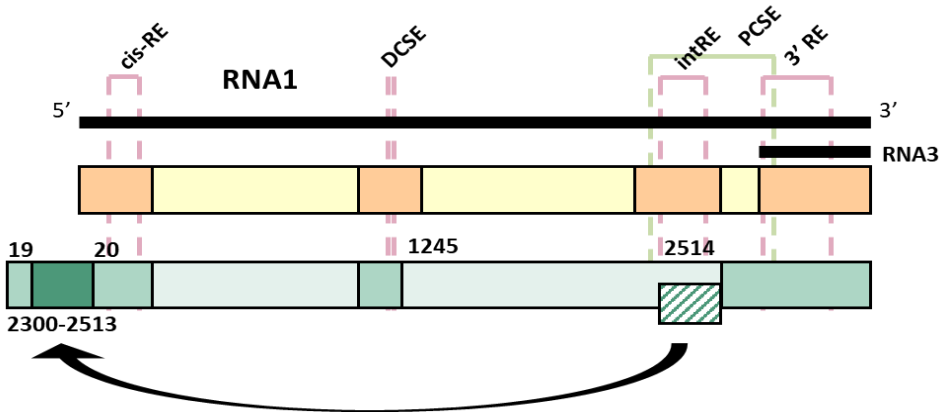


Figure 2.11: Graphical model of a complex rearrangement of a DI-RNA1.

Sequencing elucidated a complex rearrangement of a defective genome of RNA1. Orange bar (middle) indicates a 'common' DI-RNA1 where light orange indicate portions of the genome that are deleted. Green bar (bottom) is a depiction of the identified DI-RNA where the nucleotides corresponding to the intRE are reinserted in the 5' end of RNA1. *Cis-RE:* cis-Regulatory Element; *DSCE/PSCE:* Distal/Proximal Subgenomic Control Elements; *intRE:* internal Response Element; *3' RE*: 3' Response Element

The ClickSeq data also shows a frequent recombination event, 2513^21, which appears first in passage 4 and is among the 10 most common events in the final 5 passages. This matches precisely the 3' junction site of the insertion event detected in the MinION data. However, the event 20^2300 corresponding to the 5' junction site was not detected in our initial *ViReMa* analysis of the ClickSeq data as this would have required a search seed length of less than 20nts. Repeating the *ViReMa* analysis using a shorter seed length of 17 does indeed reveal the presence of the 20^2300 recombination event. This event is rarely observed in either of the other two replicate ClickSeq data (7 and 31 total reads across all passages of replicates 1 and 3 respectively). These data indicate that in a number of defective RNA genomes, the intRE element has been deleted and subsequently re-inserted at the 5' end of the defective RNA genome. As the intRE element is required for regulation of RNA replication, presumably this maintains the ability for this highly-rearranged defective RNA to replicate.

MinION Nanopore Sequencing Reveals the Emergence, Diversity, and Evolution of DI-RNAs

These data provide a comprehensive overview of the different species of defective RNAs that are present during viral passaging. Illustrating such a complex set of data is a challenge as each sample contains a large number of genome arrangements (6030 and 3639 for RNA1 and RNA2 respectively) and frequencies of these species vary substantially over time. I found that illustrating these data as a stacked area plot gave the most informative summary of the changes of the many different type of DI-RNA species over time. Due to the moderate error-rate of the nanopore read data, the exact identification of a recombination event and thus annotation of that genome may be incorrect. This would result in an over-estimation in the potential number of unique structural variants. Therefore datasets were filtered by requiring genomes to be

represented by three or more reads. While removing a lot of noise, this has the drawback where rare defective RNAs might be lost. Stacked area plots for genomes represented by three or more reads are shown in **Figure 2.12**. The stacked area plots for the unfiltered datasets are shown in **Appendix B.3**. This representation reveals key components of the evolution of the DI-RNA species.

The stacked area plot for RNA1 (**Figure 2.12A**) shows that the composition of DI-RNAs in the viral population changes over time and new species appear at each passage. For example, the most abundant defective RNA1 species in passage 5 is '*1_317^1091_1242^2301_3107*' but reduces in relative frequency in later passages. The most abundant species in the final passage 9 is '*1_313^941_1241^2325_3107*', which appears at low levels as early as passage 2, but does not begin to accumulate until passage 6 (**Appendix B.5**). Why this DI-RNA only begins to accumulate at later passages despite being present in the early passages is not clear. The 'complex DI-RNA' that deletes the intRE in RNA1 referred to in the previous section (*'1_342^1083_1245^2514_3107'*) is also observed (annotated in **Figure 2.11A**) first appearing at passage 5.

As can be seen in the stacked area plot for RNA2 (**Figure 2.12B**), the general composition of DI-RNA species is established at passages 4-5. Subsequently, the relative frequencies of the DI-RNA fluctuate but the overall diversity changes little with few new species appearing after passage 4. This is also observed when calculating the Shannon Diversity index (**Figure 2.9C**) whereby entropy reaches a maximum at passage 5 and decreases thereafter. Interestingly, this range of fluctuations resemble the sinusoidal patterns of DI-RNA abundance that have been observed in other studies of RNA viruses where the ratio of the frequency of DI-RNAs to wild-type genome has been measured through longitudinal studies[185].

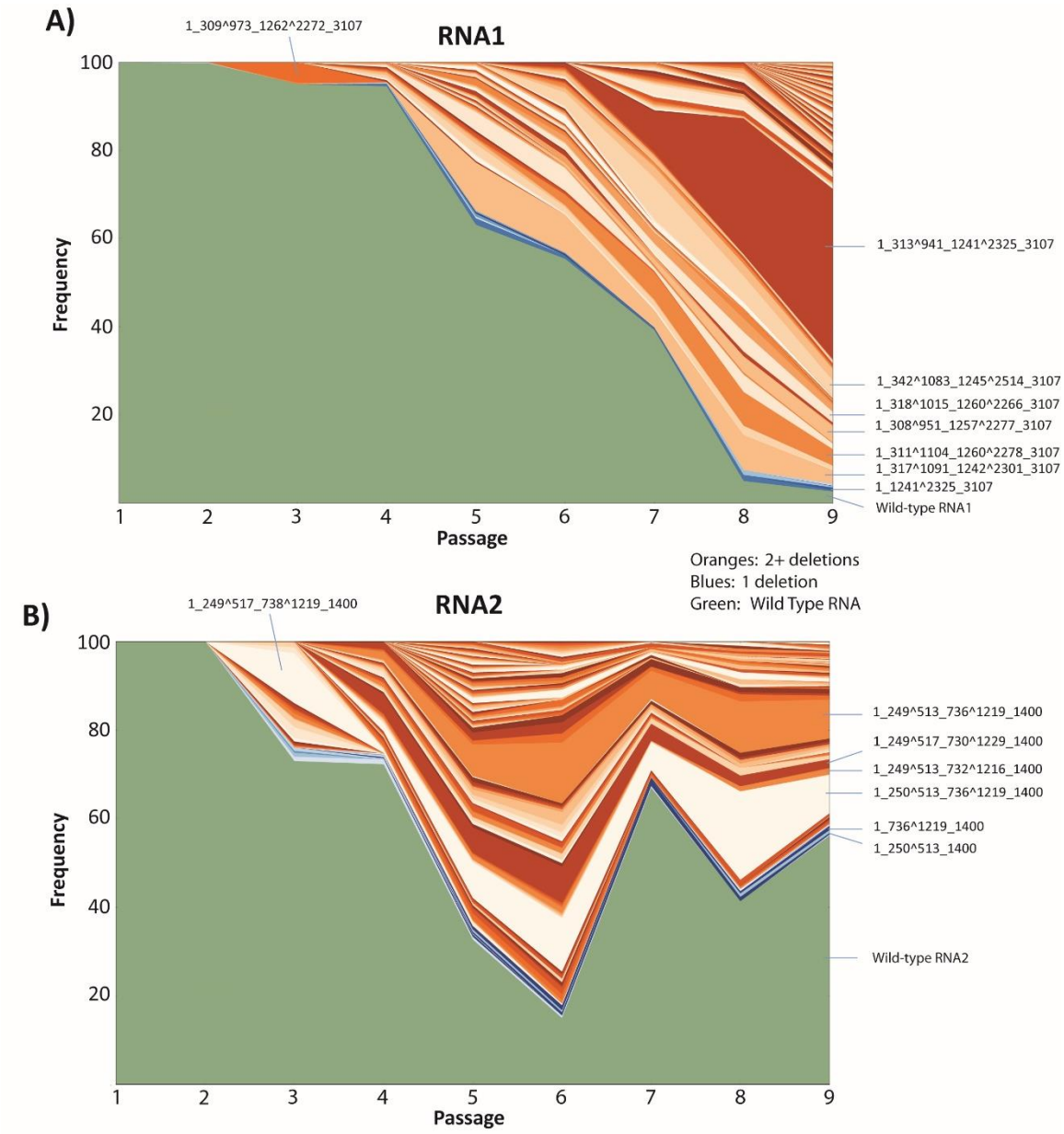


Figure 2.12: Evolutionary pathways of full-length RNA genomes.

Stacked-area plots for **(A)** RNA1 and **(B)** RNA2**.** The passage number is indicated on the x-axis and the stacked frequencies of each detected defective RNA is shown in the y-axis. Each non-contiguous color represents a specific genome characterized by MinION nanopore sequencing. Wild-type genomes are colored green, genomes with one deletion are colored in shades of blue, and genomes with two or more deletions are colored in shades of oranges (using the same color scheme as in **Figure 2.9B**). Raw data and annotations are in **Appendix B.5**.

SPECIFIC INFECTIVITY CORRELATES WITH ABUNDANCE OF DEFECTIVE RNAS.

The reduction in full-length infectious viral genomes and the accumulation of defective RNAs during passaging is likely to correspond to a decrease in the specific infectivity of the virus samples. To determine the effect of defective RNAs characterized by combined ClickSeq and nanopore sequencing of replicate 2 upon specific infectivity, I performed a 50% tissue culture infectious dose (TCID50) assays for each passage 1-9 for both the original inoculum used to infect each sample and for the particles purified from each passage[170]. The TCID50 assay is used to determine the dose required to give a 50% chance that cells in culture will be successfully infected as determined by CPE and is typically used to determine viral titer and the effective MOI of the inocula transferred from passage to passage. The results from the TCID50 assay for each passage are shown in **Figure 2.13A** and **Figure 2.13B**. This assay indicated that the TCID50 value (and thus PFUs (Plaque Forming Units)/ml) drops considerably during passaging by over four orders of magnitude. The corresponding effective MOI (PFUs per cell) also drops from 38.5 to 0.0003 during passaging (**Figure 2.13B**).

To determine whether the drop in effective MOI was driven by reduced total particle yield or from reduced specific infectivity (i.e. virus particles per PFU), TCID50 analysis was performed of our purified and quantified virus stocks (described in *methods*). This allowed me to determine and normalize the number of virus particles delivered per cell between each passage. As the particle-per-PFU ratio has previously been estimated at 300-400 particles-per-PFU[171 152], the assay was setup beginning with 300'000 particles per cell in 96 well format and performed 8 10-fold serial dilutions. In this assay, the number of viral particles required to induce CPE decreased by over 400-fold during passaging (**Figure 2.13A**) with a trend very similar to that for the unpurified inocula. Together, these data indicate virus specific infectivity drops with a corresponding

increase in the defective RNA population. There was an exception at passage 7 where TCID50 actually increased ~5 fold from the previous passage. This could be correlated to our observation of a decrease in the amount of defective RNA2 species in the MinION analysis (**Figure 2.9B** and **Figure 2.12B**).

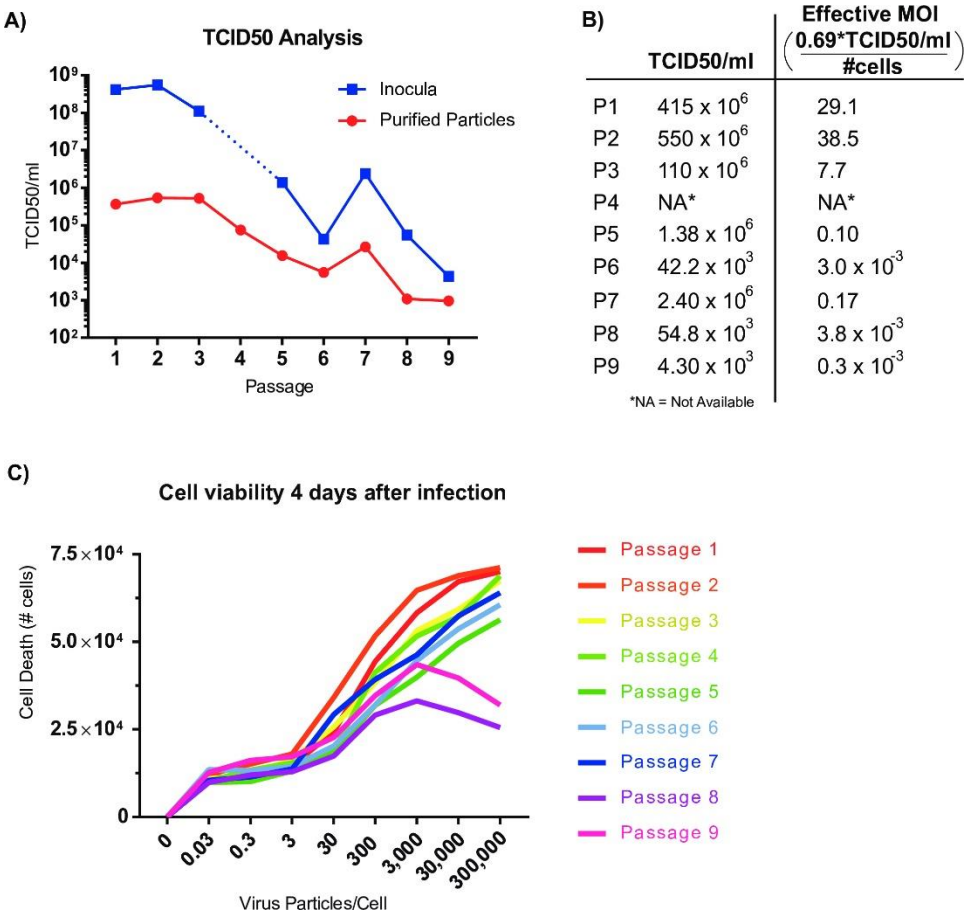


Figure 2.13: TCID50, specific infectivity and Cytopathic effect (CPE) of each passage for replicate 2.

(A) The 50% tissue culture infectious dose (TCID50) of each passage of replicate 2 is shown (inocula used in serial passaging in blue squares). Additionally, virus was purified and quantified by OD_{260nm} before infection allowing normalization particles per cell across all replicates for further TCID50 analysis (in red circles). **(B)** TCID50/ml values were used to calculate effective MOI values during serial passaging. *Passage 4 sample was not available. **(C)** 10^5 cells were infected with quantified and serially diluted virus. The number of virus particles per cell plated is indicated on the x-axis. After 4 days of infection, the remaining number of cells was analyzed by flow-cytometry to count remaining viable cells. Cell death is indicated by the difference in count compared to non-infected wells (0 particles/cell). Each line indicates a different passage from replicate 2 as indicated in the color key.

I further characterized each well of the TCID50 assay of our purified particles using flow-cytometry to give a quantitative assessment of cell survival and death in response to virus dose. We calculated the number of live cells that remained after infection at each dilution and for each passage (**Figure 2.13C** and **Appendix B.3**). Reduced overall CPE was observed in later passages at the highest virus dose as well as an increase in the number of viral particles require to induce the same amount of CPE (**Figure 2.12C**). Together these trends reflect a reduced specific infectivity during viral passaging, in agreement with the TCID50 assays. Interestingly however, for the highest particle concentrations in passages 8 and 9, less cell death was seen at the highest doses (300,000 and 30,000 particles per cell) than for cells infected with the same inoculum (and therefore same ratio of full-length to defective RNAs) but at a lower dose (3,000-30 particles per cell). This observation indicates the protection of cells from infection and/or CPE when supplied with a large dose of viral particles that contain a large proportion of DI-RNAs.

CONCLUSIONS

In this chapter I sought to provide a thorough and comprehensive analysis of the frequency and identity of recombination events present during the serial passaging of Flock House virus in cell culture in order to elucidate the pathways and mechanism of DI-RNA emergence and evolution. I began with a homogenous inoculum derived from plasmid cDNAs of each of the FHV genes. In the inoculum and in the early passages, I found a wide range of low-frequency recombination events corresponding to deletions and duplications that are dispersed throughout the viral genomic RNAs. Here I can be confident that these species do not constitute sequencing artifacts as the RNAseq libraries were made using 'ClickSeq'[157] that has previously been demonstrated to reduce artifactual recombination in RNAseq data to fewer than 3 events per million reads. Further confidence in the low rates of artifactual recombination in this study is provided internally by inspecting the numbers of inter-RNA recombination events (RNA1 to RNA2 and *vice versa*), which are always low. Furthermore, the majority of the detected inter-RNA recombination events correspond to genomic RNA hetero- and homo-dimers, which have previously been characterized as replication intermediates[183].

Within only 2-3 passages, however, deletion events similar to those previously observed in DI-RNAs appear in all three passaging replicates. In subsequent passages, these recombination events begin to accumulate rapidly so as to predominate over full-length viral RNAs. This observation of the emergence of DI-RNA species, followed by their rapid accumulation is consistent with existing theories on the evolution of DI-RNAs that postulate that a wide range of potential DI-RNA species are generated by non-programmed RNA recombination and that only a handful are successfully replicated and thus accumulate[46]. While the short-read data provide high-resolution characterization of individual recombination events, it is through the use of the Oxford Nanopore Technologies (ONT) MinION that I am able to reconstitute the complex full-length

genomic landscape of FHV during passaging and determine the relative abundances of the genomic RNAs in each passage. As a result, I was able to determine that by the final passage only ~2% of the mapped reads are full-length viral RNA1, which corresponded with a large reduction in specific infectivity. Additionally, the nanopore data revealed complex rearrangements of RNA genomes, including the excision of an entire functional RNA motif and its reinsertion in the 5'UTR of RNA1.

The variation in recombination boundaries of the DI-RNAs suggests that a range of deletions can be tolerated. However, it is important to note that while each replicate is its own distinct lineage, each replicate passaging experiment was derived from the same initial inoculum. The experiment was designed this way to determine if the same DI-RNAs would be generated independently or if completely different deletions would arise and be selected for even though the environmental conditions are practically the same. Here, the latter was observed (**Table 2.2**). Few of the RNA1 recombination events observed in the inoculum are observed again in subsequent passages. Additionally, even though the event '738^1219' was found in RNA2 the inoculum, this was not the final predominant species in any replicate. Therefore, the evolution of DI-RNAs was not pre-determined by the presence of rare DI-RNA species in the common inoculum (a founder-effect), but rather by the selection of well-replicating DI-RNAs that arose later during serial passaging. Nonetheless, the final recombination events are highly similar between replicates and to previous reports from different laboratories. Therefore, this indicates that either the DI-RNAs have emerged due to a common mechanism of formation, the presence of a common selectivity filter, or both.

In addition to providing a thorough analysis of the pathways of defective RNA formation and evolution, there are two unexpected and critical observations made through this study. First: while a wide range of recombination events early on during passaging was observed, only a

limited number of events are subsequently amplified and later *define* DI-RNAs. Moreover, these limited sets of events are similar between replicates, and to previous studies. This suggests that while a large pool of potential defective RNAs are generated, only a small number are capable of accumulating. Secondly: I did not observe the amplification of DI-RNAs with only one deletion. In contrast, 'mature' DI-RNAs accumulate rapidly. Nonetheless, there is evidence for intermediate DI-RNAs as early as the inoculum sample. This indicates that intermediate defective-species are either non-competitive and do not accumulate or are not formed as a pre-cursor to mature DI-RNAs. These two observations provide important insights into the potential mechanisms of DI-RNA emergence and evolution.

While there is a strong selection pressure for DI-RNAs to retain essential functional genomic elements, it is also postulated that a shorter defective RNA would be replicated more quickly and thus more competitively [3]. In this analysis, while the ~250-550 deletion in RNA2 (for example) is very common, I did not observe the accumulation of deletion events that are smaller than this (e.g. 300-450). This is despite the fact that there is detection and observation of such species in low frequencies both in early and late passages, suggesting that they are indeed generated but are not selectively amplified. This may in part be due to selecting for DI-RNA genomes that are as small as possible, while retaining the minimal amount of genetic material to form functional genetic elements. However, it may also be the result of a negative selection pressure or restrictive barrier that is released only after excising specific portions of the viral genome. An example of this scenario has been demonstrated for tomato bushy stunt virus (TBSV) associated DI-RNAs whereby deletion of a translation enhancer functional element removes the competition between translation and replication, thus favoring replication of the smaller DI-RNA[85]. Therefore, the final structure of DI-RNAs may depend both on the retention of essential

functional RNA elements, as well as the removal of restrictive barriers that attenuate RNA replication.

One model for the evolution of DI-RNAs is through the step-wise accumulation of deletion events through a series of individual recombination events[85]. The MinION data reveals that the defective RNAs that accumulate (rapidly over the course 2-3 passages) contain multiple deletion events. However, I did not see the rapid accumulation of the intermediate DI-RNAs, despite evidence for their presence early in viral passaging. This suggests that the mature DI-RNAs have a competitive advantage over their presumed intermediate precursors. If this is the case, the multiple deletions may function epistatically either through an undefined cooperative/additive mechanism or through the release of multiple restriction barriers, as proposed above. If multiple restriction barriers are required to be excised for the formation of DI-RNAs, small or multipartite RNA viruses, such as FHV or influenza[80], may therefore generate DI-RNAs more readily by requiring fewer intermediate steps than long, monopartite RNA viruses. Moreover, if intermediate defective RNAs fail to accumulate, this reduces the likelihood that mature DI-RNAs can subsequently be generated and may place substantial limitations on the ability of some viruses to generate DI-RNAs altogether.

An alternative reason for the rarity of precursor/intermediate defective RNAs is that the mature DI-RNAs are generated in one single event. We are yet to determine the molecular mechanism of recombination that leads to DI-RNA formation. Both template-switching, secondary-structure jumping, and non-replicative mechanisms have been proposed, and indeed these mechanisms need not be mutually exclusive. My observation of nucleotide preferences at recombination junctions (**Figure 2.5A**) may arise through any of these potential mechanisms. Alternatively, it is possible that multiple reassembly/deletion events occur in a single step, in a manner reminiscent of chromosome shattering (chromothripsis)[186]; or '*virothripsis'*. Within the

confined invaginations of the mitochondrial membranes that form the replication factories of RNA viruses such as FHV[187], the fragmentation the RNA virus genome followed by incorrect re-stitching of these genome pieces, either through forced-copy choice template switching or a non-replicative mechanism, could create the DI-RNAs observed here including the complex rearrangements observed for RNA1.

A defective-*interfering* RNA is a defective RNA that has the ability to compete with or otherwise attenuate the replication and proliferation of the wild-type helper virus. In this study I also demonstrate that the viral swarm, even after only a few passages, is replete with many varieties of defective RNAs. With a single sequencing experiment, I would not be able to determine whether these defective RNAs are accumulating, diminishing or make-up a static component of the viral intra-host diversity. However, as I performed serial passaging with sequential sequencing experiments, I could determine which defective RNAs are accumulating (for example the 'mature' DI-RNAs) and which are not (e.g. the putative intermediate, or 'immature' defective RNAs). It would be impractical to validate each of the many hundreds of detected defective RNA species with molecular virological experiments to determine whether they truly can attenuate or interfere with wild-type virus replication and to therefore categorize that species as a defective-*interfering* RNA. Indeed, I cannot exclude the possibility that multiple DI-RNAs act co-operatively within the viral intra-host diversity and are mutually dependent upon one another. However, the demonstration here of an accumulation during serial passaging is strong evidence that these species are *interfering*, as their accumulation essentially dilutes the pool of wild-type functional virus. With this in mind, I believe it would be suitable to describe the mature defective RNAs as defective-*interfering* RNAs (DI-RNAs), and the 'immature' only as defective RNAs.

It is remarkable that FHV is able to maintain a viable infection despite being burdened with such a gross excess of DI-RNAs in the final passages presented here. By performing TCID50 assays of the original inocula used between each of the passages and of the particles purified from each passage, I showed that there is a dramatic reduction in specific infectivity during passaging corresponding with the increase in the DI-RNA content. DI-RNAs have generally been demonstrated to arise at high MOIs, as was the scenario with our first passages. However, the calculated MOI drops rapidly after DI-RNAs have formed to levels that might be expected during typical *in vivo* viral passaging scenarios. However, in this experiment I was actually passaging a large number of virus particles between cells, but with a low specific infectivity. It is also interesting to observe that for the final passages there appears to be a protective property of the DI-RNAs as determined by flow-cytometry, but only when administered at the highest doses corresponding to over ~30,000 particles per cell. However, the role that DI-RNAs might play *in vivo* is not clear, as these very high doses may not be physiologically relevant. DI-RNAs for FHV similar to the ones described here have been observed to arise in experimental fruit fly infections[129] and so the mechanism of formation and/or selection is likely to be similar in cell culture and *in vivo*. However, whether RNA viruses such as FHV have evolved to favor the spontaneous formation of DI-RNAs and if so whether these DI-RNAs play an important role in modifying the life-cycle of the virus, is yet to be determined.

CHAPTER 3: STRUCTURAL ANALYSIS OF THE DEFECTIVE FHV PARTICLE ELUCIDATES SELECTIVE PRESSURES IMPOSED UPON VIRAL GENOMES

INTRODUCTION

In **Chapter 2**, I was able to show that Flock House virus accumulates defective viral genomes during serial passaging. Not only do defective genomes become the most predominate genome under certain circumstances, but it also appears that a slightly variable, but specific, 'species' of genome prevails where these fully formed genomes (ones with multiple deletions) appear and accumulate very quickly. Interestingly, the genomic arrangements I have characterized resemble the ones that many groups prior to us have also seen, implying a common mechanism of formation or selection[44, 45, 54, 86, 188]. While I saw the appearance of potential intermediates to these DI-RNAs (genomes with single deletion events), their levels remained extremely low during passaging. Overall this seems to indicate: 1) 'mature' FHV DI-RNAs are formed in one step, removing multiple portions of the genome at the same time, 2) there is a selective pressure only allowing the accumulation of a certain species, or 3) a combination of both.

Here, I sought to explore the potential mechanisms behind the formation and accumulation of mature defective genomes. The vast accumulation of this type of defective genome may purely be a product of the speed at which a smaller genome replicates. I previously showed that the most abundant DI- genomes seem to remove nucleotides towards the boundaries of conserved regulatory elements (such as replication and packaging signals) while still maintaining them. A great example of this idea is the defective genome with the complex rearrangement of sequence that was identified using Nanopore sequencing. This DI-RNA1 genome had a deletion event removing the internal response element (intRE) which is an event larger than commonly seen in other DI- species. Importantly, this element was translocated to the

5’ end of the same genome. This seems to indicate that even though there are mechanisms that allow for the formation of smaller genome size there is still pressure to maintain certain structural elements. While it is clear that viral genomes need to maintain these regulatory elements, is still isn’t clear what effect the length of a defective genome has on its replication kinetics. Some studies seem to indicate that length has no influence on the speed at which the polymerase transcribes genomes[189], while others have suggested that the new composition of a defective genome can actually increase their affinity for the polymerase (therefore increasing their replication)[92].

Furthermore, the mechanisms of packaging also impose a restrictive barrier that a defective genome has to overcome in order to be transmitted. The process of packaging viral genomes into a capsid is a critical step of the life cycle of a virus. This is the step where the virus must have a mechanism to identify the right genome(s) to package. For Flock House virus, it is known that the packaged viral RNAs play an important role in the structure and geometry of the particle[190]. Here, we know that viral RNA conforms in a double-helix to form a dodecahedral cage within the capsid and the disordered regions of the N- and the C- terminus of the capsid protein are key for this interaction[191-195]. While overall many aspects of the virus life cycle have been characterized, the mechanisms behind particle formation and assembly of this virus are still relatively unclear.

To better understand what selective pressures that packaging imposes upon the selection of defective genomes, I characterized the structure and appearance of particles containing defective genomes. To do so, I created populations of virus that had a large relative frequency of DI-RNAs via high MOI passaging. The genomes of these populations were thoroughly characterized using parallel ClickSeq and Nanopore sequencing technologies described previously (**Chapter 2**). Subsequently, in collaboration with Dr. Albert Heck and his group, we applied the

novel technique of native ultra-high mass spectrometry to determine the mass of a defective particle. Here, we were able to show that populations of particles full of defective genomes resolved to the same mass as wild type particles. Furthermore, in collaboration with Dr. Peter Stockley and his group, we were able to image these particles using cryo-electron microscopy and showed no significant/discernable difference in the capsid structure or the amount of packaged RNA in defective particles. Collectively, these data imply that the FHV particle has a system to measure the amount of RNA that is encapsidated, suggesting that packaging plays an important role in the selection of defective genomes. Interestingly, using the long-read sequencing data produced on the MinION I was able to calculate the lengths of DI- genomes and indeed, combinations of defective genomes would be able to sum up to a wild type amount of RNA.

Furthermore, I wanted to test the implications that length of the genome had on the speed of replication. Therefore, I made clones of DI-RNAs of varying length and transfected them into cells to see if length had an implication on replication speed and genome accumulation. While more studies are needed, I was able to show that defective genomes of different lengths had the ability to be replicated.

METHODS

VIRUS CULTURING, ISOLATION, AND PURIFICATION

S2 (*D. melanogaster*) cells were cultured at 28°C in Schneider's *Drosophila* Media supplemented with 10% fetal bovine serum and 1% Penicillin-Streptomycin using standard laboratory procedures. To generate Flock house virus (FHV) populations with high concentrations of defective interfering genomes virus was passaged at high multiplicity of infection. For a target MOI of 3000, 47µg of purified FHV was added per $1x10^7$ cells and cultured in T-25 flasks. Virus was grown for 3 days. Virus isolation was performed as described in **Chapter 2**[86]. Briefly, virus was released from cells by a 1% Triton X-100 treatment followed by a freeze thaw cycle. Virus was then isolated by ultracentrifugation on a 30% sucrose cushion (spun at 40,000 RPM for 2.5 hours) and the pellet was resuspended in 10mM Tris (pH 7.4). Resuspended virus was then applied atop a 10-40% sucrose gradient, spun at 40,000 RPM for 1.5 hours, and viral band was collected. Virus was then treated with 1U DNase and 1U RNase to remove any non-packaged nucleic acids. The virus sample was then concentrated on a 100,000 NMWL centrifugal filter and washed with at least 2 volumes of 10mM Tris (pH 7.4). Purified virus was then aliquoted for different analyses. For genome characterizations RNA was extracted from purified virus using the Zymo Research Direct-zol RNA extraction kit as per the manufacturer's protocol.

NORTHERN BLOTTING

RNA northern blot (NB) analysis was performed using a modified version of Thermo Fisher Scientific's Northern Max Kit (Cat #AM1940). Extracted RNA was mixed in 2 volumes of a denaturing solution (95% formamide, 0.02% SDS, 1mM EDTA), denatured at 65°C for 15 min, and immediately snap cooled on ice. Samples were then loaded on a denaturing 1% agarose gel and

ran at ~5V/cm. The outer ladder and reference lanes were cut off, stained in ethidium bromide, and imaged using a UV transilluminator. The remaining gel was soaked in an alkaline buffer (0.01N NaOH, 3M NaCl) for 20min. RNA is then transferred to a nylon membrane for 4 hours following the standard transfer procedure as directed in the Northern Max protocol. Sample RNAs are crosslinked to the membrane using a Stratagene UV Stratalinker 2400 using the 'auto crosslink' setting. The membrane was pre-hybridized with ULTRAhyb-Oligo buffer (Thermo Fisher) for 30min at 42°C using ~5mL/100cm^2 of membrane. Cy5 or Cy3 probes were designed to target conserved regions of the positive strand of the FHV RNAs (ordered from Integrated DNA Technologies). Probe sequences as follows; targeting nts 2961-2980 of (+)RNA1: 5'- Cy5-GAGTGTTGGTTTTGCCTCCT; nts 271-221 of (+)RNA2: Cy3-GAAACGCCAAACCAGGTTGACTTAATCT-GGTTAGCGCCGCCATGTTCAT. Each probe was diluted to 5pM in fresh ULTRAhyp-Oligo buffer and allowed to hybridize to the membrane at 42°C overnight. Subsequently, the membrane was subjected to a series of wash steps: twice at room temperature for 5min using Low Stringency Wash solution (Northern Max Kit), once at 42°C for 15min with High Stringency Wash solution (Northern Max Kit), once at room temperature for 30min with Odyssey Blocking Buffer (with 1% SDS), twice at room temperature for 5min with TBST, and finally rinsed/stored in TBS. The blot was exposed and imaged using a Typhoon FLA 9000 (GE). Band intensities were quantified using ImageJ software[196].

VIRAL RNA SEQUENCING

Viral genomes were characterized using both short-read ClickSeq sequencing and long-read Oxford Nanopore Technologies' MinION sequencing. For details on the protocols and corresponding data analysis see **Chapter 2: Methods**.

MASS SPECTROMETRY

Mass spectrometry analysis was performed by Tobias Wörner as a collaboration with the Albert Heck laboratory from Utrecht University. Purified FHV was buffer exchanged to 75mM ammonium acetate, (pH 7.4), with Amicon 100,000 NWCO filter units, performing 4 consecutive dilutions steps at 1,000 x g and 4°C, yielding a final monomer capsid concentration of 5µM (assuming a capsid mass of 43.71 Da).

Native mass spectra were recorded on a QExactive Ultra High Mass Rage spectrometer (QE-UHMR) allowing measurements up to 100,000 *m/z*[172]. Proteins were directly infused into the mass spectrometer with in-house made gold-coated nanoelectrospray needles. Capillary needle voltage was kept at 1,350V and the in-source trapping was employed at -50V for better focusing of large ions[197]. Xenon was used in the collision cell at a pressure of approximate 9 x 10^{-10} (UHV readout) and ion transfer optics were tuned for the transmission of large ions. For desolvation, ions were activated using 300V collisional activation, transient length was set to 64ms and spectra were generated by transient averaging. The instrument was calibrated using cesium iodide (CsI) clusters sprayed up to 11,000 *m/z*. Charge-states were determined by minimizing the standard deviation of the mass over the charge-state distribution.

CRYO ELECTRON MICROSCOPY

Cryo-EM data was generated as previously described by Hesketh *et al*.[198] and performed by members of the Peter Stockley laboratory at the University of Leeds. Briefly, Cryo-EM grids were glow-discharged for 30 seconds and subsequently 3µl of virus sample was applied to the grid, incubating for 5min, and repeated 3 times. Grids were blotted immediately after final application and plunge frozen in liquid ethane using a Leica EM GP device. Data was collected on

an FEI Titan Krios EM (Astbury Bioctructure Laboratory, University of Leeds) at 300kV, with an electron dose of 110e^-/Å^2, and a magnification of 75,000x. Exposures were acquired with an object sampling of 1.065Å/pixel using a FEI Falcon III direct electron detector and the EPU automated acquisition software. Each exposure contained 79 frames over a two second exposure time.

Image processing was done using the RELION 2.0/2.1 pipeline[199] and MOTIONCOR2[200] was used to create drift-corrected averages. Contrast transfer function was determined using gCTF[201]. Particles were picked manually and classified using reference-free 2D classification[202] and the generated 2D class averages were used as templates for automatic particle picking by dAutomatch[203]. RELION was used to classify particles using several rounds of reference-free 2D classification and 3D classification, with imposed symmetry. The starting model was reconstructed using the negative stain reconstruction filtered to ~60Å, after which, best classes were used for the subsequent rounds of classification. To mask the model and correct for the B-factor of the map, post-processing was employed[204]. Final resolution was calculated using the 'gold standard' Fourier shell correlation (FSC = 0.143) and local resolution was estimated using RELION's local resolution feature[199].

TRANSFECTIONS FOR REPLICATION KINETICS

S2 cells were counted and 1x10^6 cells were plated in each well in a six well plate format. Cells were transfected with the indicated combinations of 2µg of pMT-FHV plasmid using Lipofectamine 3000 Transfection Reagent as per the manufacturer's protocol. Plasmids included: pMT-FHV RNA1 (NC_004146), pMT-FHV RNA2 (NC_004144), pMT-FHV RNA2-Δ1 (containing a deletion of RNA2 at nts 249-517) pMT-FHV RNA2-Δ2 (containing a deletion of RNA2 at nts 736-1219) and pMT-FHV RNA2-Δ1-2 (containing two deletions of RNA2 at nts 249-517 and 736-1219). Plasmid transcription was induced 24 hours post transfection with the addition of 50mM $CuSO_4$

to the culture media. A fraction of cells were collected at 24, 48, 72, and 96 hours post induction (hpi). TRIzol reagent was applied to cells and RNA was extracted using the Zymo Research Direct-zol Kit as per the manufactures protocol (with the DNaseI step included in the extraction).

RESULTS

CHARACTERIZATIONS OF VIRAL POPULATIONS BY LONG- AND SHORT-READ SEQUENCING

To generate populations of Flock House virus particles that contained a high percent of defective genomes, S2 cells were infected with three distinct lines of previously characterized FHV inocula at extremely high MOIs. A control, 'wild type', population was generated by expanding a homogeneous passage 0 population (particles formed from a plasmid transfection) at a lower MOI. Infections proceeded for three days, after which, particles were harvested and purified. Subsequently, RNA was extracted from a fraction of the purified particles to characterize the identity and composition of DI-RNAs in each population. Using gene specific primers against both RNA1 and RNA2, total viral RNA was reverse transcribed (RT-PCR) and PCR amplified to create double stranded cDNA libraries. Samples were then processed and sequenced using Oxford Nanopore Technology's (ONT) long read MinION sequencer. Using the standard protocols, nanopore reads were collected over approximately a 20 hour time period, generating >1.2 million reads of which 742,740 passed ONT's filter. Reads were then mapped to the Flock House virus genome using BBMap, of which at least 95% mapped to either RNA1 or RNA2 (**Table 3.1**).

ONT MinION	Wild Type	DIP_{P6}	DIP_{P7}	DIP_{P8}
Total Reads	27,721	47,280	60,829	35,589
RNA1	*46.4%*	*39.9%*	*52.0%*	*57.7%*
RNA2	*48.9%*	*57.8%*	*46.3%*	*40.3%*
UnMapped	4.7%	2.3%	1.7%	2.0%

Table 3.1: Read mapping with ONT's MinION sequencer.

The number of demultiplexed nanopore reads passing quality filters for the selected samples are shown. Reads were end-to-end mapped to the FHV genome using the BBMap suite. The percent of total reads mapped to each genome (RNA1 or RNA2) as well as unmapped reads are shown.

Long read sequencing allows for the identification of the location of deleted nucleotides, number of deletions per genome, and relative abundance of each species. The number of deletions per genome and their relative frequency is shown in **Figure 3.1A.** Here, the 'wild type' population contained at least 86% full length (zero deletions) RNA1 genomes and >92% full length RNA2 genomes, as indicated by the green bars. The three 'defective enriched' (DIP) populations mapped <8% full length RNA1 and <36% full length RNA2 genomes for each sample. As MinION sequencing requires a series of processing and amplification steps, Northern blot (NB) analysis on

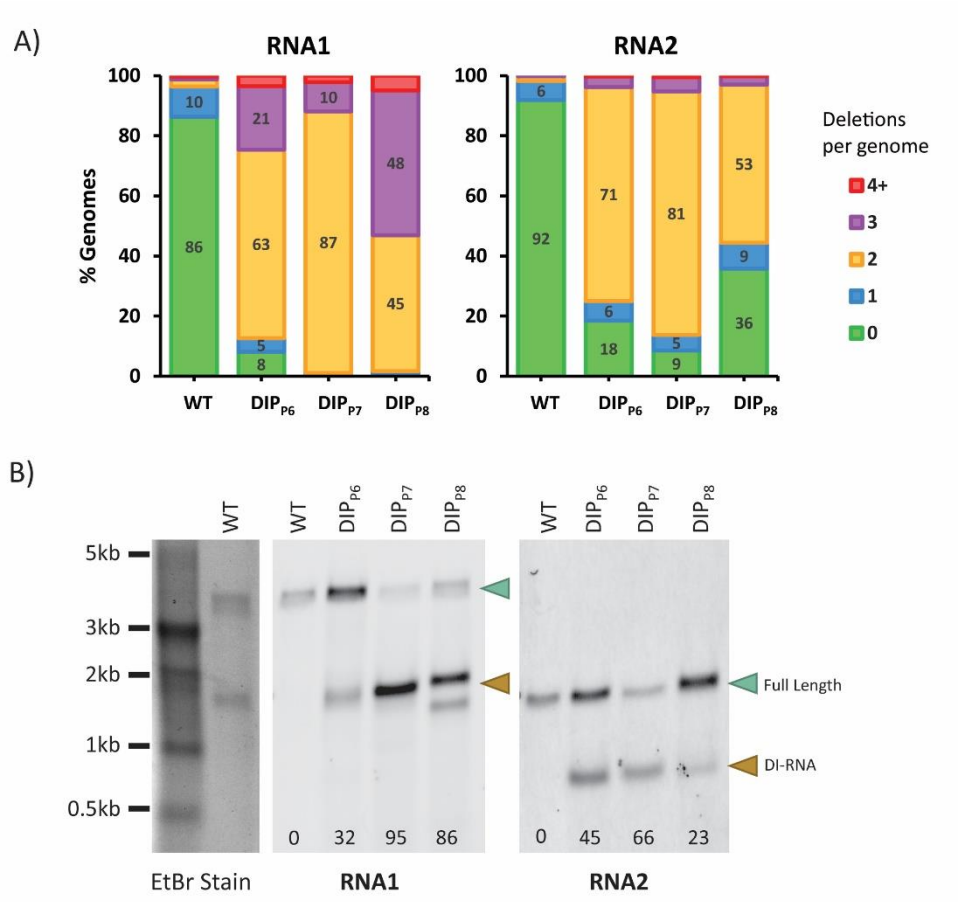


Figure 3.1: Characterizations of DVGs in various FHV populations.

(A) Full length genome characterization using ONT's MinION indicates the relative frequency of DI-genomes within the population. **(B)** Northern blot analysis of FHV RNA1 and RNA2 extracted from purified FHV particles. Percent abundance of DI-RNA species was calculated based off total lane band intensity and indicated under each sample. A ssRNA ladder and total RNA from WT particles was ran concurrently on a 1.2% agarose gel and stained with ethidium bromide post electrophoresis (left).

RNA purified from particles allowed for confirmation of the relative abundance of defective genomes per population (**Figure 3.1B**). Northern blot probes were designed to target a conserved region of the positive strand of either RNA1 or RNA2 and contained one Cy5 (for RNA1) or Cy3 (for RNA2) fluorophore per probe. NB analysis confirmed that the DIP populations contained between 32-95% defective RNA1 and between 23-66% defective RNA2 genomes. Overall this indicated that all three populations of FHV had a relatively high percentage of defective genomes.

While long read MinION sequencing allowed for full genome analysis and correlation between deletion events, the high error rate (~7%) of the platform meant that the precise nucleotide junctions of recombination events was not as clear. Therefore, to specifically determine the precise nucleotide boundaries of the recombination events I applied the high throughput, short read Illumina RNA-seq approach using ClickSeq. Purified RNA samples were processed using the standard ClickSeq protocol to produce NGS libraries and sequenced on an Illumina NextSeq 550 obtaining single end 150 nucleotide long reads (SE 1x150). At least 8.3 million reads per sample were generated; of which ~97% passed quality filtering and trimming (**Table 3.2**). Reads were then mapped to the Flock House virus genome and the host (*D. melanogaster*) using Bowtie end-to-end mapping[166]. ViReMa[45] was then applied to the remaining reads to identify recombination events. The percent of reads mapping to each genome for each sample is shown in **Table 3.2**. 'Wild type' particles only mapped 0.1% of reads to recombination events while the three DIP population mapped 3.3%, 6.8%, and 6.0% recombination events, respectively.

Illumina (ClickSeq)	Wild Type	DIP_{P6}	DIP_{P7}	DIP_{P8}
Raw Reads	11,110,063	8,843,494	8,315,603	15,837'010
Filtered Reads	10,811,564	8,603,423	8,058,652	15,349'497
FHV Mapping	91.7%	89.5%	72.4%	80.1%
RNA1	*68.0%*	*74,9%*	*65.2%*	*58.8%*
RNA2	*23.7%*	*14.6%*	*7.2%*	*21.3%*
Host Mapping	6.1%	2.7%	12.5%	8.0%
Recombination Events	**0.1%**	**3.3%**	**6.8%**	**6.0%**
RNA1 - RNA1	*0.0%*	*2.4%*	*5.8%*	*5.5%*
RNA2 - RNA2	*0.0%*	*1.0%*	*1.0%*	*0.5%*
UnMapped	0.0%	0.0%	0.1%	0.0%

Table 3.2: Read mapping with ClickSeq.

Quantity of reads generated from an Illumina NextSeq run for each sample are tabulated. Reads were mapped to either FHV or the host using Bowtie. Remaining reads were then processed using ViReMa to identify FHV recombination events. Data is shown as percent of reads per total number of filtered (processed) reads.

Using the data generated from ViReMa, I calculated the frequency that each specific nucleotide is deleted across the whole genome. Heat map plots were generated for each sample and are shown in **Figure 3.2** whereas the yellow color indicates regions that are removed more frequently. As I showed previously in **Chapter 2**, there are certain deleted regions that are conserved in the DI- genomes between all samples (deletions laying outside of important regulatory elements), but the frequency and precise locations of the recombination events are slightly variable between all samples. To show how the short read ClickSeq data correlated with the long read MinION sequencing data, the top three most abundant full length genomes (MinION) are displayed below the recombination heat maps (ClickSeq) for each respective DIP sample (**Figure 3.2**). Overall, the correlation between the two data sets is good with Pearson coefficients of 0.77, 0.92, and 0.82 for RNA1; and 0.91, 0.94, and 0.87 for RNA2 of each sample set respectively. To illustrate this, I bolded the top five most common recombination junctions identified in the ClickSeq data on the most common genomes identified with long read sequencing.

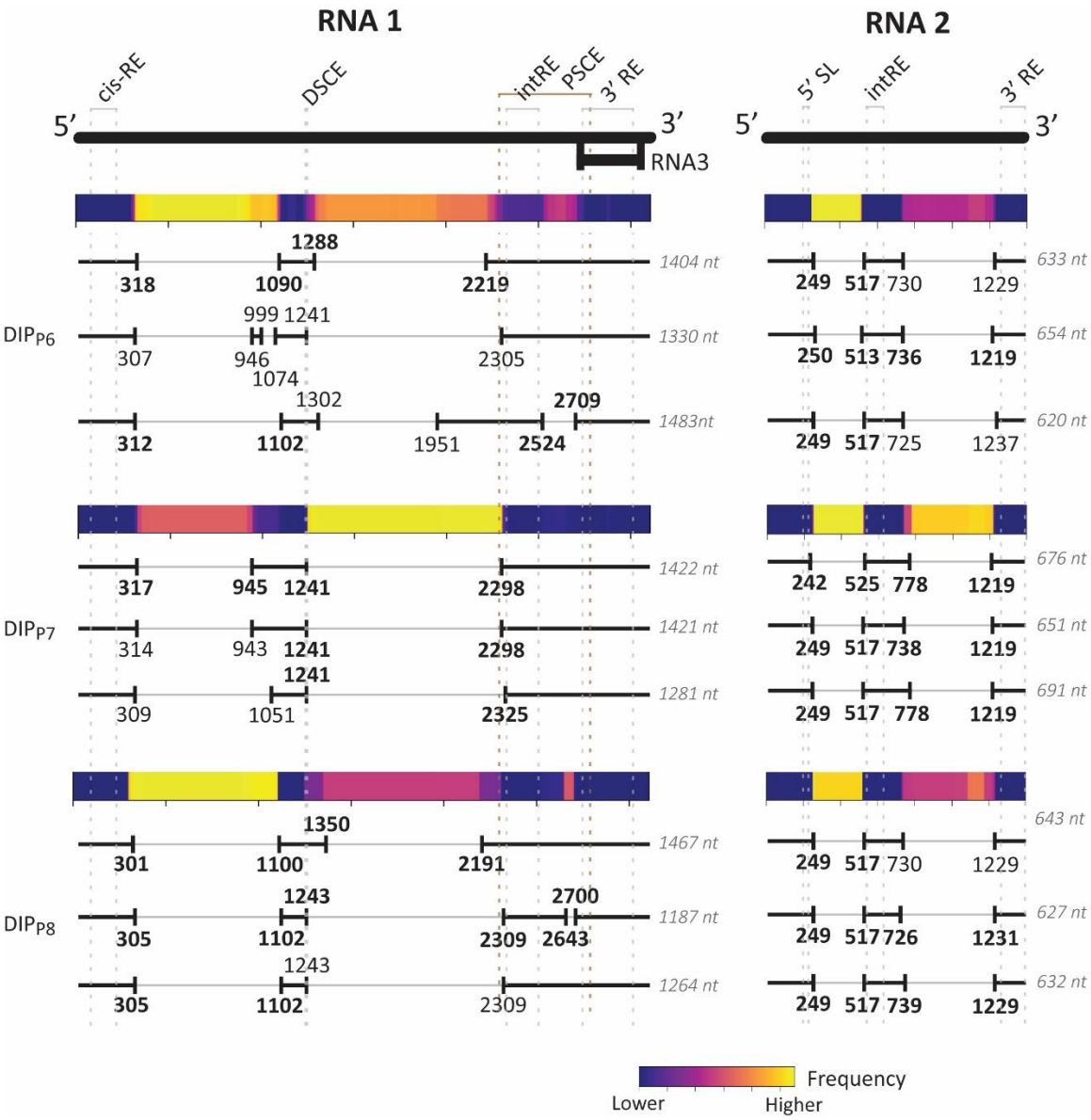


Figure 3.2: Conservation HeatMaps and most common DI- Genomes.

Colored heatmaps represent the frequency with which specific nucleotides in the FHV genome are deleted after recombination using sequencing data extracted from ClickSeq. Line models are graphical representations of the top 3 most common defective genomes identified with MinION sequencing. Location of deletion events are depicted with a lighter grey line. Numbers indicate locations of the recombination boundaries. Nucleotide length of each genome is shown to the right of each representation. Vertical dashed lines indicate boundaries of functional RNA motifs. *cis*-RE: cis-Regulatory Element; DSCE/PSCE: Distal/Proximal Subgenomic Control Elements; intRE: internal Response Element; 3' RE: 3' Response Element; 5' SL: 5' Stem Loop

NUCLEOTIDE LENGTH OF DEFECTIVE GENOMES IS PRESERVED IN MOST DI-RNAS

With MinION sequencing the length of each defective genome identified can be calculated. The histogram plot shown in **Figure 3.3** indicates the frequency of genomes with a specific calculated length. Defective RNA2 genomes cluster heavily around a length of 625-675 nucleotides for all three DIP populations. For the defective genomes in RNA1 there is slightly more variation in length between each sample. For DIP_{P6}, DI-RNA lengths cluster around 1400nt while the population DIP_{P7} significantly favors genomes that are 1400nt in length. The third population, DIP_{P8}, has two highly represented genome lengths of 1175nt and 1450nt (which can also be seen in the Northern blot analysis of **Figure 3.1B**).

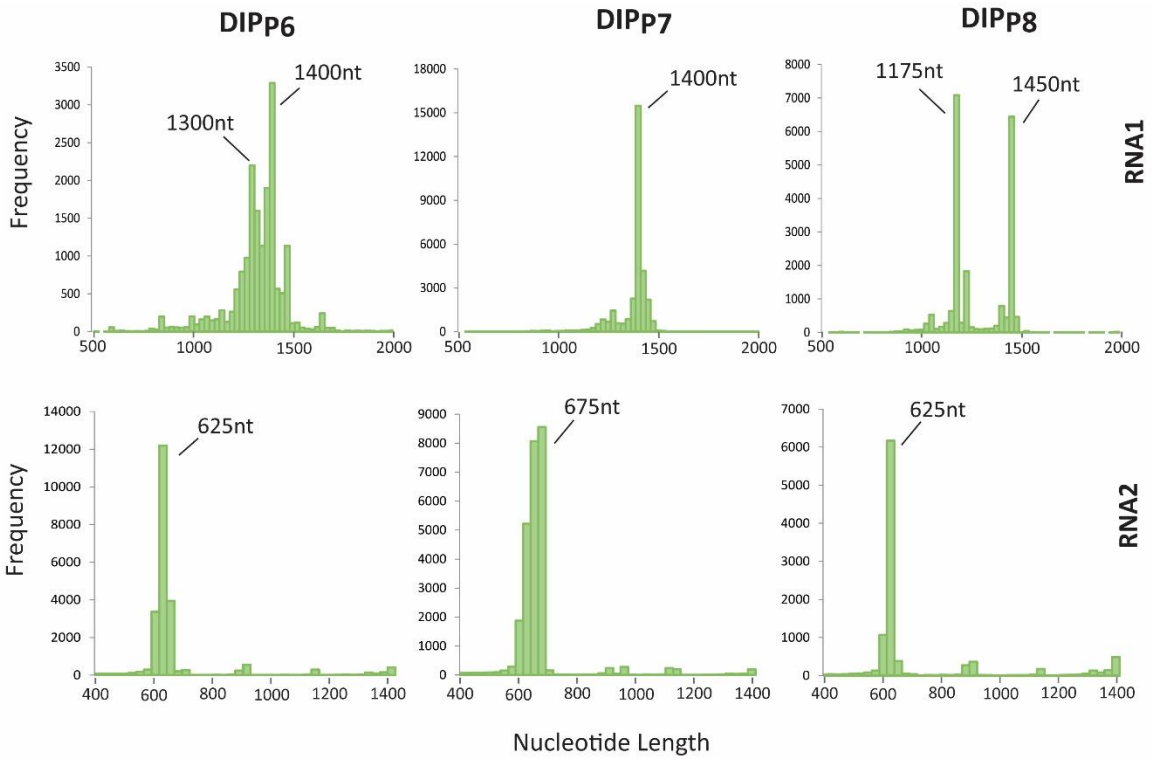


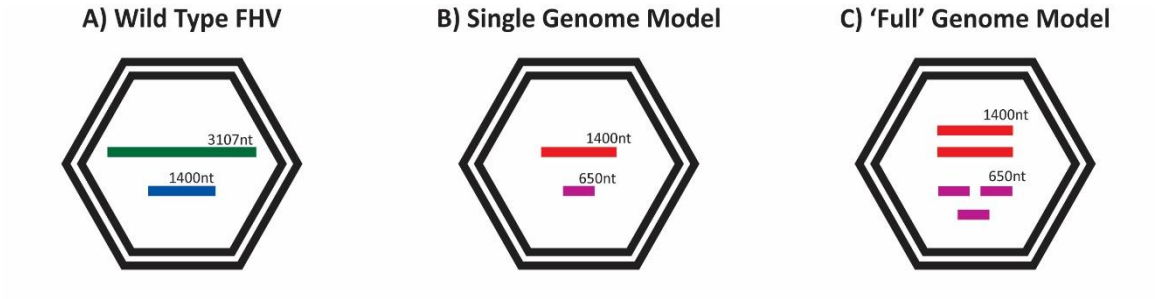
Figure 3.3: Histogram of defective genomes nucleotide lengths.

The nucleotide length of defective genomes for each RNA obtained by MinION sequencing is calculated and plotted. Lengths are binned every 25nt and most frequent lengths are indicated. Frequency is provided as the count of all reads. A full length RNA1 is 3107nt and a full length RNA2 is 1400nt.

HYPOTHESIZED PACKAGING OF DEFECTIVE VIRAL GENOMES

To understand what a Flock House virus defective particle 'looks like' and to determine the amount of RNA that DIPs package I wanted to employ a novel variation of native mass spectrometry (native MS). Native MS is a powerful tool that can characterize complexes in their natural states, preserving protein interactions, enabling the extraction of more than just molecular weight information as with denaturing MS approaches[205]. Historically, standard native MS techniques have a limit to the size of the complex that could be accurately and sensitively analyzed (range of 20-250kDa). This is particularly important when trying to analyze entire viruses which fall far outside of this range. To further complicate things, MS relies on the positive charge of the specimen to pull particles through the detector. This makes it exceptionally difficult to run entire native viruses as the protein-nucleic acid complex results in a low number of charges. Therefore, trying to not only analyze an extremely large complex but a low charge complex has been an extreme challenge.

Dr. Albert Heck from Utrecht University and his group (in collaboration with Thermo Fisher Scientific) have modified the current native MS instrument (a Q Exactive Plus (QE) Orbitrap) to overcome these exact problems. Here, they have applied technical modifications to the QE Orbitrap to better resolve larger protein-nucleic acid complexes in what they call ultra-high mass range spectrometry (QE-UHMR)[172]. Amazingly, with this technique they were able to accurately resolve (within ± 1kDa) the masses of prokaryotic 30S, 50S, and 70S ribosome particles (masses of 0.8- to 2.3MDa), hepatitis B virus, and more importantly, intact Flock House virus particles which has a mass of 9.3MDa[172].



Flock House virus (WT)			
180	x	capsid protein β	7058.0 kDa
180	x	peptide-γ	791.3 kDa
1	x	RNA1	1000.6 kDa
1	x	RNA2	449.6 kDa
240	x	Ca^{2+}	9.6 kDa
Total			**9309.1 kDa**

Flock House virus (DIP)			
180	x	capsid protein β	7058.0 kDa
180	x	peptide-γ	791.3 kDa
1	x	DI-RNA1	449.6 kDa
1	x	DI-RNA2	208.7 kDa
240	x	Ca^{2+}	9.6 kDa
Total			**8517.2 kDa**

Flock House virus (DIP)			
180	x	capsid protein β	7058.0 kDa
180	x	peptide-γ	791.3 kDa
2	x	DI-RNA1	899.2 kDa
3	x	DI-RNA2	626.1 kDa
240	x	Ca^{2+}	9.6 kDa
Total			**9384.2 kDa**

Figure 3.4: FHV particle packaging models and expected masses.

(A) Calculated mass of each individual component of a wild type FHV particle containing one RNA1 genome and one RNA2 genome and summed up total of an entire particle. **(B)** The theoretical breakdown and mass of a 'Single Defective Genome' particle where only one of each defective RNA is packaged. Length of the defective genomes are based on the most frequent length identified experimentally. **(C)** The 'Full Genome' particle packages multiple genomes to encasidate the same number of nucleotides as a wild type particle (4,507nt). There can be many combinations of genomes (both full length and defective) to fulfill this requirement.

I therefore set out to determine the mass of the defective FHV particle. A wild type FHV particle, that packages one each RNA genomes should resolve to a mass of ~9.31MDa (component breakdown shown in **Figure 3.4A**). As the mechanism into FHV RNA packaging is still largely unknown, there are many possibilities into how these defective genomes are packaged into particles. One hypothesis is that only one genome of each RNA is packaged into a particle. In this 'single genome model', a defective genome containing particle could contain combinations of each of the genomes. This could result in a particle with an approximate mass of 8.52MDa (one DI-RNA1 and one DI-RNA2, **Figure 3.4B**), 9.07MDa (one full RNA1 and one DI-RNA), or 8.76MDa (one DI-RNA1 and one full RNA2). If this was the case, we would expect to see heterogeneity in the native MS data resolving multiple lower mass peaks. This model implies that the virus particle

has a mechanism to sense each genome and selectively packages one of each. Furthermore, packaging would have little implications on the formation and accumulation of defective genomes as long as that genome retained the specific encapsidation signal. Another hypothesis is that particles will package RNA until they are 'full'. Here, particles will capture as many genomes as they need to fulfill a certain nucleotide requirement, which in the case of FHV, is 4507 nucleotides. While there could be a huge range of combinations when packaging multiple RNA segments, one example is shown in **Figure 3.4C.** For example, packaging two DI-RNA1 genomes and three DI-RNA2 genomes would sum to a mass of approximately 9.38MDa (using the most common DI genomes lengths identified experimentally: 1400nt for RNA1 and 650nt for RNA2). This is just one example as I have shown with the sequencing data the lengths of defective RNAs within a population are not that static. Therefore, the combination of which genomes are being packaged concurrently dictates if the total number of nucleotides is fulfilled. With this hypothesis, the spectra produced from these particles would be homogenous and practically indistinguishable from the spectra produced by wild type particles. Lastly, while highly unlikely, particles could be quite promiscuous having no selection in the genomes that they package, producing a wide, heterogeneous spectra with little to no distinct peaks.

NATIVE MASS SPECTROMETRY INDICATES DI- PARTICLES PACKAGE RNA TO 'CAPACITY'

In collaboration with Dr. Heck and his graduate student, Tobias Wörner, we were able to apply their ultra-high mass native mass spectrometry approach to determine the mass of defective FHV particles. Here, we used the previously characterized populations of particles that contain high percentages of defective genomes (DIP_{P6}, DIP_{P7}, and DIP_{P8}). As expected and previously shown[172], the wild type FHV population resulted in well-resolved, sharp peaks around 42,000 *m/z* with a calculated mass of 9,357 ± 1 kDa (**Figure 3.5**). This agrees with the predicted mass of a complete particle which has a theoretical mass of 9,309 kDa when summing up all the individual components. Surprisingly, the spectra produced by the defective heavy populations were almost identical to the wild type spectra where masses were resolved to 9,357 ± 2 kDa for DIP_{P6} and 9,356 ± 2 kDa for DIP_{P8}. Both populations of defective particles only showed a single group of well resolved peaks with no smaller groups at lower *m/z* values indicating that the

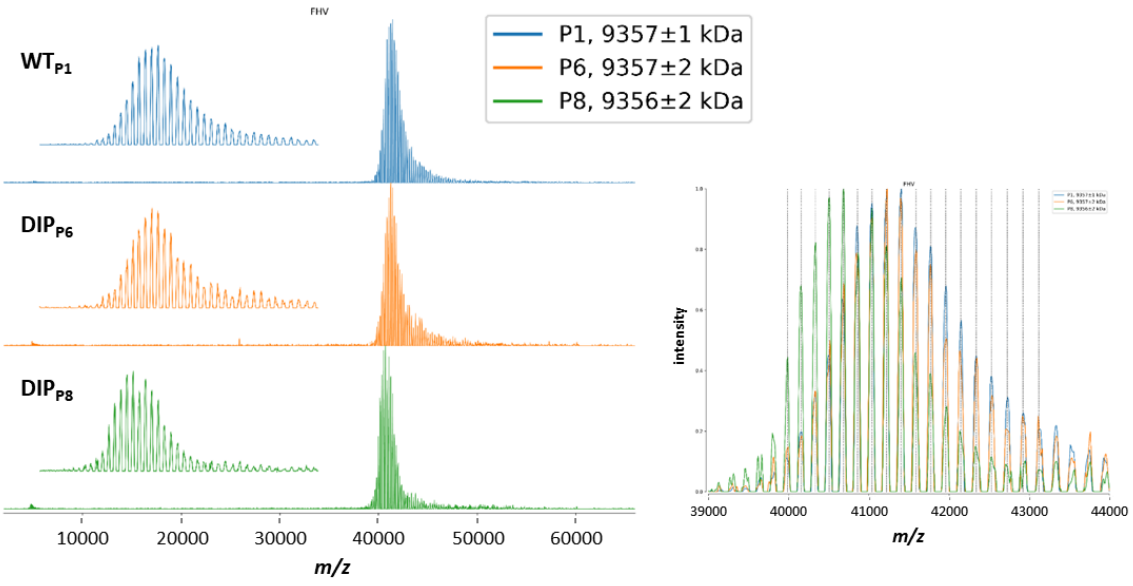


Figure 3.5: Ultra-high mass native MS of defective particles.

(A) Native mass spectra of wild type FHV (blue) and particles full of defective genomes (orange and green). **(B)** Overlay spectra detected around 42,000 *m/z* indicates all 3 populations of particles resolve peaks at similar charge states. (work performed by T. Wörner)

analyzed populations were relatively homogenous in mass. This data indicates that FHV particles, whether they are full of wild type genomes or defective ones, have a propensity to package a certain number of nucleotides supporting a 'full' genome model (**Figure 3.4C**).

CRYO-ELECTRON MICROSCOPY ELUCIDATES THE STRUCTURE OF THE FHV DEFECTIVE PARTICLE

Native MS provided an insight into the potential packaging mechanism of defective genomes into FHV particles where I found DIPs have a similar mass compared to wild type particles. However, I also wanted to determine if there were any structural differences between wild type and DIPs. Therefore, I wanted to apply a cryo electron microscopy (cryo-EM) approach to determine the structure of a defective FHV particle. In collaboration with Dr. Peter Stockely and his group at the University of Leeds, we have been able to reconstruct two independent populations of DIPs (**Figure 3.6**). Cryo-EM models were resolved to 4.3Å for wild type particles, and 4.6Å and 4.1Å for DIP_{P7} and DIP_{P8} respectively. To determine the average density within a FHV particle, average radial density was calculated. Density was determined as a function of radius and plotted for the wild type particle, DIP_{P7}, and DIP_{P8} (**Figure 3.7**). The increase in radial density

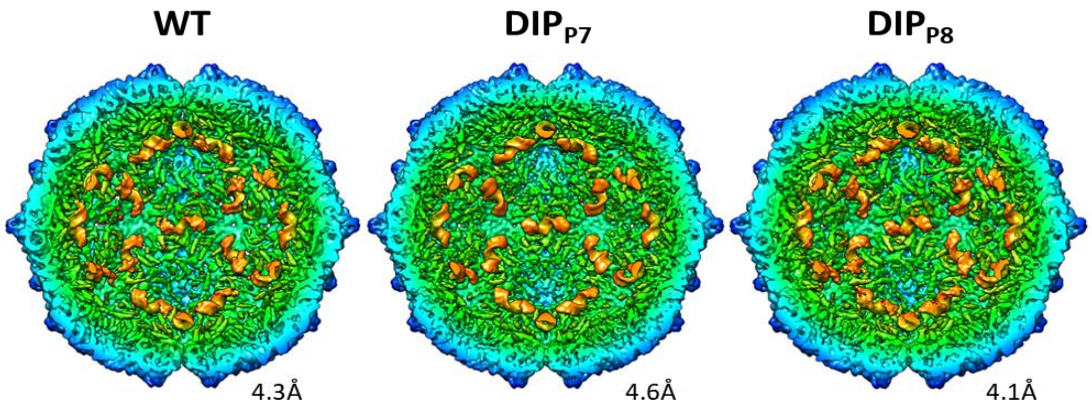


Figure 3.6: Symmetrical reconstruction of the defective FHV particle.

Cryo-EM reconstructions of wild type FHV particles and two distinct populations of particles containing high populations of defective genomes. Structures corresponding to the capsid protein are colored in blue-to-green while packaged RNA is indicated in orange. (work performed by D. Maskell)

labeled ‘A’ corresponds to the internal (unstructured) RNA. Peak B corresponds to the structured dodecahedral RNA cage, and peak C corresponds to the density of the outer capsid protein. Analysis of the models indicates no significant difference between all three populations of particles. These data correspond with the native MS data indicating that DIPs have a similar RNA density and structure within particles.

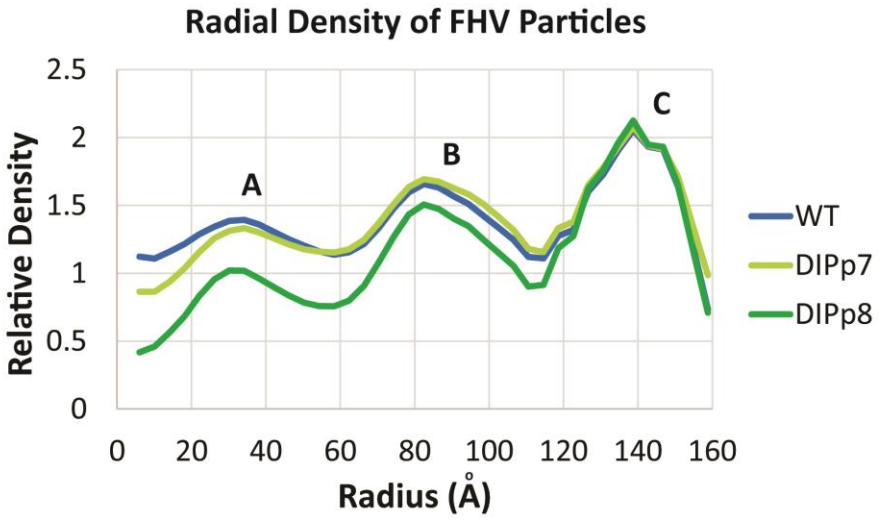


Figure 3.7: Radial density of defective FHV particles.

Average density as a relative function of radius is plotted for the three populations of FHV particles using the reconstructed cryo-EM models. Peaks of increased density are labeled to corresponding structures: A- unstructured internal RNA, B- dodecahedral cage, C- capsid protein.

LENGTH DEPENDENCE ON THE REPLICATION OF DEFECTIVE VIRAL GENOMES

Structural analysis of defective particles indicates that packing is one of the barriers to the accumulation of certain DI-RNAs. I further wanted to explore the hypothesis that accumulation of multiple deletion RNAs is a product of their shorter size. Other researchers have speculated that shorter genomes can be replicated much faster as it takes the polymerase less time to copy an entire transcript[189]. The faster replication would result in the accumulation of these shorter genomes within the cell and stoichiometry would dictate that these genomes would

be packaged into particles more often and therefore, passaged to the next cell. I wanted to explore the hypothesis that the quick accumulation of these double deletion genomes is a product of replication kinetics.

To test this I decided to focus on the defective genomes produced in RNA2 as studies have shown that the defective genomes of this RNA are less variable (**Chapter 2**). Three defective RNA pMT- plasmids of RNA2 based off the most common deletion events identified previously were created: 1) Δ1: a deletion of nucleotide 249 to nucleotide 517, 2) Δ2: a deletion of nucleotide 736 to nucleotide 1219, and 3) Δ1-2: a clone containing both of the previous deletion events (as shown

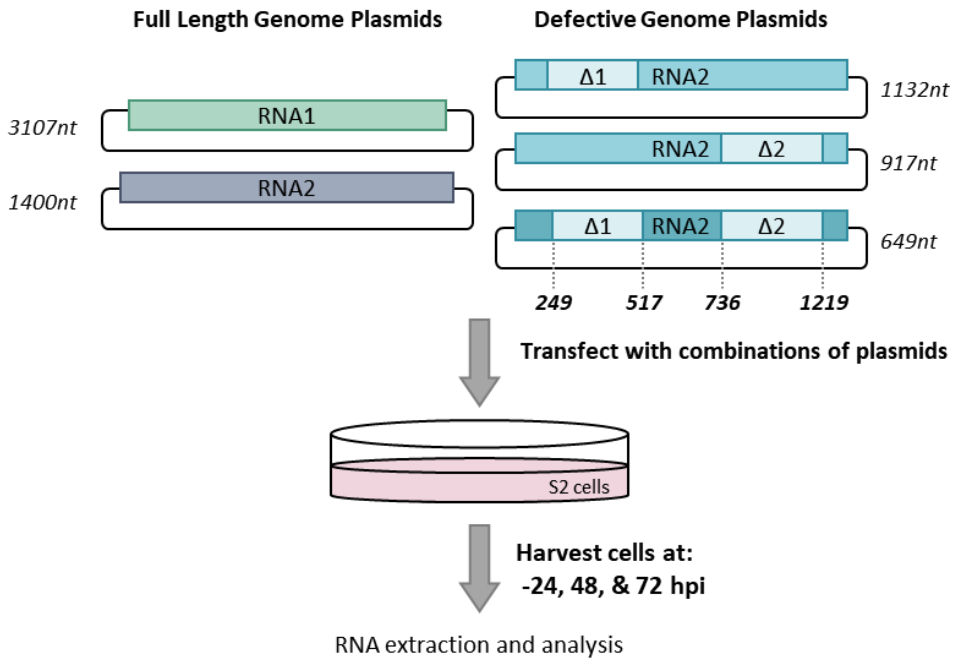


Figure 3.8: Experimental design to determine influence of genome length on replication.

pMT vector plasmids containing the FHV RNA2 genome were engineered removing certain portions of the RNA2 genomes. These mimic the most common deletion events found during *in vivo* FHV infections. The RNA2-Δ1 clone is lacking nucleotides between 249 and 517. RNA2-Δ2 is a deletion event between nucleotides 736 and 1219. RNA2-Δ1-2 is a clone containing both deletion events. Nucleotide lengths of the RNAs is indicated next to the plasmid schematics. Combinations of the indicated plasmids were transfected into S2 (*D. melanogaster*) cells and RNA expression was induced with copper sulfate 24 hours post transfection. Cells were then harvested at 24, 48, and 72 hours post induction and RNA from cells was extracted for analysis.

in **Figure 3.8**). Combinations of these plasmids as well as full length plasmids containing either RNA1 or RNA2 were transfected into S2 cells and cells were harvested at 24, 48, and 72 hours post induction of expression (hpi). RNA was extracted from cells and analyzed by Northern blot and by MinION sequencing (**Figure 3.8**).

To determine if each of the defective genomes could be replicated by the viral polymerase, cells were transfected with plasmids containing RNA1 (which encodes for the RNA dependent RNA polymerase) and each of the deletion genomes individually. NB analysis was then performed on the total cellular RNA extracted 72 hours post induction (**Figure 3.9A**, lanes 1-3) and quantified relative abundance of each ΔRNA2 is shown at the bottom of each lane (normalized to RNA1 intensity to adjust for loading). Replication analysis revealed that all three deletion genomes are replicated by the RdRp but generally the double deletion (Δ1-2) genome is more abundant than either of the two other DI- genomes. More replicates would be needed to calculate significance.

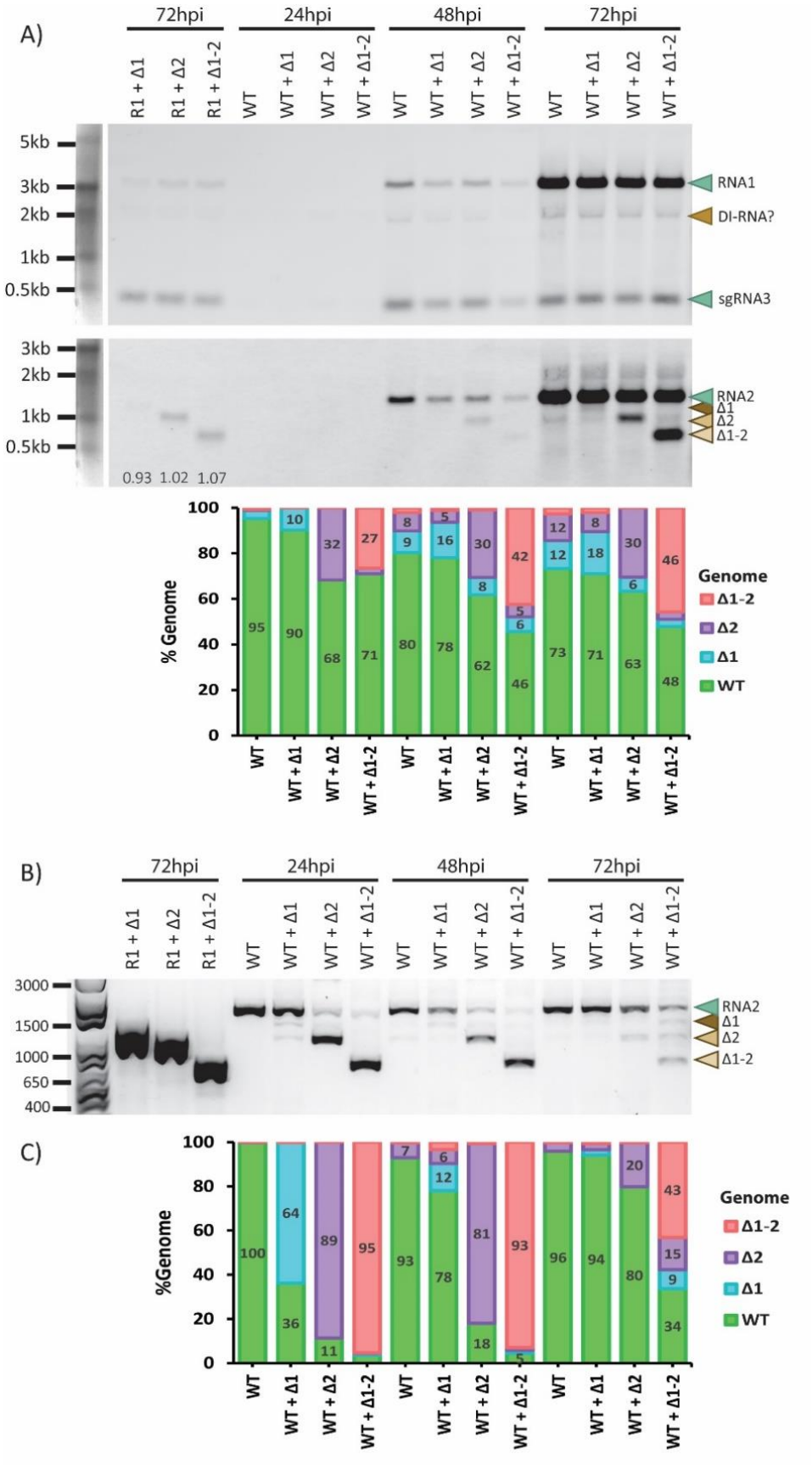
I further wanted to test the replication kinetics of the defective RNA2 genomes in the presence of full length, wild type RNA2. Here, I co-transfected plasmids containing RNA1, RNA2, and one of each ΔRNA2 genome into cells, harvested at multiple time points post induction, and analyzed RNA abundance by northern blot (**Figure 3.9A**, lanes 4-15). At 24 hpi (**Figure 3.9A**, lanes 4-7), no RNA was visible despite total cell RNA was quantified and loaded equally across all lanes. By 48 and72 hours post induction, RNA quantities corresponding to the transfected plasmids are within the threshold of sensitivity of the NB analysis and therefore can be visualized. Relative abundance of each genome (WT, Δ1, Δ2, and Δ1-2) for all lanes was quantified and plotted in the bar graph below as a percent of that lane's intensity. At these time points, the transfected defective genomes can be replicated in the presence of wild type RNA2. The double deletion genome (Δ1-2) was relatively more abundant in its population when compared to the other

defective genomes (**Figure 3.9A**). The first deletion (Δ1) genome appears to be the least favored. Interestingly, while shorter genomes are trending to be faster replicators, none of these genomes are more abundant than wild type RNA2. This seems to imply that while length has some implications on the speed of replication there are other characteristics of the genome that regulate its accumulation.

While the 24 hpi sample set did not have any visible bands corresponding to the transfected RNAs, quantification of the NB analysis still produced values corresponding to those genomes and values trended with our findings of later time points. To validate and confirm this result, I reverse transcribed the RNA, PCR amplified using gene specific primers against RNA2, and visualized the products on an agarose gel to improve the sensitivity to these lower quantity RNAs (**Figure 3.9B**). These cDNA products were furthermore processed and sequenced using the long-read MinION platform to precisely calculate the prevalence of each genome within each condition (**Figure 3.9C**). This data indicates that at 24 hpi, the Δ1 genome accounts for 64% of all RNA2 genomes, 89% for Δ2, and 95% for Δ1-2. Similarly, I calculated the abundance of each genome for the subsequent two time points. Comparing the data generated from MinION sequencing and the NB analysis, I see similar trends when comparing the abundance of the defective genomes but generally MinION sequencing over represents the frequency of these shorter genomes.

Figure 3.9: Defective genomes of varying lengths are replicated by FHV RdRps

(A) Northern blot analysis of RNA extracted from cells transfected with the indicated combinations of FHV plasmids and collected at multiple hours post plasmid induction (hpi). WT indicates RNA1 and RNA2 plasmids. Top blot shows RNA1 (Cy5), bottom blot for RNA2 (Cy3). An ethidium bromide stained ssRNA ladder is shown to the left of the blots. Bands corresponding to each of the indicated RNA2 genomes were quantified and percent abundance is shown in the stacked bar graph. **(B)** RNA from cells was reverse transcribed, PCR amplified using RNA2 gene specific primers, and ran on a 1.2% agarose gel. **(C)** MinION sequencing of RNA2 cDNA (from **B**). Frequency of each genome is shown. (next page)


A)
72hpi
24hpi
48hpi
72hpi
R1 + Δ1
R1 + Δ2
R1 + Δ1-2
WT
WT + Δ1
WT + Δ2
WT + Δ1-2
5kb
3kb
2kb
1kb
0.5kb
RNA1
DI-RNA?
sgRNA3
RNA2
Δ1
Δ2
Δ1-2
0.93 1.02 1.07
% Genome
Genome
Δ1-2
Δ2
Δ1
WT
B)
3000
1500
1000
650
400
C)
%Genome

Importantly, while I am generally able to show that each of the variations of defective genomes have the ability to be replicated this experiment has many caveats; therefore, I am not able to draw any solid conclusions from this data. Primarily, transfection efficiency can significantly change the outcome of abundance of each genome for each condition. To address this, gene levels could be normalized to quantities of plasmids transfected into cells or more accurately clones could be made to contain each genome within one plasmid.

CONCLUSIONS

In this chapter, I sought to elucidate the selective pressures that are imposed upon DI-RNAs to get a better understanding of why there is an accumulation of a certain 'species' of genomes. I began by generating three populations of FHV that had a high frequency of defective genomes, which I characterized using long read MinION sequencing and short read ClickSeq. To take a closer look at what these populations of defective particles look like I used a combination of native ultra-high mass spectrometry and cryo-electron microscopy. Here, I was able to show that DIPs have the same mass as wild type particles and are visually indistinguishable. This seems to indicate that the mechanisms involved in packaging can impose a restrictive barrier to the types of defective genomes that get encapsidated. While we have yet to examine the species of DVGs present inside cells during infections, only genomes that meet a length requirement get packaged and transmitted.

Particle assembly is a key component of the viral life cycle so ensuring that a particle encapsidates the correct genomes is critical. While studies have been able to identify the packaging signals in genomes and the regions of the capsid that interact with RNA, the exact mechanism behind this process is still unclear[54, 191-195]. This study suggests that FHV has a type of 'molecular measuring tape' where the virus is able to measure the correct number of nucleotides to package within its capsid. Indeed, analysis of the capsid protein has indicated that the arginine-rich motif (ARM) located in the N-terminus of capsid protein may be implicated in this process[206].

CHAPTER 4: METHODS DEVELOPMENT: NEXT GENERATION SEQUENCING

INTRODUCTION

Next generation sequencing (NGS) is a powerful tool used to determine the nucleotide sequence of genetic materials. Since the technology's initial introduction to the market in 2005 a handful of different platforms have been released all with one goal: provide researchers with a high throughput, accurate, and inexpensive way to sequence nucleic acids (reviewed by Goodwin *et al*.[190]). Broadly, to go from sample to tangible results the NGS pipeline is a three step process: 1) sample preparation (also called library preparation), 2) obtaining the sequences using any platform of choice (ie. Illumina, Ion Torrent), and 3) data analysis.

Sample preparation and library construction is a fundamentally step in NGS as nucleic acids need to be converted to a form compatible with the sequencing platform. Very simply, the RNA or DNA needs to be fragmented to a shorter size, converted to double stranded DNA, and sequencing adapters need to be attached onto those fragments. While this is an over simplification of the process, it is arguably the most important step in the entire process. There are many different protocols currently on the market which all generally follow this same basic process (reviewed by Head *et al*.[207]). Unfortunately, almost all steps in the library preparation process have been reported to introduce artifacts and biases in the sequencing data (reviewed by van Dijk *et al*.[208]).

Routh *et al*. initially developed a fundamental different approach for NGS library preparation that is based upon click-chemistry, called 'ClickSeq'[157]. Here, I discuss the improvements and optimizations that I have made upon this original method. Furthermore, I have also helped developed two variations of ClickSeq. The first, called Poly(A)-ClickSeq (PAC-Seq), targets and identifies the 3'UTR/Poly(A) junction of mRNAs, and the second, Polymerase Profiling

with ClickSeq (2PC-seq), is a method for profiling an actively replicating polymerase. In this chapter I introduce each technique and provide an understanding to what led us to develop that method. I then provide a detailed protocol. Lastly, I show how we apply the method to help answer biological questions.

CLICKSEQ‡

Replacing fragmentation and enzymatic ligation with click-chemistry to prevent sequence chimeras.

OVERVIEW

In nature, DNA is composed of long polymers of deoxyribose sugars each carrying a nucleobase that are linked together by phosphate groups. However, a number of recent studies have generated DNA and RNA with unnatural triazole-linked backbones that can approach the natural properties of DNA or RNA[209-218]. Such molecules are generated by 'click-ligating' azido- and alkyne-functionalized nucleic acid strands together via Copper-catalyzed Azide-Alkyne Huisgen Cycloaddition (CuAAC), the prototypical Click-Chemistry reaction[219] (**Figure 4.1A**) or Strain-promoted Azide-Alkyne Cycloaddition (SPAAC)[220] (**Figure 4.1B**). There are many variations and types of triazole-linkages (**Figure 4.2**). Some closely mimicking the native structure and base-to-base distance of natural DNA (compare structures A and B in **Figure 4.2**), while others inserting large bulky chemical groups (e.g. structure E in **Figure 4.2**). It is thus remarkable that such nucleic acid templates have been shown to be biocompatible in a number of settings; in vitro using reverse transcriptases[210] and DNA polymerases[157, 217, 221], and in vivo in E. coli[213, 218] and eukaryotic cells[209].

Using click-chemistry to click-ligate nucleic acids together has allowed the generation of DNA templates that might otherwise be unobtainable using biological ligation, for instance, in the

‡ This section was adapted from: Jaworski, E., and Routh, A. (2018). ClickSeq: Replacing Fragmentation and Enzymatic Ligation with Click-Chemistry to Prevent Sequence Chimeras. Methods Mol Biol 1712, 71-85. (see **Appendix C** for publisher's copyright permission)

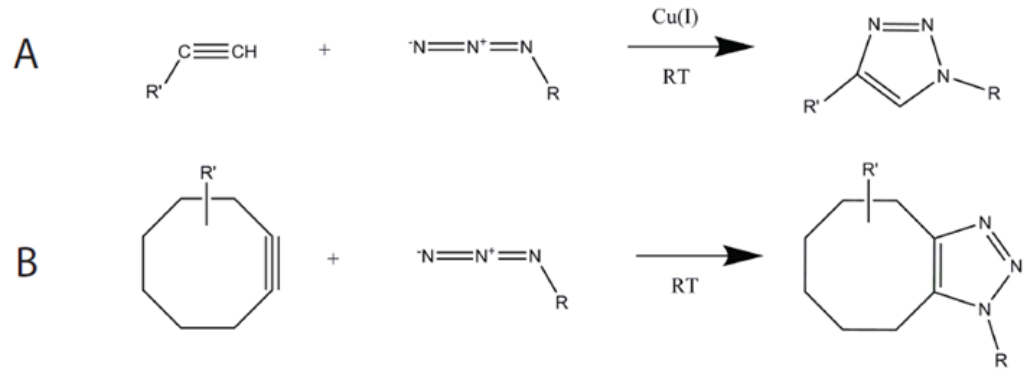


Figure 4.1: Prototypical click-chemistry reactions.

(A) Copper Catalyzed Alkyne-Azide Cycloaddition (CuAAC); and **(B)** copper-free Strain Promoted Alkyne-Azide Cycloaddition (SPAAC).

solid-phase synthesis of very long oligonucleotides. Routh *et al.* recently demonstrated that click-chemistry can be used for the synthesis of Next-Generation Sequencing (NGS) libraries in a process we dubbed 'ClickSeq'[157] (see schematic in **Figure 4.3**). The novel innovation was to supplement randomly-primed RT-PCR reactions with small amounts of 3'-azido-nucleotides to randomly terminate cDNA synthesis and release a random distribution of 3'-azido blocked cDNA fragments. These are then 'click-ligated' to 5' alkyne-modified DNA adaptors via CuAAC. This generates ssDNA molecules with unnatural yet bio-compatible triazole-linked DNA backbones that can be used as PCR templates to generate RNA-seq libraries.

'ClickSeq' was developed in order to address a critical limitation in NGS datasets: abundant artefactual chimeras[222]. Artefactual recombination occurs primarily due to template switching of the reverse transcriptase during RT-PCR and the inappropriate ligation of DNA/RNA fragments during cDNA synthesis[223]. In ClickSeq, azido-terminated cDNA fragments cannot provide substrates for forced copy-choice template-switching during RT-PCR as they lack a free 3' hydroxyl group required for DNA synthesis. Additionally, the azido-terminated cDNA can only be ligated to orthogonally-provided alkyne-labelled DNA oligos and not to other cDNAs.

Consequently, two of the main suspected sources of artefactual recombination in NGS are eliminated[208]. When studying Flock House virus, which is known to undergo extensive recombination in vivo[45, 143], it was demonstrated that artefactual recombination in NGS was reduced to fewer than 3 events per million reads allowing for the confident detection of rare recombination events (as demonstrated in **Chapter 2**)[86, 157]. Consequently, ClickSeq allows us to explore biological systems where the rate of recombination may be much lower than what was previously detectable.

For on-going research, ClickSeq has become a routine method for making RNA-seq libraries in our laboratory. In addition to advantages in avoiding chimeric read formation, ClickSeq does not require any template sample fragmentation. Overall, I have made improvements to this protocol to make it a simple and cost-effective procedure; once the initial click-specific reagents have been purchased the main expense is at the level of plastic-ware and PCR enzymes. This allows for the screening of multiple conditions so that a fully optimized library can be produced.

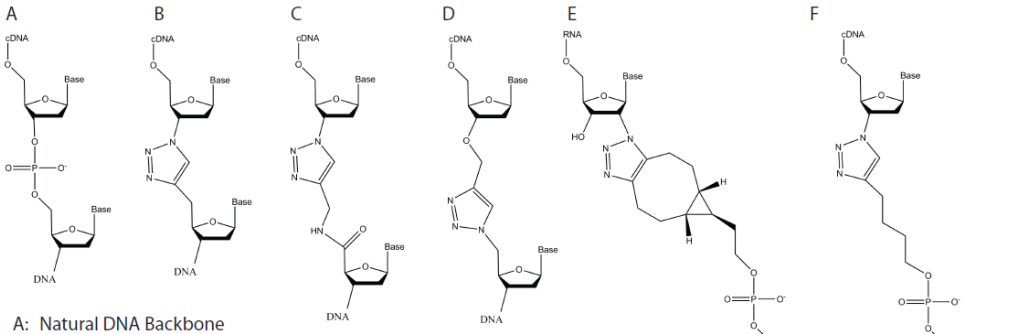


Figure 4.2: Examples of triazole-linked nucleic acids.

A varied range of unnatural triazole-linked nucleic acids generated by click-ligation and demonstrated to be bio-compatible have been reported. B[215], C[221], D[213], E[210], F[156]

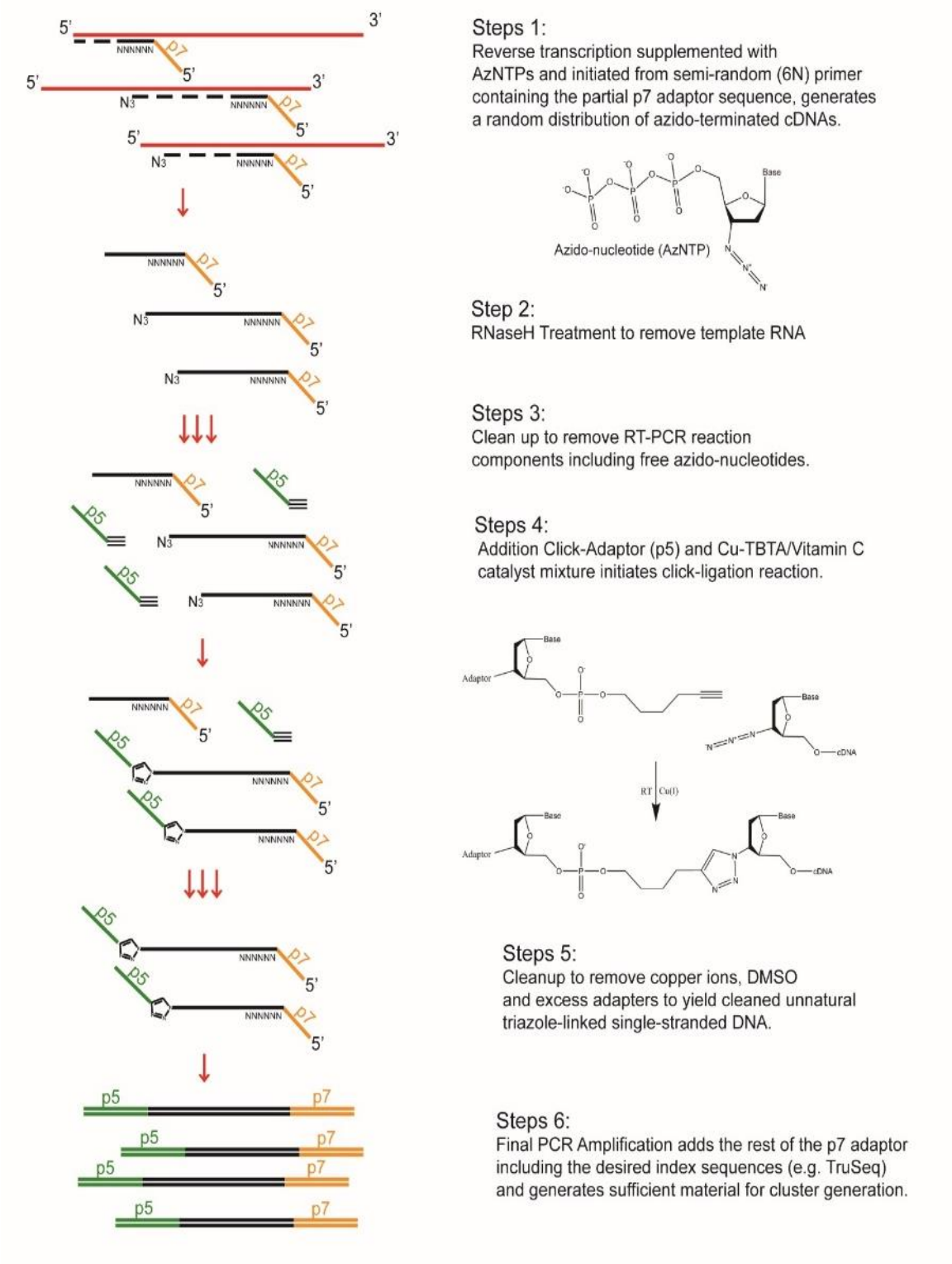


Figure 4.3: ClickSeq Protocol Schematic.

Schematic of the 'ClickSeq' protocol illustrating the individual steps and the Click-ligation reaction.

PROTOCOL

MATERIALS

2.1 Reverse transcription components:

- Deoxyribonucleotide set (dNTPs) (10mM in water)
- 3'-Azido-2',3'-dideoxynucleotides (AzNTPs) (10mM in water) (Trilink Biotechnologies, N-4007, N-4008, N-4009, N-4014). Reagents are stored frozen and mixed thoroughly prior to use. During reverse transcription, the ratio of AzNTPs to dNTPs determines the distribution of cDNA fragment lengths generated (see **Note 4.3**). AzNTP:dNTP mixtures are made by making appropriate dilutions of 10mM AzNTPs in 10mM dNTPs. For example, for a 1:20 10mM AzNTP:dNTP solution add 1µL 10mM AzNTPs to 20µL 10mM dNTPs.
- Reverse transcriptase: Our choice is Superscript II or III (Life Technologies) which is provided with standard reaction buffers.
- RNaseOUT Recombinant Ribonuclease Inhibitor (Life Technologies)
- RNaseH (NEB, or any other)

2.2 Click-chemistry components:

- Click-adapter stock is resuspended in 10mM Tris pH 8.0 and 0.5mM EDTA at 100µM; working solutions of Click-adapter at 5µM in water
- Copper(II)-Tris(benzyltriazolylmethyl)amine complex (Cu-TBTA) 10mM in 55% aq. DMSO (Lumiprobe) or home-made

- 50mM L-Ascorbic Acid is prepared by dissolving 0.44 grams powdered L-Ascorbic Acid in 50mL water. Aliquots are dispensed into 200µL microcentrifuge tubes and stored at -20°C. One aliquot is used fresh per experiment and discarded after use.
- 100% DMSO (e.g. sigma)
- 50mM HEPES pH 7.2

2.3 PCR Reaction

- OneTaq DNA Polymerase 2X Master Mix with standard buffer (NEB, M0482)

2.4 Other reagents and equipment

- Standard non-stick 1.5mL microcentrifuge tubes
- Standard 0.2mL PCR tubes
- E-Gel Precast Agarose electrophoresis system with 2% Agarose gels (Life Technologies).
- Blue light Transilluminator (e.g. Safe Imager 2.0 Blue-Light Transilluminator, Life Technologies)
- 100bp DNA ladder
- Solid Phase Reversible Immobilization (SPRI) magnetic beads, homemade (see **Note 4.16**) or AMPure XP beads (Beckman Coulter, A63880)
- Zymo Gel DNA Recovery Kit (Zymo Research, D4007)
- Qubit fluorimeter (Life Technologies).
- Standard Thermocyclers
- Standard Tabletop centrifuges

2.5 Primers and Oligos:

Primer Name	Sequence	Stock Solution	Working Solution
3'Genomic Adapter-6N (partial p7 Adaptor) (see **Note 4.1**)	GTGACTGGAGTTCAGACGTGTGCTC TTCCGATCT**NNNNNN**	100µM in Water	100µM in Water
Click-Adapter (p5 Adaptor)**	5'Hexynyl-**NNNN**AGATCGGAAGAGCGTCGTGT AGGGAAAGAGTGTAGATCTCGGTG GTCGCCGTATCATT	100µM in TE*	5µM in water
Universal Primer Short {UP_S} (p5 Adaptor)	AATGATACGGCGACCACCGAG	100µM in TE	5µM in water
3'Indexing Primer (remaining p7 Adaptor)***	CAAGCAGAAGACGGCATACGAGAT **XXXXXX**GTGACTGGAGTTCAGACGT GT	100µM in TE	5µM in water

- *TE = 10mM Tris pH 8.0, 1mM EDTA
- **The Click-adapter can be purchased from Integrated DNA Technologies (IDT). HPLC purification is required by the vendor and recommended.
- ***Underlined portion of the 3' Indexing Primer corresponds to the index sequence. Any provided or customized indexes may be used here (we use the Illumina TruSeq Indexes).

METHODS

3.1 Reverse Transcription

1. Input RNA: in principle, any input RNA can be used to generate RNAseq libraries. We have successfully sequenced viral genomic RNA, total cellular RNA, poly(A)-selected RNA, and ribo-depleted RNA. RNA should be provided in pure water, following standard precautions to avoid RNase activity. In our lab, we usually aim to provide ≥100ng of RNA (see **Note 4.2**). No sample fragmentation is required.

2. The reverse transcription is performed using standard protocols, with the exception that the reaction is supplemented with small amounts of azido-nucleotides (AzNTPs). Set up the RT-PCR reaction as follows for a 13µL reaction:

a. ≥100ng RNA

b. 1µL dNTP:AzNTP mixture at 10mM (see **Notes 4.3** and **4.4**)

c. 1µL 3'Genomic Adapter-6N primer at 100µM

d. H_2O to a final volume of 13µL

3. Incubate mixture at 95°C for 2 mins to melt RNA and immediately cool on ice for >1 min to anneal semi-random primer. This high melting temperature is tolerated as small amounts of RNA fragmentation does not diminish efficiency of library generation.

4. Add the following at room temperature for a final reaction volume of 20µL (see **Note 4.5**):

a. 4µL 5X Superscript First Strand Buffer

b. 1µL 0.1M DTT

c. 1µL RNase OUT Recombinant Ribonuclease Inhibitor

d. 1µL Superscript III Reverse Transcriptase

5. Incubate with the following steps:

a. 25°C for 10 mins, (this step should be skipped if using a template-specific primer, see **Note 4.1**)

b. 50°C for 40 mins,

c. 75°C for 15 mins, and

d. Hold at 4°C

6. To remove template RNA, add 2U RNase H and incubate at 37°C for 20 mins, 80°C for 10mins, and then hold at 4°C.

3.2 Azido-terminated cDNA purification

After cDNA synthesis and RNA digestion, the azido-terminated cDNA must be purified away from the AzNTPs present in the RT-PCR reaction mix. These small molecules will be in molar excess of azido-terminated cDNA by many orders of magnitude and will compete for ligation to the alkyne-modified 'click-adaptor' if not completely removed. This can be achieved in a number of ways (see **Note 4.15**); I prefer to use Solid Phase Reversible Immobilization (SPRI) magnetic beads due to their simplicity of use and high throughput ability.

1. Add 36µL SPRI beads to 21µL of the RT-PCR reaction, mix well, and incubate 10mins at room temperature.
2. Pellet beads on a magnetic rack
3. Wash beads twice with 200µL 80% EtOH without disturbing pellet.
4. Air dry magnetic pellet for 5mins.
5. Elute library in 10µL 50mM HEPES pH 7.2 or water (see **Note 4.6**).

3.3 Click-ligation

Following purification of the single-stranded azido-terminated cDNA, the click-ligation reaction is performed to join the 5' alkyne-modified click-adapter on to the 3' end of the azido terminated cDNA. This generates a longer single stranded cDNA with a triazole-ring and a long hexynyl linker in place of a phosphate backbone (see **Figure 4.2F**).

1. First, dilute the azido-terminated cDNA in DMSO and add a large molar excess of the click-adapter using the following volumes:
    a. 10µL azido-terminated cDNA (in HEPES)
    b. 20µL 100% DMSO (see **Note 4.7**)

c. 3µL Click-Adapter at 5µM in water (note: EDTA will chelate copper required in click reaction and so must be minimized)

2. Next, generate the catalyst and accelerant mixture (for multiple samples, prepare a master mixture):

    a. 0.4µL Vitamin C at 50mM

    b. 2µL Cu-TBTA in 55% DMSO.

3. Upon addition of Vitamin C, the Cu-TBTA reagent will turn from a light blue to colorless liquid, indicating the reduction of the Cu(II) ions to Cu(I). Wait 30-60s to ensure full reduction of the copper ions (see **Note 4.8**).

4. Add 2.4µL of the Vitamin C and Cu-TBTA mixture to each cDNA sample to initiate the click ligation reaction.

5. Allow reaction to proceed at room-temperature for at least 30 mins, repeat steps 3.3.2 – 3.3.5 (see **Notes 4.9, 4.10,** and **4.11**).

<u>3.4 Click-ligated cDNA purification:</u>

To remove the components of the click-ligation I use SPRI magnetic beads.

1. Add 68µL SPRI beads to 37.8µL of the click-reaction, mix well, and incubate at room temperature for 10 mins.

2. Pellet beads on a magnetic rack.

3. Wash beads twice with 200µL 80% EtOH without disturbing pellet.

4. Air dry magnetic pellet for 5mins.

5. Elute library in 20µL 10mM Tris pH 7.4 or water.

3.5 Final PCR Amplification:

I have screened a number of cycling conditions and have found the following to give the best results (see **Note 4.14**):

1. Mix at room temperature for a 50µL reaction:
    a. 10µL Clean Click-ligated DNA (in 10mM Tris pH 7.4) (see **Note 4.13**).
    b. 2µL 3' Indexing Primer (1 barcode/sample) at 5µM
    c. 2µL Universal Primer Short {UP-S} at 5µM
    d. 11µL H2O
    e. 25µL 2X One Taq Standard Buffer Master Mix
2. Cycle on a standard thermocycler using the following steps (see **Note 4.14**):
    a. 94° 1 min,
    b. 54° 30sec,
    c. 68° 10 min;
    d. {94° 30sec, 54° 30sec, 68° 2 min} x 12-20 cycles
    e. 68° 5min;
    f. 4° ∞
3. Purify the PCR product with another SPRI bead protocol:
    a. Add 50µL SPRI beads to 50µL of the PCR reaction, mix well, and incubate at room temperature for 10 mins.
    b. Pellet beads on a magnetic rack.

c. Wash beads twice with 200µL 80% EtOH without disturbing pellet.

d. Air dry magnetic pellet for 5mins.

e. Elute library in 20µL 10mM Tris pH 7.4 or water.

3.6 Gel extraction and size selection

Size selection is a critical component for NGS library preparation. Fragments that are too short will not yield map-able cDNA fragments. While fragments that are too long will not cluster properly on the sequencing platform. There are a couple of ways to size select sample libraries. I have found that the most accurate method is by running the amplified cDNA library on an electrophoresis gel and cutting the appropriate band sizes based off a molecular weight ladder. Conversely, SPRI beads following the size select protocol is also possible and is advantageous when processing many samples (see **Note 4.17** for SPRI bead protocol).

1. Add 20µL eluted cDNA library onto a 2% agarose precast pre-stained e-gel. For multiple samples, run empty wells in between each sample to prevent cross-contamination of final libraries. Also run a 100bp MW ladder (e.g. NEB) for size reference.
2. Run using 1-2% agarose protocol for 10mins (E-Gel iBASE Version 1.4.0; program #7)
3. After the run has completed, image gel on blue transilluminator and keep image for records, (e.g. **Figure 4.4A**).
4. Crack open precast gel cassette, and with a fresh/clean scalpel or razor blade, excise the desired cDNA library sizes. In ClickSeq, the total length of adapters are 126bp. Therefore, minimum cDNA library size should be 176bp for 1x50bp SE Illumina run. Example in **Figure 4.4B** shows a library excised from 200-300bp (lanes 2-3) for a 1x75bp SE Illumina run on a HiSeq or cut from 400-600bp (lanes 4-7) for a 1x150bp SE Illumina run.

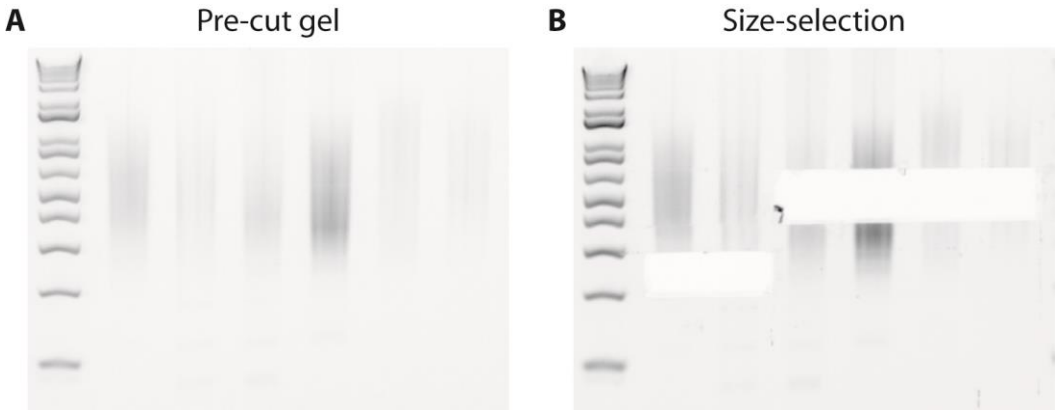


Figure 4.4: Gel electrophoresis of a final cDNA library.

(A) The library should appear as a smooth smear as per the shown example. **(B)** A library of the desired size is excised, and an image is retained for records. Different lengths of library should be cut for different applications. For example: for a 1x75bp run cut library from 200-300bp; for a 1x150bp run cut between 400-600bp.

5. Weigh excised gel and mix 3:1 volume for weight Zymo Agarose Dissolving Buffer (ADB) (e.g. 180µL ADB for 60mg agarose)

6. Incubate at 50°C for approximately 10mins. Make sure that the agarose has entirely dissolved before proceeding. Take care not to incubate at temperatures greater than 50°C, as this may partially melt some dsDNA fragments and result in improper quantification.

7. Purify the PCR product with the Zymo DNA clean protocol:

    a. Apply melted agarose in ADB to silica column, and centrifuge for 30-60s at 14,000 RPM, as per the manufacturer's protocol.

    b. Wash with 200µL ethanol-containing wash buffer and centrifuge for 30-60s at 14,000 RPM as per the manufacturer's protocol. Repeat for two washes.

    c. Elute by centrifugation for 60s at 14,000 RPM into fresh non-stick Eppendorf tubes using 6-10µL 10mM Tris pH 7.4 or water.

8. Quantify yield of final size selected cDNA library using a QuBit fluorimeter as per the manufacturer's protocol.

3.7 Sequencing and ClickSeq-specific data preprocessing

ClickSeq libraries can be submitted for either paired-end or single-end sequencing on Illumina platforms using the adaptor sequences described here. The first read is obtained from the Illumina universal primer end (p5) end of the cDNA fragment which is the location of the triazole ring in the original cDNA. The second read starts from the indexing (p7) adaptor, which is the site of the random priming in the RT-PCR. During data preprocessing, I recommend trimming the first 6 nucleotides from the beginning of both the forward and reverse reads, which correspond to the random 'N' nucleotides included in the sequencing adaptors (see **Materials 2.5**). Additionally, in the forward read, I have found that there is sequence bias in the 4th to 6th nucleotides for the forward read. These nucleotides correspond to those flanking the unnatural triazole linkage. In particular, position 5 can be occupied with an 'A' in up to 80% of the sequence reads. This position corresponds to the base complementary to the terminating azido-nucleotide introduced during RT-PCR, suggesting that either AzTTP is inserted more readily than the other azido-nucleotides during reverse transcription or that the click-ligation reaction favors terminal azido-thymine in the cDNA. Alternatively, it is possible that the PCR amplification step may preferentially insert an 'A' opposite the triazole-linkage regardless of the complementary base. I have not found this to adversely affect the evenness of our sequence coverage, however future optimization may be required to eliminate any potential bias.

NOTES

1) Template specific primers can be used in place of semi-random primers at this step. Simply exchange the 'NNNNNN' nucleotides for the sequence of choice. Proceed to 50°C without

initial 25°C incubation immediately after addition of reverse-transcription enzyme to reduce off-target amplification (**Methods 3.1.5**).

2) I have successfully generated RNAseq libraries from as little as 20pg of starting Flock House virus RNA. However, the number of PCR cycles used for final library amplification must be greatly increased (up to 36 cycles), which will inevitably introduce sequence bias and duplication.

3) Optimal AzNTP:dNTPs ratios must be determined empirically for a given procedure; but as a general rule, 1:20 is suitable for ~100-200nt inserts (e.g. 1x100bp SE Illumina); and 1:35 for >250nt inserts (e.g. 2x300 PE Illumina).

4) Care must be taken when aiming to make libraries with long insert lengths and thus with large ratio of dNTPs to AzNTPs. Smaller RNA fragments will allow the reverse transcriptase to reach the end of the RNA fragment without the incorporation of an AzNTP, resulting in an un-clickable product. As a result, these fragments will be strongly under-represented in the final cDNA library.

5) At this stage, a master mix can be made. For example, if making five libraries, mix 22µL 5X Superscript First Strand Buffer, 5.5µL 0.1M DTT 5.5µL RNase OUT 5.5µL Superscript III Reverse Transcriptase, and then add 7µL to each RNA/primer/dNTP mixture from **Method 3.1.2**. Superscript II and III Reverse Transcriptases seem to be stable during this short high-salt incubation.

6) Do not elute in any manufacturer provided elution buffer which contains Tris or in a buffer that contains EDTA. Amine-rich buffers such as Tris solutions may reduce the efficiency/yield of the click-ligation reaction[224] and EDTA will chelate the copper ions required for click reaction catalysis. I have found that HEPES elutes DNA well. Other buffers that are slightly

alkaline (to release DNA from the silica matrix) and that are compatible with the click-ligation reaction may also be suitable (e.g. potassium phosphate).

7) To determine the optimal concentration of DMSO during the click-reaction, side-by-side comparisons of libraries were made using the same input azido-terminated cDNA click-ligated in the presence of 10%, 20%, 30%, 40%, and 50% DMSO. The greatest output of final library was achieved in the presence of 50% DMSO (**Figure 4.5**).

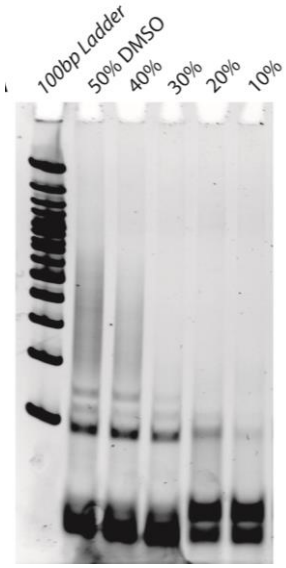


Figure 4.5: Optimizations of the ClickSeq protocol: DMSO concentrations.

Demonstrated by gel electrophoresis. Increasing concentrations of DMSO in the Click-Ligation reaction improve the final yield of the amplified cDNA library. (Work performed by A. Routh)

8) With such small volumes, the ascorbic acid reducing agent in the click-ligation reaction (whose role it is to maintain the copper catalyst in its required +1 oxidation state) is highly vulnerable to oxidation by atmospheric oxygen. Therefore, avoid introducing bubbles during pipetting and keep Eppendorf tubes closed as often as is possible. Smaller (e.g. PCR) tubes are similarly preferable.

9) The click-ligation reaction can be performed successfully at temperatures up to 90°C. However, caution must be taken using high-temperature or extended incubation due to the possibility of copper-mediated oxidative damage to the cDNA. This may result in an increased error-rate in base calling[225]. Moreover, I have found that extended or heightened incubation temperatures do not improve yield.

10) The click-ligation reaction can be re-supplemented with fresh catalyst/accelerant solution to ensure maximal click-ligation at regular intervals. I routinely make two total additions at 0mins and 30mins; proceeding to the cleaning step after 60mins. No substantial improvements in yield were found with three or more (**Figure 4.6**).

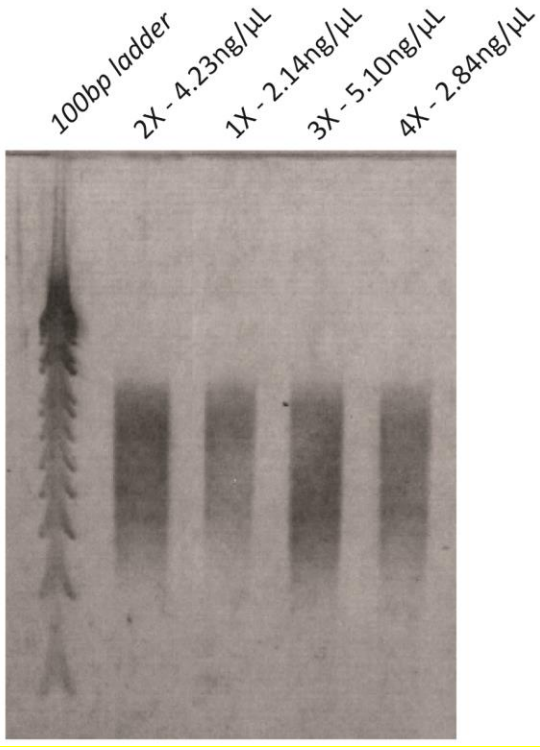


Figure 4.6: Optimizations of the ClickSeq protocol: Cu(I) Additions.

Demonstrated by gel electrophoresis. The optimal number of copper(I) additions at the click ligation step was tested. Two rounds of Cu(I) addition to the click-ligation reaction yielded to be most optimal. Final library yields are indicated above each lane.

11) Performing the click-reaction on the Zymo silica column itself resulted in a reduce yield, but nonetheless was feasible and results in reduction of work-flow. To 'Click-on-column': instead of eluting the azido-terminated cDNA at **Step 3.2.4**, make a mixture containing the click-ligation components as detailed in **Step 3.3.1** and **3.3.2**, except replace the azido-terminated cDNA with HEPES pH 7.2 buffer. Add 10µL of this mixture to the column without spinning and leave at room temperature for up to 1 hour. Then, add 280µL of the Zymo DNA binding buffer and incubate again at room temperature for 15mins to ensure that the cDNA remains bound to the silica matrix. Next, wash the column 2 times in wash buffer as per the standard procedure and elute the click-ligated DNA in 10mM Tris pH 7.4. Proceed directly to the final PCR amplification **Step 3.5**.

12) In the original 'ClickSeq' publication[157], the click-ligated cDNA products were not purified away from the components of the click-reaction. While this made for a simpler flow-through, it results in having to perform the final PCR reaction in a very large volume (200µL) in order to dilute away the large amounts of DMSO to acceptable levels. Additionally, without cleaning, the catalytic copper ions from the click-ligation would remain the PCR reaction mixture and may induce DNA damage due to the high cycling temperatures used during PCR[225]. Therefore, I prefer to purify the click-ligated cDNA. This may be achieved in a number of ways (e.g. SPRI beads, Zymo DNA clean columns, EtOH precipitation, etc.).

13) I often find it useful to amplify only half of the total purified click-ligated DNA so that a second library can later be made with fewer or more PCR cycles in case the yield of the final library is found to be inadequate or over-amplified.

14) Recently, researchers have reported that up to an 80% read-through of triazole-linked DNA templates can be achieved using non-thermostable Klenow polymerases with very long

incubations[226]. To determine whether longer PCR cycling conditions would improve final library yield, I performed the PCR cycles as described, but either with a 1min extension time in the initial and all subsequent cycles, or 2min, 5min or 10min extension time in the initial cycle followed by 2min extension times in all subsequent cycles (**Figure 4.7A**). The longer extension time in the first cycle improved yield by ~2 fold. I additionally screened for the optimal annealing temperature (**Figure 4.7B**). An annealing temperature of 54° was selected based off the calculated primer melting temperatures (NEB Tm calculator) and the condition that provided the strongest library intensity.

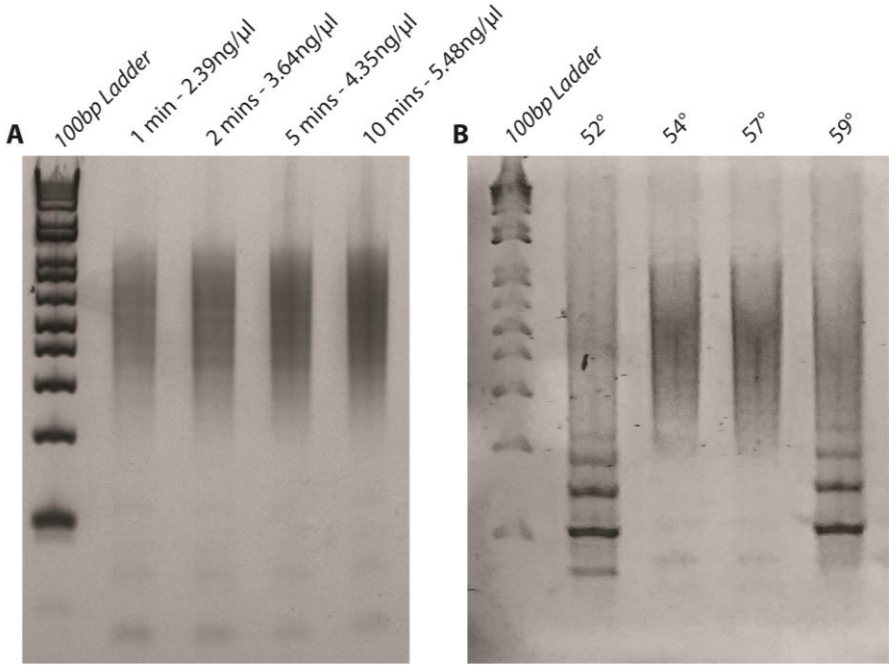


Figure 4.7: Optimizations of the ClickSeq protocol.

Demonstrated by gel electrophoresis. **(A)** Increasing the extension time during the first cycle of the PCR amplification improves yield. Final library yields after the PCR amplification are indicated above each lane. **(B)** PCR primer anneal temperatures were screened. A Tm of 54° was found to be optimal.

15) Reaction products can also be cleaned at any stage using a variety of methods. I prefer SPRI beads due to their simplicity, high throughput capabilities, and cost effectiveness. I have also

found that the Zymo DNA clean protocol works well due to its ability to elute in small volumes with minimal carry-over (Zymo Research DNA Clean and Concentrator-5 D4013):

1. **Step 3.2**: take 21µL RT-PCR reaction, and add 140µL Zymo DNA binding buffer (7:1 binding buffer:DNA). **Step 3.4**: the click-ligation reaction is first diluted with 60µL water to a total volume of 100µL prior to addition of the DNA binding buffer in order to dilute the DMSO then 700µL Zymo DNA binding buffer is added (7:1). **Step 3.5.3**: take the 50µL PCR reaction and add 250µL Zymo DNA binding buffer (5:1).
2. Apply to silica column, and centrifuge for 30-60s at 14'000 RPM, as per the manufacturer's protocol.
3. Wash with 200µL ethanol-containing wash buffer and centrifuge for 30-60s at 14'000 RPM as per the manufacturer's protocol. Repeat for two washes.
4. Elute by centrifugation for 60s at 14'000 RPM into fresh non-stick 1.5mL microcentrifuge tubes using:
    - **Step 3.2**: 10µL 50mM HEPES pH 7.2 or water (see note 4.6).
    - **Step 3.4**: 20µL 10mM Tris pH 7.4 or water.
    - **Step 3.5.3**: 20µL 20mM Tris pH 7.4 or water.

16) SPRI magnetic beads (such as AMPure XP, Beckman Coulter) give users an efficient and high throughput way to clean up reaction products throughout this protocol. Beads can homemade which will drastically reduce costs associated with the cleanup steps. DeAngelis et al. originally described the method for making homemade SPRI beads but I follow Faircloth and Glenn's 'Serapure' protocol (sourced from Rohland and Reich)[227-229].

17) SPRI magnetic beads can be used in place of gel electrophoresis cDNA fragment size selection (in lieu of the PCR purification **Step 3.5.3**). I have found that this method does not provide as well-defined fragment boundaries but has the advantage of allowing the user to process many samples at once. The 'SPRIselect User Guide' provided by Beckman Coulter (document #B24965AA) is very informative. I highly recommend that the user gets acquainted with the protocol and, if using homemade SPRI beads, test the precipitative qualities of their stock to adjust volumes accordingly. **Figure 4.8** shows an example of how SPRI bead size selection effects the same NGS library. Small changes in SPRI bead volume will change the fragment size that is precipitated. For a 1x150bp SE run on an Illumina HiSeq we select fragments ~400-600bp in length. Therefore, using this batch of SPRI beads I would follow the dual size selection protocol with a left side ratio of 0.7x and a right side ratio of 0.5x (**Figure 4.8**, Lane 4).

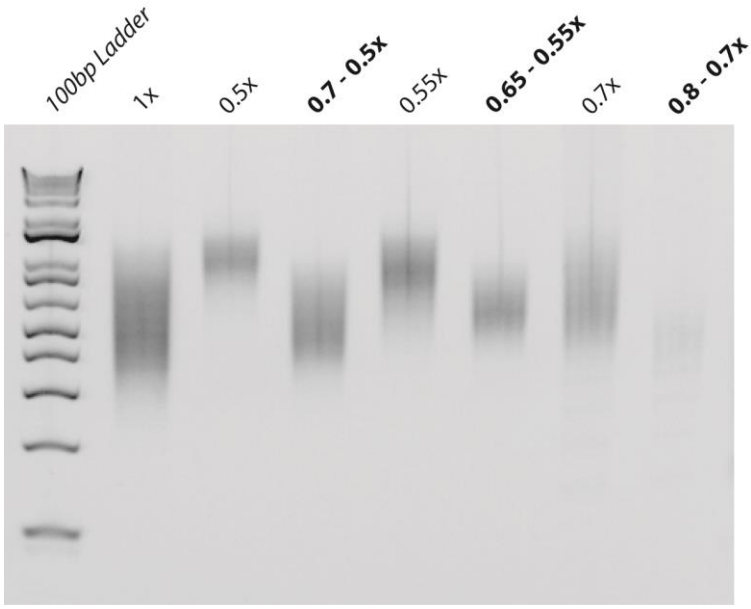


Figure 4.8: SPRI bead fragment size selection.

Demonstrated by gel electrophoresis. Left side selections are shown in lanes 2, 3, 5, and 7 (normal font). A larger ratio of SPRI beads increases the efficiency of binding smaller fragments (thereby precipitating shorter fragments). Double size selection can be performed (indicated by bolded lanes 4, 6, and 8) if a particular fragment range is desired.

18) I have tested the sensitivity of the ClickSeq protocol to pick up and identify genetic mutations and variations within a sample. Here, I generated and isolated RNA from a homogenous population of wild type Flock house virus and a FHV mutant, RNA2- A226G. I then mixed the quantified RNAs at a ratios of 100:0, 99:1, 90:10, 50:50, 10:90, 1:99, and 0:100, WT-to-mutant. Samples were then processed for sequencing using the standard ClickSeq protocol and sequenced on an Illumina HiSeq (1x150 SE). **Table 4.1** shows the theoretical and calculated error rate of the 226 nucleotide position based on the mixed ratios. Each sample had over 230,000X coverage at nt 226 of RNA2. Small errors can be accounted for due to the inherent error rate of the Illumina platform (<0.1%), pipetting errors that could have been introduced during the initial pooling step, or natural mutations introduced by the virus. To determine the reproducibility of ClickSeq and to determine the cut-off for sensitivity of discovery replicate ClickSeq libraries were generated using the same sample RNA, from three independent FHV samples. Recombination events were identified using ViReMa (as described in **Chapter 2**) and the comparison between the replicate was plotted and shown in **Figure 4.9**. Pearson correlation coefficients are indicated for the three independent samples. Further

WT:MT	Theoretical	Calculated	Nt Coverage
100:0	0%	0.2%	304,824
99:1	1%	1.2%	476,807
90:10	10%	11.2%	344,319
50:50	50%	53.0%	463,477
10:90	90%	91.7%	377,518
1:99	99%	99.5%	231,566
0:100	100%	99.9%	344,870

Table 4.1: Sensitivity of ClickSeq in identifying nucleotide errors.

RNA extracted from a wild type and mutant (RNA2-A226G) FHV populations were mixed at the specified ratios. The theoretical percentage that nucleotide position 226 of RNA2 should be a G is governed by the mixed ratio. Samples were sequenced and frequency was calculated (number of mapped 'G's / total nucleotide coverage at position 226). Nucleotide coverage at position 226 for each sample is indicated.

reproducibility of ClickSeq to identify the same recombination events between experimental triplicates is shown in **Chapter 2**, **Figure 2.4**.

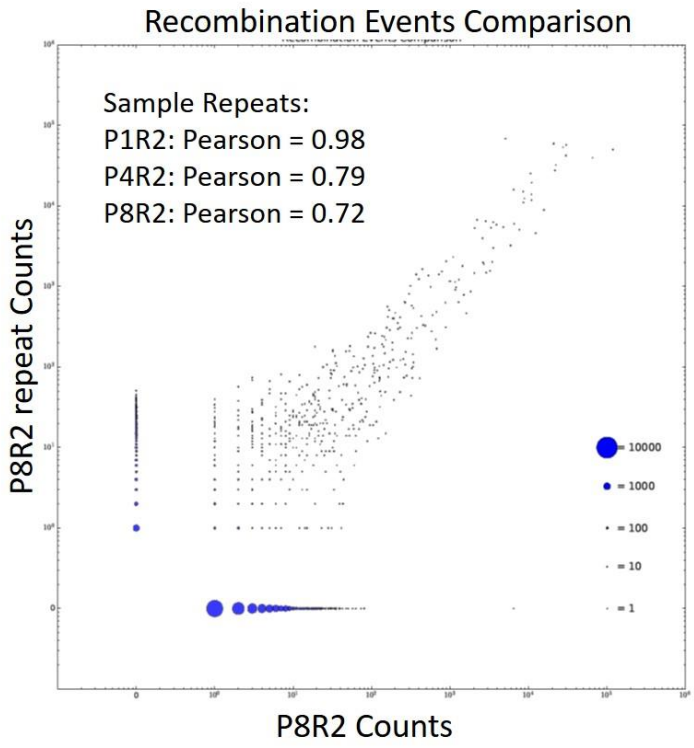


Figure 4.9: Reproducibility of ClickSeq to identifying FHV recombination events within the same sample.

RNA extracted from the same experimental samples were used to generate ClickSeq libraries in replicate. Recombination events were identified using the methods described in **Chapter 2.** Comparison of the identified recombination events are plotted for sample P8R2. Pearson correlation coefficients from three independent FHV samples are indicated.

POLY(A)-CLICKSEQ[§]

A click-chemistry based method for next generation 3'-end sequencing without RNA enrichment or fragmentation.

OVERVIEW

Poly(A) tails, with a few exceptions, are ubiquitous to all eukaryotic mRNAs and have important functions in localization signalization, translation, and stability (reviewed by Proudfoot[230]). Interestingly, they can also be found at the 3' end of many RNA viruses such as picornaviruses[231], influenza virus[232], and HIV[233]. During elongation, cellular mRNAs get poly(A) tails as part of pre-mRNA processing. Here transcripts are cleaved co-transcriptionally and subsequently tails are added on by the poly(A) polymerase. Cleavage of the 3' ends of pre-mRNAs is driven by three sequence elements: 1) a polyadenylation signal (PAS), with a typical sequence motif of AWUAA[234], 2) a cleavage site, which is typically a CA dinucleotide[235], and 3) a downstream sequence element (DSE), which is typically U/UG rich[236]. While it is generally accepted that these three elements help dictate the efficiency of polyadenylation, many groups have shown that the polyadenylation process is actually quite dynamic. This process is called alternative polyadenylation (APA) where the pre-mRNA processing machinery can generate distinct 3' ends of mRNA resulting in isoforms of varying lengths (reviewed by Tian and Manley[237]). APA results increasing the diversity of the transcriptome, can affect the stability of that mRNA, and is used by the cell as a mechanism to control gene expression (reviewed by Di Giammartino *et al*.[238]). Furthermore, studies have shown that APA is regulated in different tissue types[239], during

[§] This section is partially adapted from: Routh, A., Ji, P., Jaworski, E., Xia, Z., Li, W., and Wagner, E.J. (2017). Poly(A)-ClickSeq: click-chemistry for next-generation 3′-end sequencing without RNA enrichment or fragmentation. Nucleic Acids Res 45, e112. (Nucleic Acids Research is an open access journal applies the Creative Common Non-Commercial license)

development[240], as well as under disease/stress conditions[241]. Therefore, profiling the position of the polyadenylation site is critical to understand a wide range of biological studies.

As a result, a number of strategies have been developed with the specific goal of enriching for the junction of the encoded 3′UTR ends and the beginning of the non-templated poly(A) tail (reviewed by Szkop and Nobeli[242]). Common themes found in several of these techniques are the enrichment for poly(A)+ RNA from total RNA, fragmentation of mRNA using a variety of approaches (e.g. enzymatic, heat, sonication), and attachment of an adaptor to the 3′ end either through the use of a splinted oligo or directly to the terminus of the poly(A) tail. These initial steps can also involve the use of a biotin-containing oligonucleotide to allow for purification of the desired library intermediates using streptavidin magnetic beads. These approaches typically utilize between 1M and 20M reads and have the advantage of allowing precise mapping of the position of the poly(A) tail addition. However, these approaches often entail complex experimental pipelines and purification strategies that can impart sample bias and reduce throughput capacity. Importantly, these challenges can reduce the number of core facilities offering these types of sequencing technologies thereby limiting their application only to laboratories with more than routine experience in sequencing library preparation.

Here, in collaboration with Dr. Eric Wagner's group, I present a novel approach that provides a number of advantages over other methodologies due to its simplicity, cost-effectiveness and speed while providing high-quality, unbiased sequencing libraries. This approach is a subtle alteration of the ClickSeq technique where we sought to utilize this technique to target sequencing to only the 3′ ends of polyadenylated RNAs: 'Poly(A)-ClickSeq'; or PAC-seq. For PAC-seq, rather than using a random primer, is initiated using oligo (dT) primers without a non-T anchor during reverse transcription. This primer also contains an overhang corresponding to a portion of the Illumina p7 adaptor (illustrated in **Figure 4.10A**). By priming directly from poly(A)

tails, we can specifically reverse transcribe polyadenylated RNAs directly from crude RNA extracts without any prior sample purification or poly(A) enrichment. Moreover, by avoiding the use of a non-anchored oligo(dT) primer, in principle the primer can anneal anywhere with the poly(A) tail. Therefore, complementary cDNA transcripts will contain 'T's derived from the template as well as 21 'T's derived from the RT-primer.

In 'ClickSeq', cDNA synthesis can terminate opposite any nucleotide. In PAC-seq, however, the critical innovation required to specifically sequence the junctions of RNA 3′UTRs and their poly(A) tails is to omit AzTTP from the reaction mixture (i.e. we provide a mixture of AzVTPs and dNTPs). Without AzTTP present in the RT-PCR reaction mixture, reverse-transcription cannot terminate opposite an 'A' in the RNA template. Rather, reverse-transcription must continue until non-A residues are found (**Figure 4.7A**). Therefore, cDNA synthesis is stochastically terminated at a distance upstream of the 3′UTR/poly(A) junction tailored by adjusting the ratio of AzVTPs to dNTPs. This design allows for cDNA chain termination to occur only in the residues just upstream of poly(A) tail, essentially 'homing-in' on the junction of the 3′UTR and the poly(A) tail. We have found that a ratio of 1:5 AzVTPs:dNTPs reliably yields cDNA fragments ranging from 50 to 400nts in length.

To finalize PAC-Seq libraries, we purify the azido-terminated cDNA, 'click-ligate' the 5′ Illumina adaptor, and then PCR amplify an NGS library containing the desired demultiplexing indices (**Figure 4.10B**). The total size of all the adaptors including the oligo(dT) primer is 150bp. Therefore, cutting cDNA fragments 200–400nt in length will yield inserts 50–250nts in length (**Figure 4.10C**). Each of the cDNA fragments will therefore contain: the full Illumina p5 adaptor; cDNA corresponding to the 3′UTR of the RNA transcript, the length of which is determined by the stochastic termination of RT-PCR; the poly(A) tail; and finally the Illumina p7 indexing adaptor (**Figure 4.10D**). For optimal yield of reads containing poly(A) tails, libraries must be carefully size

selected depending upon the sequencing platform and length of reads sequenced. Sequencing is initiated from the p5 adaptor. Therefore, if fragments are too large and the cDNA insert is longer than the length of the sequencing read, the 3′UTR/poly(A) tail junction will not be reached. Ultimately, this protocol allows researchers to simply and efficiently: 1) identify APA site selection, 2) hone into the 3'UTR/poly(A) tail junction (to determine cleavage site with nucleotide precision), and 3) count transcript levels.

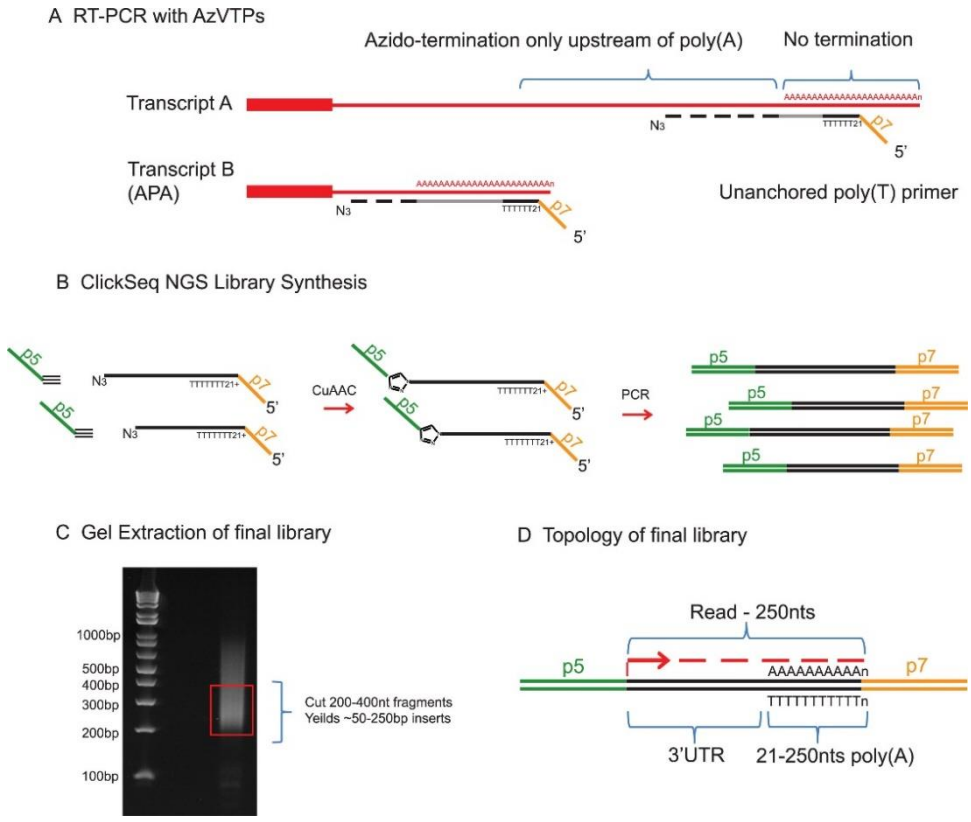


Figure 4.10: Schematic overview of Poly(A)ClickSeq (PAC-seq).

(A) RT-PCR is launched from a non-anchored Poly(T) primer containing a portion of the Illumina p7 adaptor. RT-PCR is performed in the presence of AzATP, AzGTP and AzCTP, but not AzTTP, thus only allowing chain termination to occur upstream of the poly(A) tail in the 3′UTR. **(B)** 3′-Azido-blocked cDNA fragments are 'click-ligated' to 5′-hexynyl–functionalized DNA oligos containing the p5 Illumina adaptor. This yields triazole-linked ssDNA which can be PCR-amplified using primers to the p5 and p7 Illumina adaptors. **(C)** The cDNA library is analyzed by gel electrophoresis and should consist of a smear of DNA products centered ∼200–300bp. Appropriate cDNA fragment sizes are cut out of the gel and purified to yield a final library. **(D)** The final library consists of DNA fragments containing the Illumina p5 adaptor, a portion of the 3′UTR, a stretch of As derived from both the RNA template and the poly(T) primer, and finally the p7 Illumina Indexing primer.

PROTOCOL

Here I describe the PAC-seq protocol. This is highly similar to the highly optimized ClickSeq protocol described above. Therefore, to avoid redundancies I only describe the specific steps that are different to those of ClickSeq. All other steps are identical.

MATERIALS

2.1 Reverse Transcription Components:

- During reverse transcription, a ratio of AzVTP:dNTPs is used (instead of AzNTPs)(see **Methods 3.1.2** below for recipe)

2.5 Primers and Oligos:

Primer Name	Sequence	Stock Solution	Working Solution
3' Illumina_4N_21T (partial p7 Adaptor)	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**NNNN**TTTTTTTTTTTTTTTTTTTTT	100µM in Water	100µM in Water

METHODS

3.1 Reverse Transcription

1. Input RNA: For poly(A)seq, we usually aim to provide 4µg of RNA total cell. No sample fragmentation is required. No sample purification/rRNA depletion/selection is required. Total crude extract can also be used to generate RNAseq libraries. No extraction methods are required as little as 10^4 cells can be used.
2. For a 1:5 5mM AzVTP:dNTP solution mix the following:
    a. 10µl 10mM dNTPs
    b. 2µl 10mM AzATP
    c. 2µl 10mM AzCTP

d. 2µl 10mM AzGTP

e. 4µl H_2O (NOTE: do not add AzTTP)

3. The reverse transcription is performed using standard protocols, with the exception that the reaction is supplemented with small amounts of azido-nucleotides (AzVTPs) and a specific poly(T) primer. Set up RT-PCR reaction as follows for a 13µl reaction:

    a. 2µl 5mM 1:5 AzVTP:dNTP mixture (see **note 4.19**)

    b. 1µl 3' Illumina_4N_21T primer at 100µM

    c. 125ng-4µg RNA (see **Note 4.20**)

    d. H_2O to a final volume of 13µl

4. Incubate mixture at 65°C for 5 mins to melt RNA and immediately cool on ice for > 1 min to anneal poly(T) primer.

5. Add the following at room temperature for a final reaction volume of 20µL:

6. 4µL 5X Superscript First Strand Buffer

    a. 1µL 0.1M DTT

    b. 1µL RNase OUT Recombinant Ribonuclease Inhibitor

    c. 1µL Superscript III Reverse Transcriptase

7. Incubate with the following steps:

    a. 50°C for 40 mins,

    b. 75°C for 15 mins, and

    c. Hold at 4°C

8. To remove template RNA, add 2U RNase H and incubate at 37°C for 20 mins, 80°C for 10mins, and then hold at 4°C.

3.2-3.5 As per the standard protocol

3.6 Fragment size selection

Any method can be used as previous described. For PAC-seq libraries I recommend using gel electrophoresis size selection as extracting the appropriate cDNA fragment size is critical for this method. Gel size excision should be 200-300bp for a 1x150 Illumina run or 200-400bp for a 1x300 Illumina run.

3.7 Sequencing and Poly(A)-ClickSeq specific data preprocessing

It is recommended that PAC-Seq libraries be submitted for single-end sequencing on Illumina platforms using the adapter sequences described here. The first read is obtained from the Illumina universal primer end (p5) of the cDNA fragment (**Figure 4.10D**). This will read through the cDNA fragment (3'UTR of a transcript) followed by the poly(A) track. Paired-end sequencing is not recommended as the second (paired) read starts from the p7 indexing adapter which will begin reading through the pol(A) tail. We have found that the high abundance of As during the initial rounds of sequencing results in a failed run due to the Illumina platform requiring diversity on the flowcell. During data processing, the first 6 nucleotides (that were derived from the 'Click Adapter', as per the standard ClickSeq protocol) should be trimmed. Reads that are shorter than 60nts should be discarded as they are too short to contain both a poly(A) tail (when requiring poly(A) tails to be >21nts) and enough nucleotides to provide unambiguous mapping. Additionally, reads that have a have poly(A) tails less than 15nts in length and total read length less than 40nts are filtered out. The As are trimmed off the reads. This is to ensure that the sequenced A track is a true poly(A) tail (as the 21T primer used during reverse transcription can

bind non-specifically or partially prime from shorter internal A sequences). The remaining reads are then mapped to the host using a standard mapping program, we prefer to use HiSat2[243]. Lastly, the 3' most nucleotide position is called as the poly(A) site. To read a more in-depth protocol for data processing of Poly(A)-ClickSeq data for differential gene expression and poly(A) site selection analysis see Elrod and Jaworski, *et al.* [244].

NOTES

19) An optimal AzVTP:dNTP ratio should be determined for a given procedure. We have found that a ratio of 1:5 is suitable for cDNA inserts of 50-400bp in length (e.g. 1x150 SE Illumina). Other ratios such as 1:3 or 1:4 can be used for longer inserts (**Figure 4.11**).

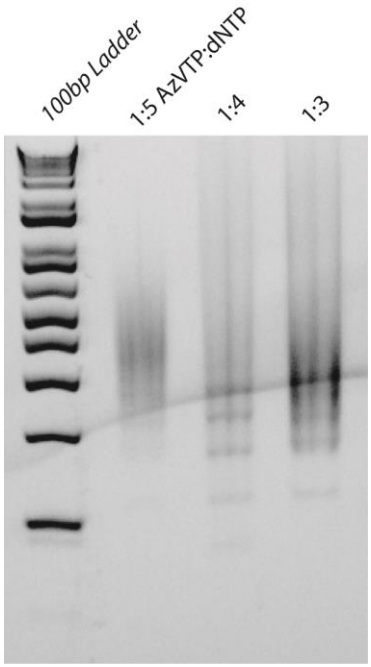


Figure 4.11: PAC-seq Optimizations.

Demonstrated by gel electrophoresis. The ratio of AzVTPs to dNTPs during the RT-PCR reaction will affect the distribution of cDNA fragment length. Decreasing the ratio of dNTPs will result in longer fragments.

20) In order to determine the input sensitivity range of PAC-seq we have tested a range of input RNA. Libraries were generated following the standard PAC-seq protocol using RNA isolated from HeLa cells ranging from 2µg to 125ng. Samples were sequenced and the sensitivity to identify poly(A) sites between samples was calculated. Overall, a pairwise comparison between samples indicated Pearson correlation coefficients ranging between 0.92 and 0.98. Therefore, we have found that the quality of data is not altered by the amount of input RNA and as little as 125ng of total cellular RNA will provide adequate results. (work performed by: P. Ji and explained in detail in Routh *et al.*[245])

METHOD APPLICATION

For proof of principle and a demonstration of application see Routh *et al.*[245]. Here, PAC-seq was used to identify the poly(A) sites of RNA extracted from both human (HeLa) and Drosophila (S2) cells. In these models PAC-seq was able to accurately identify the locations of previously annotated polyadenylation sites. Furthermore, using HeLa cells deficient of the pre-mRNA cleavage factor, CFIm25, the sensitivity of PAC-seq to analyze alternative polyadenylation sites was validated.

POLYMERASE PROFILING WITH CLICKSEQ

A click-chemistry based next generation sequencing technique for profiling nascent elongated RNA transcripts.

OVERVIEW

RNA polymerases (RNAP) are highly conserved enzymes that produce RNA from a genomic DNA template sequence in a process called transcription. They are highly regulated and play an vital role in gene transcription and regulation. Polymerase activity is modulated at individual genes, which is regulated by interactions of transcription factors with other regulatory factors (reviewed by Fuda *et al*.[246]). Transcription carried out by the RNA polymerase can be broken down into three phases: initiation, elongation, and termination. It was initially accepted that initiation was the major regulatory step and polymerase pausing only occurred at the promoter-proximal regions but it turns out that elongation is also a discontinuous process where pausing occurs globally and frequently[247]. While the exact mechanisms behind the pausing is not well known it is thought that this pausing provides opportunity for regulation and coordination with other processes such as mRNA maturation (ie. splicing), 3' end processing, and transcript export modulation, which occurs through contact with the elongation complex[248, 249]. Ultimately, the pausing during elongation plays a huge role in gene regulation during cell differentiation, proliferation, and under disease states[250].

In order to better understand the mechanisms that govern polymerase pausing and the role that pausing plays in transcriptional regulation, a handful of sequencing strategies have been developed. These techniques broadly fall into two categories: 1) nuclear 'run-on' sequencing, and 2) native transcript sequencing.

Some common methodologies for nuclear run-on sequencing include GRO-seq (global run-on-sequencing)[251], PRO-seq (precision nuclear run-on-sequencing)[252], BRIC-seq (BrU-immunoprecipitation chase-deep sequencing)[253], Bru-seq[254], BruDRB-seq[255], and 4sUDRB-seq[256] (BrU: 5'-bromo-uridine; DRB: 5,6-dichlorobenzimidazole 1-beta-D-ribofuranoside; 4sU: 4-thiouridine). These methods are able to identify genes actively transcribed by the polymerase and they broadly follow the same basic principle. First, nuclei from cells are extracted and isolated where polymerase activity is halted (ie. freezing, DRB). Extracts are then treated with 'run-on' components that restart the polymerase in the presence of labeled nucleotides (labeling varies per protocol, ie. BrU, biotin-NTP, 4sU) which can be incorporated into the elongating transcript. Finally, newly synthesized transcripts can be identified and extracted based upon the unnatural incorporated modification and sequencing libraries can be generated. These methods are able to map transcriptionally engaged RNA polymerases and determine the relative activity of transcription. Unfortunately, these processes are limited to cell culture studies and can produce many artifacts due to the many manipulation steps.

Native transcript sequencing includes methods like NET-seq (native elongating transcript sequencing)[257], short nuclear RNA sequencing[258], and 3'NT method (3' ends of native transcripts)[259]. NET-seq, the more common of the approaches flash freezes the RNA polymerase. Cells are then lysed and chromatin DNA is fragmented. This process then exploits the intrinsic stability of the RNA polymerase complex and uses immunoprecipitation to pull down the RNAP and its associated transcript. Subsequently, the RNAs are extracted, adapted sequences are attached to the newly synthesized fragments, and then further processed for sequencing. This technique provides nucleotide resolution of the elongating transcript but can be limited to systems that have available RNAP antibodies. Overall, these approaches are extremely complex and can time consuming, with some protocols taking 4-5 days[260].

Here, I present my approach to studying polymerase activity during transcription that provides some advantages over previous methods due to its simplicity and cost-effectiveness. This approach is a variation of our ClickSeq technique where I sought to identify, with nucleotide resolution, the exact location of elongating polymerases with a method I call 'Polymerase Profiling-ClickSeq'; or 2PC-seq. For 2PC-seq, sample processing is a critical component; cells are flash frozen using liquid nitrogen to instantly halt transcription and total cellular RNA is extracted using standard methods (ie. an acid-guanidinium-phenol type reagent) (**Figure 4.12A**). Ribosomal RNA is then depleted from the total cellular RNA pool. Subsequently, a single stranded RNA ligase (T4 RNA ligase) ligates a miRNA cloning linker onto the 3' ends of nascent RNAs. Rather than using a random primer (as with standard ClickSeq), for 2PC-seq, reverse transcription is initiated using a primer that is the reverse compliment of the miRNA cloning linker. Reverse transcription is stochastically terminated as the reaction is spiked with azido terminated NTPs (AzNTPs). This primer also has the partial Illumina p7 adapter sequence (**Figure 4.12B**). The standard ClickSeq protocol is then followed where the p5 Illumina adapter is 'clicked' onto purified cDNA fragments and PCR amplified to make final NGS libraries (**Figure 4.12C**).

The total size of the adapters and miRNA cloning linker is 147bp. Therefore, cutting cDNA fragments 200-400nt will yield inserts ~50-250nts in length. Final cDNA libraries will contain: the full Illumina p5 adapter; cDNA corresponding to the nascent elongated RNA; the miRNA cloning linker; and finally, the full Illumina p7 adapter (**Figure 4.12D**). Just as with the PAC-seq protocol, libraries must be carefully size selected depending on the length of reads sequenced. Sequencing is initiated from the p5 adapter so in order to determine the precise 3' end of the RNA transcript the sequencing read needs to reach the 3'end/miRNA cloning linker junction. Therefore, if the cDNA insert is too long the junction will not be reached.

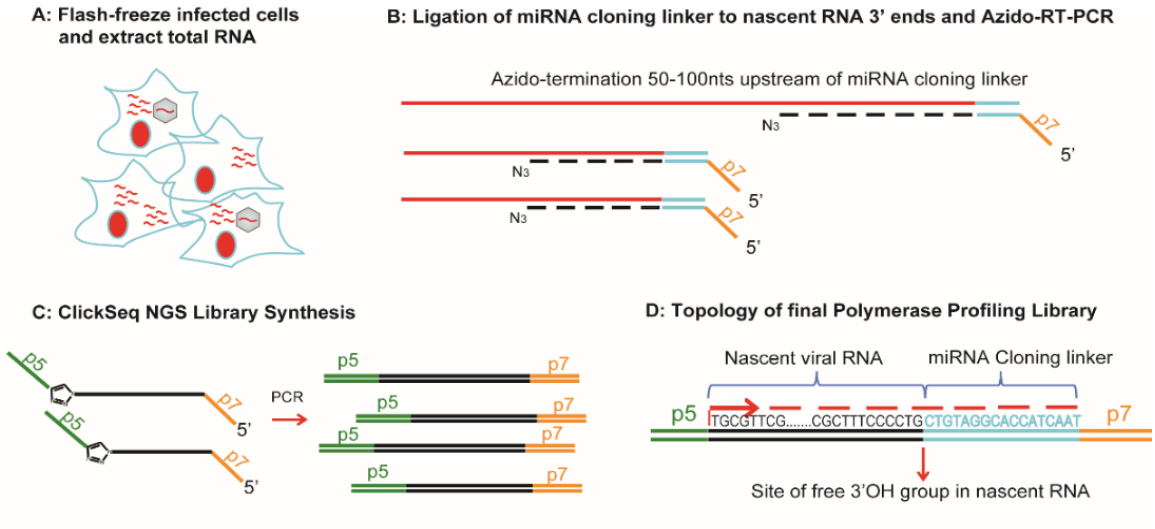


Figure 4.12: Schematic overview of 3' end sequencing.

(A) Infected cells are flash-frozen in LN_2 to trap polymerases during RNA replication. **(B)** Nascent RNAs have free 3'OH groups at the site of last replication. Without fragmentation, a miRNA cloning linker is ligated onto the 3'OH end and a ClickSeq library is made priming from this linker **(C).** Final NGS libraries have the structure depicted in **(D)**. The position of the 3'OH site is determined by the junction between mapped reads and the miRNA cloning linker

PROTOCOL

Here I describe the 2PC-seq protocol which is highly similar to the highly optimized ClickSeq protocol described above. A critical difference in this protocol is the sample handling steps prior to library preparation. Therefore, to avoid redundancies I only describe these specific steps and the ones that are different to those of ClickSeq. All other steps are identical.

MATERIALS

2.5 Primers and Oligos:

Primer Name	Sequence	Stock Solution	Working Solution
3' univ_miRNA_Illumina (partial p7 Adaptor)[1]	GTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT**ATTGATGGTGCCTACAG**	100µM in Water	100µM in Water

- [1]The bolded sequence is the reverse compliment of the miRNA cloning linker

2.6 Additional Components:

- TRIzol Reagent (Life Technologies), TRI-Reagent (Zymo Research), or any similar acid-guanidinium-phenol reagent
- Ribo-Zero rRNA Removal (Illumina)
- miRNA Cloning Linker (NEB, S1315S) (sequence: 5´ rAppCTGTAGGCACCATCAAT–NH_2 3´)
- T4 RNA Ligase I (ssRNA Ligase) (NEB, M0204), with standard 10X reaction buffer
- 50% PEG 8000 (NEB, B1004A) (comes with the T4 RNA Ligase module)
- Zymo Research Direct-zol (Zymo Research, R2051)
- Zymo Research RNA Clean & Concentrator -5 (Zymo Research, R1015)

METHODS

3.0 Sample Preparation and RNA Processing

1. To properly capture actively replicating polymerases, cells from live cultures need to be flash frozen to halt replication quickly. This can be accomplished by exposing the cells to liquid nitrogen (LN_2) which can be done in a variety of different ways. For suspension cells, take at least $1x10^5$ cells in media and pellet (10min at 1000 x g, or as described in the cell line's protocol). Once the cells are pelleted, remove supernatant and submerge the tube in LN_2. After the tube reaches temperature (LN_2 ceases bubbling) then remove tube and apply TRIzol reagent (volume based off the number of cells used). For adherent cells, remove media and dunk entire tissue culture flask in LN_2 until flask reaches temperature. (Attention: check that the tissue culture flask you are using is compatible with flash freezing). TRIzol reagent can then be applied directly to the flask and transferred to a tube for continued processing. Alternatively, adherent cells can be dislodged (either through mechanical scraping or with enzymatic methods such as Trypsin), transferred to a tube, pelleted, supernatant removed, and then flash frozen in LN_2. Personally, I do not recommend this method as the extensive manipulation of the cells can potentially disrupt polymerase activity, thereby skewing results.
2. Total cell RNA can then be extracted as desired. I prefer to use the Zymo Research Direct-zol RNA extraction Kit for its ease of use.
3. Depletion of ribosomal RNA is performed using the standard Ribo-Zero rRNA Removal Kit (Illumina, Document #15066012 v02).
    1) Clean Magnetic Beads
        i. Add 225µL Ribo-Zero beads per reaction and wash on magnetic stand twice with 225µL H_2O (RNase-free), vortexing to resuspend between washes

ii. Re-suspend beads in 65µL Magnetic Bead Resuspension Buffer

2) Hybridize probes to sample RNA

i. Mix for a final 40µL reaction:

a. 4µL Ribo-Zero Reaction Buffer

b. 10µL Ribo-Zero Removal Solution

c. >2.5-5µg RNA

d. H_2O to a final volume of 40µL

ii. Incubate with the following steps:

a. 68°C for 10min

b. Room temperature for 5min

3) Remove rRNA

i. Combine 65µL of cleaned Magnetic Beads (from step 3.0.3.1) to 40µL of the hybridized RNA (from step 3.0.3.2).

ii. Incubate with the following steps:

a. Room temperature for 5min

b. 50°C for 5min

iii. Immediately place on magnetic stand and transfer supernatant to a new tube

4. Clean up RNA. This can be done a number of ways; I prefer to use the Zymo Research RNA Clean & Concentrator-5 following the standard protocol, eluting in 15µL H_2O.

5. After removal of ribosomal RNAs, the sample should contain full length RNAs as well as our target RNA. This RNA, since it was terminated during replication, should have a free hydroxyl group on the 3' end of the RNA fragment (3'-OH). Here I take advantage of an established adapter ligation protocol in order to target these molecules based off the known adapter sequence[261].

    1) Mix the reaction as follows for a 22µL reaction, then incubate overnight at 18°:

        a. 250ng-1µg RNA

        b. 0.5µL miRNA Cloning Linker (NEB) (at 0.5µg/µL)

        c. 2µL 10X T4 RNA Ligation I Buffer

        d. 0.5µL T4 RNA Ligase I

        e. 8µL PEG 8000 (50%)

        f. H_2O to a final volume of 22µL

6. Clean up RNA using a protocol that allows for RNA size selection (required to remove small fragments (e.g. tRNAs, miRNAs). I prefer to use the Zymo Research RNA Clean & Concentration-5 following the 'Purification of small and large RNAs into separate fractions' portion of the supplied protocol.

    1) Add 78µL H_2O to sample (to dilute down the PEG8000).

    2) Adjust 100µL RNA Binding Buffer with 100µL 100% ethanol (50:50 binding buffer:EtOH).

    3) Add 200µL of adjusted buffer to 100µL of the RNA sample.

4) Apply to silica column, and centrifuge for 30-60s at 14,000 RPM (The flow through contains 17-200nts long RNAs, which includes unligated miRNA linker and small host RNAs).

5) Wash with 400µL of RNA Prep Buffer and centrifuge for 30-60s at 14,000 RPM.

6) Wash with twice with RNA Wash Buffer, and centrifuge, a volume of 700µL for the first wash and 400µL for the second wash.

7) Elute by centrifugation for 60s at 14,000 RPM into a fresh non-stick tube using 15µL H_2O.

3.1 Reverse Transcription

1. Input RNA: must be processed as described above (**Method 3.0**).
2. The reverse transcription is performed using standard protocols, with the exception that the reaction is supplemented with a specific RT primer (**Materials 2.5**). Set up RT-PCR reaction as follows for a 13µl reaction:
    a. 1µl 10mM 1:35 AzNTP:dNTP mixture.
    b. 1µl 3' univ_miRNA_Illumina primer at 100µM
    c. >100µg RNA
    d. H_2O to a final volume of 13µl
3. Incubate mixture at 65°C for 5 mins to melt RNA and immediately cool on ice for > 1 min to anneal the miRNA primer.
4. Add the following at room temperature for a final reaction volume of 20µL:
    a. 4µL 5X Superscript First Strand Buffer

b. 1µL 0.1M DTT

c. 1µL RNase OUT Recombinant Ribonuclease Inhibitor

d. 1µL Superscript III Reverse Transcriptase

5. Incubate with the following steps:

a. 50°C for 40 mins,

b. 75°C for 15 mins, and

c. Hold at 4°C

6. To remove template RNA, add 2U RNase H and incubate at 37°C for 20 mins, 80°C for 10mins, and then hold at 4°C.

3.2-3.5 As per the standard protocol

3.6 Gel extraction and size selection

Any method can be used as previously described but for 2PC-seq libraries I recommend using gel electrophoresis size selection as extracting the appropriate cDNA fragment size is critical for this method. Gel size excision should be 200-300bp for a 1x150 Illumina run.

3.7 Sequencing and Polymerase Profiling data preprocessing

As with PAC-seq it is recommended that 2PC-seq libraries be submitted for single-end sequencing on Illumina platforms. The first read is obtained from the Illumina universal primer end (p5) of the cDNA fragment (**Figure 4.12D**). This will read through the cDNA fragment (3' end of an elongating transcript) followed by the miRNA cloning linker sequence. For data processing, trim the miRNA linker sequence and discard any reads that do not contain that sequence. Reads that are shorter than 40nts should be discarded as they are too short to contain enough nucleotides to provide

unambiguous mapping. Furthermore, the first 6 nucleotides off of every read should be trimmed (as per the standard ClickSeq protocol; this is the region around the triazole linker where nucleotide biases are seen). Reads can then be mapped to the host genome using any standard mapping program; I prefer to use HiSat2[243]. Pileups are generated using SAMtools[169], the 3' ends of the reads are extracted using SAMtools and custom made python scripts, and the positions plotted as illustrated in **Figure 4.13**. These represent the 3' most nucleotide added by the polymerase during elongation.

METHOD APPLICATION

My motivation to develop this variation of the ClickSeq protocol was to study and identify viral RNA polymerase pause sites. As introduced in **Chapter 1**, almost all RNA viruses produce defective viral genomes and while the evidence for their existence is plentiful, the mechanisms by which they form is less supported. One of the more accepted models for their formation is polymerase driven recombination. This could be due to a template switching event (sequence homology re-priming during replication and/or forced copy choice when the polymerase reaches the end of a template and jumps to another) or due to some 'tough' spots that force the polymerase to jump from one part of the template to another (as with strong secondary RNA structures).

We and others have shown that Flock house virus produces and maintains a slightly variable 'species' of defective genomes during high multiplicity infections (as described in **Chapter 2**)[44, 54, 86, 157, 188]. This implies that there are some sort of selective pressures for these common, but slightly variable, sets of genomes to form. Their maintenance in a population is either a product of the mechanism by which they are formed, or is governed by other factors like particle packaging and stability (as explored in **Chapter 3**), or maybe even by both. In the case of FHV, it appears that I have evidence to support that both factors could be at play.

To better study the mechanisms by which DI-RNAs are formed I needed a way to profile what the RNA polymerase was doing. As mentioned previously, there are a handful of techniques that could possibly allow us to do this (i.e. GRO-seq, NET-seq, etc.). Unfortunately, all of them have their pitfalls, making the kinds of studies I wanted to accomplish on my model unideal. For example, the NET-seq protocol requires immunoprecipitation anti-bodies against the polymerase, which for some viruses (such as FHV) aren't available on the market. Or in the case of GRO/PRO-

seq that require nuclear isolations which entails many complicated and sensitive steps (FHV replicates in invaginations of the mitochondria). Furthermore, these protocols require that the polymerase of study incorporates unnatural NTPs (such as biotinylated-NTPs). These obstacles pushed me to develop a better system that would help me answer the questions I'm asking.

I tested the 2PC-seq method on FHV replication. Here I wanted to see if I could find any polymerase pause sites and, if those sites correlated with the recombination sites I see during DI-RNA formation. Supernatant from two different passages of FHV infections were used to infect S2 cells. Infection was allowed to proceed for 9 hours, after which, cells were spun down and flash frozen using liquid nitrogen following the suspension cell protocol described previously. Total cellular RNA was extracted using TRIzol reagent followed by the Zymo Research Direct-zol Kit. 2PC-seq libraries were made following the standard protocol, and 1x250 single-end reads were acquired on an Illumina MiSeq platform. Approximately, 1.5 million reads were generated per sample of which only about 10,000 reads mapped to the FHV genome (**Table 4.2**). The remaining reads mapped to the host genome. The 3' ends of mapped reads can be extracted and plotted indicating the free 3' ends of elongating transcripts of both FHV RNAs (**Figure 4.13A**). Directionality of the read can also be extracted from the mapped data indicating whether the polymerase was transcribing the positive sense RNA (shown in blue) or the negative sense RNA (shown in red). Interestingly, many of the pause sites correspond to the boundaries of important regulatory elements found on both RNAs. Features can be seen such as increased pausing around nucleotide 2720 on the negative strand of RNA1 corresponding to the 3' end of subgenomic RNA3. Increased pausing can also be seen next to the stop codon (nucleotide 3036) on the positive strand of RNA1. **Figure 4.13B** depicts approximations of the most common DI-RNA found for both FHV genomes as well the annotations of sequence elements. For RNA1 it appears that there is increased pausing on the (+)sense RNA at the boundaries of the deleted regions (as indicated at approximately

nucleotides 250 and 1150). While this is much more apparent for Passage 5 of RNA1, this could be attributed to the fact that there are increased relative levels of DI genomes in the P5 sample. Apparent pausing sites could be correlated to the deletions found in RNA2 such as the pause site around nt 250 of the (+)RNA.

Overall, while it appears that the 2PC-seq technique is able to consistently identify polymerase pausing sites during FHV replication (as shown by the consistent spectra between both passages) much more work still needs to be done. Firstly, many more controlled studies need to be conducted in order to more confidently attribute polymerase pause sites to the recombination I see in defective genome formation. Furthermore, I would also like to confirm that the free 3' ends that I am profiling are indeed true polymerase pausing sites as opposed to breaks in the genome as I could potentially be profiling degradation products since this method has no way for specifically selecting for replicating genomes. Validation plans include using polymerase perturbing antiviral drugs as well as comparing our technique to already established polymerase profiling methods.

	Passage 1	**Passage 5**
Raw Reads	1,480,558	1,654,854
Processed Reads	1,456,714	1,622,931
FHV Mapped Reads		
RNA1	6224	6530
RNA2	3113	2805

Table 4.2: Mapping of 2PC-seq reads.

Quantity of raw reads generated for each sample is tabulated. 'Processed reads' indicates reads remaining after filtering and trimming (following the criteria discussed in **Methods 3.7**). Reads that mapped to RNA1 or RNA2 using HiSat2 are shown. This count includes both positive and negative stranded RNA.

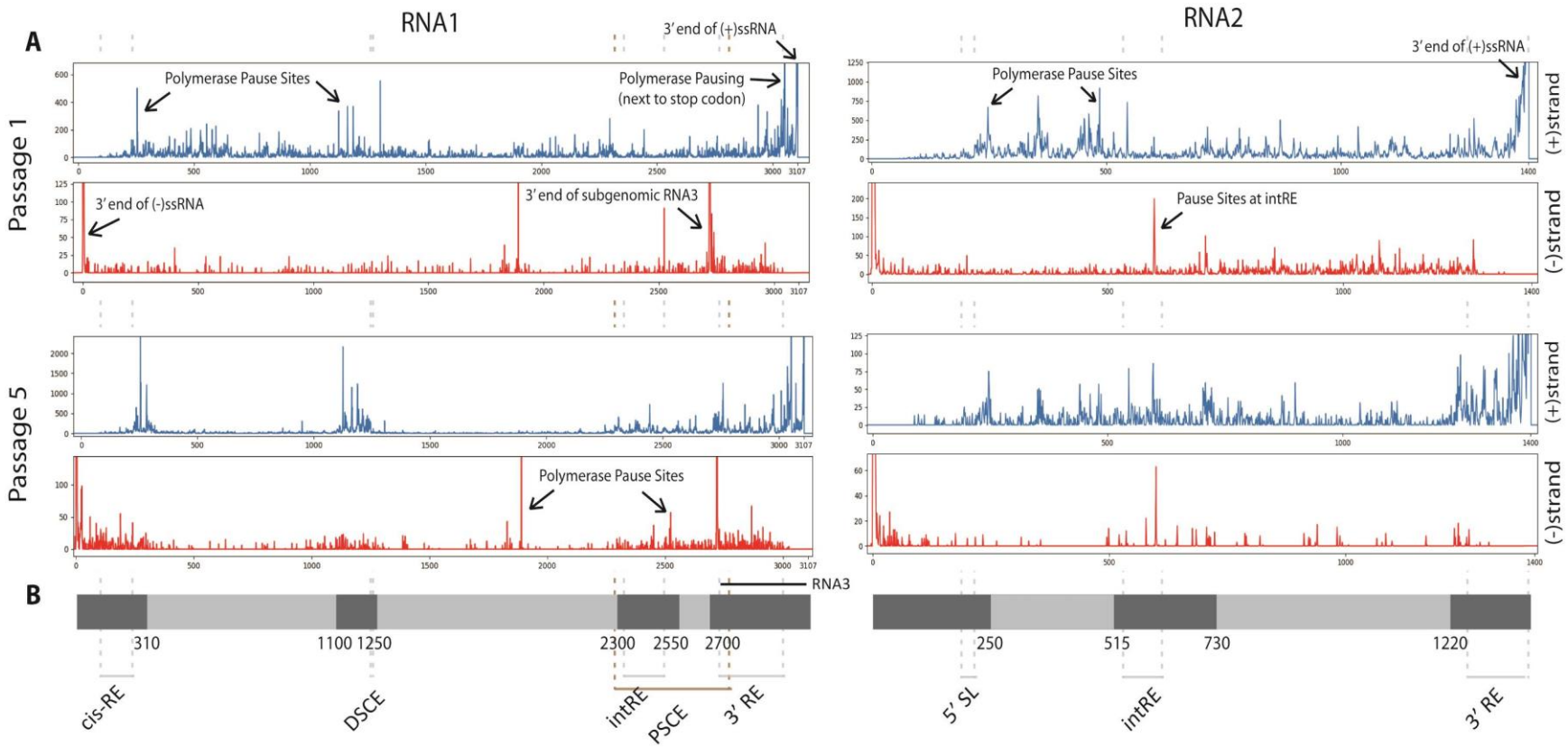


Figure 4.13: Mapped polymerase pausing sites in FHV infections.

(A) Polymerase pausing sites as indicated by frequency that a read mapped to a certain nucleotide position on the FHV genomes during infections infected with a passage 1 virus and a passage 5 virus. Positive strand mapping is shown in blue, negative strand mapping is shown in red. **(B)** Schematic of the most common deleted regions (light grey) in FHV. *Cis-RE:* cis-Regulatory Element; *DSCE/PSCE:* Distal/Proximal Subgenomic Control Elements; *intRE:* internal Response Element; *3′ RE*: 3′ Response Element; *5′ SL*: 5′ Stem Loop

CHAPTER 5: DISCUSSION

"According to the definition of a virus as simply a molecular genetic parasite, any genetic replicator, even noncellular prebiotic ones, would be susceptible to parasitic replicators or viruses. The tendency for replicators to become parasitized, and even for the parasitic replicators themselves to become parasitized, is a well-established phenomenon in virology. The parasites of parasitic replicators would correspond to the defective viruses that are observed for most types of viruses. Defective viruses are thus exactly the parasitic replicators of a functional virus, itself a parasitic replicator."

- Luis P. Villarreal, Viruses and the Evolution of Life

In the beginning of this manuscript, I quoted a poem by Augustus De Morgan; "Big fleas have little fleas…/ and little fleas having lesser fleas…". While originally this was De Morgan's philosophical response to the infinite essence of the universe, the poem has merit in biological terms as we discover parasites of parasites in nature. Generally, we think of viruses as parasites, as so by this merit, they too are susceptible to their own parasites, or defective interfering particles.

Defective interfering viruses are particles that contain broken genomes (DI-RNAs) and rely on the wild-type virus to propagate, effectively parasitizing the normal viral machinery. Since their discovery in the mid-1950s, our understanding of these defective particles has changed as we explored their mechanisms and functions in viral infections. Initially thought to be an artifact of cell culturing practices, these particles were mostly dismissed as important biological entities. Throughout the years, this idea changed as we began to see the appearance of these particles in a wider variety of viral infections. A thorough literature search has indicated that almost every

family of RNA viruses can produce defective interfering particles (**Appendix A**). Defective viral genomes have been shown to promote the establishment of viral persistence, prolong the infectious period of the host, stimulate the host's immune system, and have been proposed to be exploited in the use of vaccines or antiviral therapies[80]. More importantly, DIPs have been found in clinical samples where patient outcomes were correlated to the concentrations of these viral parasites[125]. Therefore, understanding how these genomes are formed and the roles they play during infections has critical importance.

Here, I sought to characterize and determine the molecular mechanisms behind the formation of defective interfering RNAs of Flock House virus. My overall goal was to develop a comprehensive set of tools to study this model insect virus, which can then be applied to other systems, while also gaining insights into the fundamental processes of virus development.

Previously, FHV has been demonstrated to readily produce DI-RNAs in both cell culture[44] and *D. melanogaster*[129] but the mechanism of their formation was not well known. Therefore, in **Chapter 2**, I used a combination of novel short read- and long- read sequencing technologies to characterize the stepwise progression of DI-RNA generation during viral passaging. Using ClickSeq (short- read RNAseq) I was able to determine the precise identity of RNA recombination sites. Long- read MinION sequencing allowed me to correlate these events within one genome and determine their relative frequency within a population. I was able to show evidence that fully formed, 'mature', DI-RNAs (characterized by multiple deletions) appeared early during passaging and accumulated quickly with little presence of partially formed species (genomes with a single deletion). This suggests that mature FHV DI-RNA genomes either form in one swift step, the intermediate species are not competitive, or there are selection pressures filtering for certain species.

In **Chapter 3**, I wanted to begin to explore these concepts by determining and elucidating some of the potential selective pressures imposed upon DI-RNAs to understand why there is an accumulation of these mature defective genomes. Primarily, the mechanisms of viral packaging can impose a restrictive barrier that a defective genome has to overcome[190]. Genome packaging is a vital step of a virus's life cycle and therefore, is a highly regulated process. While studies have shown that packaged RNA plays an important role in capsid stability, the exact mechanisms behind particle formation and genome packaging are still widely unclear. Therefore, I applied a combination of structural techniques to determine what defective interfering particles 'look' like. Using native ultra-high mass spectrometry, I was able to show that viral populations full of defective genomes had the same mass as wild type particles. Furthermore, I was able to confirm these results using cryo- electron microscopy where reconstructed models of defective particles were almost indistinguishable from their wild type counterparts. These data suggest that FHV has a mechanism to 'measure' the amount of RNA that is packaged into the capsid which can force the selection of certain genomes that meet a certain length requirement.

Overall, more studies are needed to determine what mechanisms are at play in the formation of DI-RNAs within a cell. Here, I would speculate that there is an enormous variation of defective viral genome species not packaged within particles. I suspect that a combination of mechanisms promote the formation of defective genomes. In **Chapter 1**, I reviewed both polymerase dependent and polymerase independent mechanisms of genomic recombination. While not empirically tested, analysis of the common recombination events found during FHV passaging can support multiple of the proposed mechanisms. For example, copy choice recombination is a polymerase driven event when the polymerase falls off the template and reattaches to another portion of the genome to resume elongation. Analysis of the nucleotide frequencies at the recombination junctions indicates a preference for As and Us, which is thought

to be a driver for polymerase jumping[79]. Furthermore, some of the most common recombination events appear to have short sequence homology at their junctions suggesting targets for re-priming. For other mechanisms, it has been suggested that strong secondary structures act as inducers of recombination events where the polymerase bypasses hairpin structures during replication. Interestingly, a handful of the identified recombination events aligned with predicted RNA secondary structures. Lastly, long read sequencing identified a DI-RNA species that contained a complex rearrangement of the FHV genome (as discussed in **Chapter 2**). While less commonly supported in the field, I suspect that this genomic rearrangement could have arose during a 'break and re-ligate' scenario. Ultimately all these mechanisms need not be mutually exclusive and could be occurring concurrently. More studies would be needed to determine what species are present within host cells and what the mechanisms are that drive their formation.

Nevertheless, whether there is a common mechanism that drives the formation of DVGs or not, I have shown that there are mechanisms for the selection of certain genomes. For example, almost ubiquitously I and others have found that defective genomes conserve important regulatory elements[46]. This suggests that maintaining certain genomic elements is the first criteria that a defective genome must meet in order to accumulate. Retaining the replication signal would be vital for that genome to be efficiently copied by the polymerase, while the packaging signal is crucial for its encapsidation. Lastly, structural analysis of Flock House virus particles seems to suggest that maintenance of specific genome length is the final checkpoint that DI-RNAs must pass for them to be packaged and propagated during infections (as discussed in **Chapter 3**). Fundamentally, these structural studies have not only helped us understand the selective pressures imposed upon DI- formation but can also help us learn about the fundamentals of viral packaging mechanisms.

While we have learned a lot from these studies there are still many questions left unanswered. Primarily, we still have little understanding into the mechanism of formation of DI-RNAs. To better understand the fundamental driving forces in their formation we need to determine what species of DI-RNAs are present within the cell, not just what is packaged, which can help determine if the events we observe in packaged particles can be attributed to the mechanism of formation or the selection process imposed. A higher intercellular diversity of species would support the hypothesis that there are many levels of selection that a DI-RNA must pass in order to propagate further. While less likely, if the species of DI-RNAs is the same as what was identified in packaged particles, then the implication is that the mechanism of formation is what drives the evolution of defective genomes and packaging selection may not be as important.

Furthermore, understanding the mechanisms by which DI-RNAs are formed would not only give us an insight into why and how these genomes arise but it would also provide information on other parts of the viral life cycle. Here, a variety of experiments can be conducted. Firstly, the most popular model in DI- genesis is a polymerase dependent model. This is thought to be driven during the elongation step of replication through re-priming at homologous sequences or by 'jumping' at strong secondary structures (reviewed in **Chapter 1**). Both concepts are sequence dependent and therefore could be tested by introducing perturbations in the sequences hypothesized to drive these events followed by characterizing the consequent DI-species. If there are observed changes in the location of recombination events this would strongly suggest that specific sequences drive recombination events. In this case it wouldn't be clear as to what exactly about the sequence is the driving force, but further experiments could be conducted to tease out the cause (e.g. RNA structure studies, rescued sequence homology experiments, identification of potential bound proteins).

The fidelity and processivity of the polymerase have also been suggested to have an impact on how accurately a template genome is transcribed. The concept of viral fidelity has long been discussed amongst virologist where it is thought that a higher mutation rate actually benefits a viral population. The idea is that, to a limited extent, a wider varying population (caused by genomic mutations) is better suited to handle and adapt to the ever-changing environment that a virus encounters as it moves within a host and from host to host. While it is widely debated if viruses have actually evolved to deliberately be more mutagenic or if a higher error rate is just a by-product of the desire to be a fast replicator, the fact is that many RNA viruses express these characteristics[262]. Whatever the driving force may be, I would like to think that fidelity could also have stark implications on the formation of DI-RNAs in these viruses. With increased speeds of replication, it wouldn't be a far stretch to say that mutations aren't the only mistakes that a polymerase makes during replication. Therefore, it would be interesting to examine the effect of fidelity on the formation of DI-RNAs in viruses.

Additionally, the less adopted model for DI- formation, a polymerase independent model of formation, should also be explored in order to better understand the mechanisms of DI-RNA formation. Here, other factors, such as the host, could play a major role in determining what species of defective genomes are generated. Differential gene expression analysis would be able to elucidate what host genes are up- or down- regulated in the presence of virus with high populations of DI-RNAs. Protein/virus RNA interaction studies would be able to determine what proteins specifically interact with the virus. Identified factors could be further analyzed in their influence in DI- formation. A similar study to the one proposed has been conducted in TBSV infections where Prasanth *et al.* identified and confirmed a wide variety of influential host factors[263]. That study was limited to yeast, an unnatural host to the plant virus so studies in more natural hosts would be required to confirm these results. Studies like this are not only import to

improve our understanding of viral defective genomes but also in the fundamental understanding of the interactions between virus and host.

When I started my studies, my initial goal was to determine the molecular mechanism that drive the formation of DI-RNAs in RNA viruses. Through this process I stumbled upon the identification of certain selective pressures that are imposed upon genomes for them to be effectively passaged and propagated. One of the major factors was the mechanism of packaging and how a virus capsid is able to 'know' that it is encapsidating the correct genome(s) and in the right amount. This has sparked the idea that viruses have a 'molecular measuring tape' that they use to identify the right genome(s) to package. Flock House virus contains an arginine rich motif (ARM) that is hypothesized to play this role where studies have begun to show that perturbations in this region result in genomic packaging defects[206, 264, 265]. Interestingly, many RNA viruses contain very similar ARM motifs where correlation studies have indicated that a virus with a larger ARM region has a longer genome. Further studies would be needed to validate this hypothesis that it is the positive charged arginines that interact with the negatively charged genome to identify it.

Fundamentally, I have always been curious as to the true origin and purpose for a virus to generate DI-RNAs. Are they direct and deliberate products that overall benefit the virus in some way or are they just an artifact of the virus's natural life cycle? Or are they deliberately generated by the host to be used as an additional piece of their arsenal to protect itself? Do these defective genomes code for functional proteins? As more studies are being conducted, we are only starting to fully understand their roles in natural infections. Historically, it has been thought that DI-RNAs help establish persistent infections, but we are only now starting to understand that they potentially play so many more roles and their origin of formation.

Lastly, throughout this process into the discovery of the mechanisms and selective pressures imposed upon the formation of DI-RNAs in Flock House virus, many technological developments have been made. Historically, DI-RNAs have been thought to be an artifact of cell culturing. This is a misconception due to the low sensitivity of detection of the employed biologic techniques. Identification and characterizations of defective RNAs included things like visualization on electrophoresis gels, Sanger sequencing lower molecular weight fragments, and cloning. While relatively accurate, these techniques were not very sensitive requiring these products to be in high abundance and only captured a small subset of the defective population. Overall, this led researchers to believe that DI-RNAs were only an artifact of cell culturing because it was only under those conditions that the techniques of the time could identify them. With the improvements in technology and the development of next generation sequencing (NGS) our detection for these low-level events significantly improved and the identification of DI-RNAs in natural infections began apparent. Unfortunately, even with the increasing amount of evidence of the presence of defective RNAs in natural infections there are still many sceptics to their overall importance.

The implantation of NGS into DI- research has vastly improved the field due to its high throughput approach, but it was still without any caveats. The standard methods for preparing RNA for sequencing introduced many biases, including artefactual recombination events. Even though these methods were able to identify viral recombination, it was hard to distinguish if an event was real or was it a product of the enzymatic ligation steps employed during NGS library preparation. From this ClickSeq was born. ClickSeq and the optimizations done to it have allowed us to confidently and accurately call the boundaries of recombination junctions. Even though ClickSeq was originally invented to better study viral recombination we are finding that this technique has many more applications in not only FHV studies but in fields outside of the field of

virology. With small modifications to the standard protocol we are able to use the technique for things like transcriptomics and poly-adenylation studies with PAC-seq and to profile the nascent 3' ends of transcripts, thought to be caused by polymerase pausing, using 2PC-seq.

Furthermore, the entire success of the experiments presented here were reliant on novel and innovative new technologies. Long-read Nanopore sequencing was fundamental in helping correlate multiple recombination events within one genome. The fundamental concept of analyzing a single viral genome is not new, but the power in this long-read sequencing is for its ability to do so on a high throughput and grand scale. Currently, there are other long-read sequencing techniques on the market (such as PacBio long-read sequencing) but high equipment costs and complicated processes made that technology relatively inaccessible. Nanopore technology has been able to make this technique extremely attractive due to its extremely low barrier of entry giving researchers the power to analyze entire sequences of any biological nature in a cost-effective way. As with any new product there are still some caveats, such as the high error rate of the sequence base calling, which with future developments is posited to improve.

Through the process of my studies I have also been able to implement the use ultra-high mass native mass spectrometry. Native mass spectroscopy is a technique that is an improvement on standard MS, allowing scientists to study whole proteins and protein complexes in their native states. This was a huge stride in proteomics research as now entire proteins were able to be measured. Informative as it was, the technology was still limited to studying smaller and highly charged proteins, which limited the types of research that could be conducted. Studying large, net-neutral complexes, like entire virus particles, was unheard of at the time. Improvements to the mass spectrometry instruments have made these types of studies possible [172]. So far, using this technology the field has only been able to measure complexes up to <10Mda, but it is only a matter of time as continued development pushes this boundary over and over again[172].

Overall, I have shown that a combination of novel techniques could be effectively applied to comprehensively study and characterize the broken genomes of viruses. While my study used Flock House virus, a model insect virus, the tools and findings that I described here could easily be translated to other viruses and even other non-viral systems. Though the goal of my project was to answer and explore the fundamentals to the mechanisms and biology of defective genomes I hope that the groundwork of connecting a wide range of seemingly different tools has been laid and will continue to be implemented in future studies.

APPENDIX A: TABLE OF VIRUSES WITH DEFECTIVE (INTERFERING) PARTICLES

Particle Sym.	Envelope?	Baltimore Class[34]	Genome	Family	Common Name	Publication
Icosahedral	No	III	dsRNA	Reoviridae		41, 113
	No	III	dsRNA	Birnaviridae	Infectious pancreatic necrosis	266
	No	IV	(+)ssRNA	Caliciviridae	Vesicular exanthema of swine	267
	No	IV	(+)ssRNA	Picornaviridae	Poliovirus	76, 268-270
					Coxsackie	271
					Hepatitis A	121
					Foot-and-mouth disease	43
	No	IV	(+)ssRNA	Secoviridae	Tomato ringspot	42
	No	IV	(+)ssRNA	Nodaviride	Flock house	44, 45, 86, 98, 129, 143, 188
	No	IV	(+)ssRNA	Tombusviridae	Tomato bushy stunt	85, 93, 272
					Cucumber necrosis	128
					Cymbidium ringspot virus	69, 273
					Turnip crinkle virus	4
	No	IV	(+)ssRNA	Bromoviridae	Brome mosaic	274
	Yes	IV	(+)ssRNA	Flaviviridae	Dengue	120, 123, 275
					Hepatitis C	38, 79, 149
					West Nile	150, 276
					Murray Valley encephalitis	277
					Bovine viral diarrhea	124
	Yes	IV	(+)ssRNA	Togaviridae	Sindbis	47, 63, 278
					Japanese encephalitis	112
					Rubella	39, 279
					Semliki Forest	40, 48, 81, 132, 133
	Yes	VI	(+)ssRNA-rRT	Retrovirus	Human immunodeficiency	147, 280
					Murine leukemia	281, 282
					Rous sarcoma	283

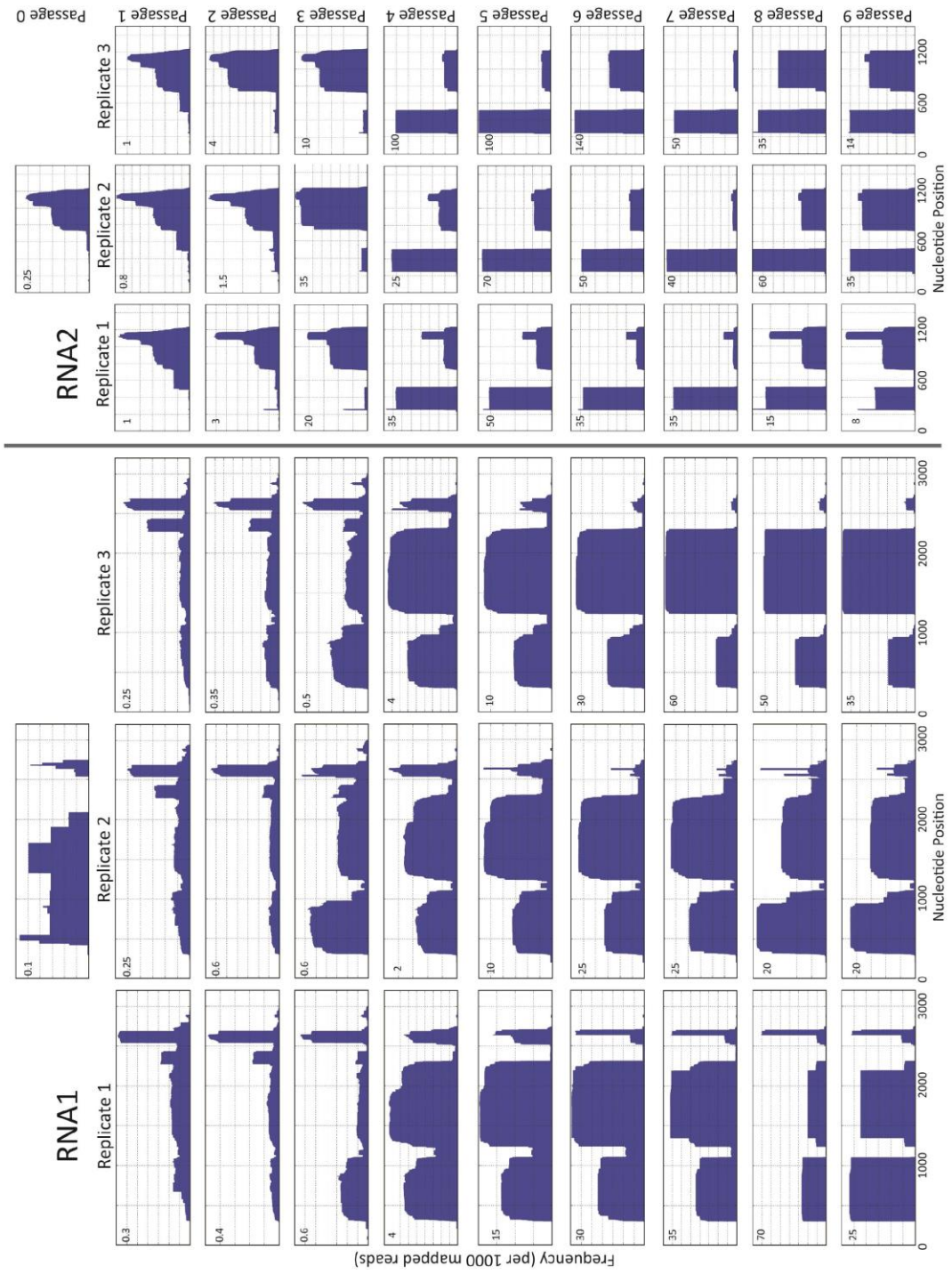
Continued on next page.

Particle Sym.	Envelope?	Baltimore Class[34]	Genome	Family	Common Name	Publication
Helical	No	IV	(+)ssRNA	*Closteroviridae*	Citrus tristeza	284
	No	IV	(+)ssRNA	*Virgaviridae*	Tobacco rattle	285
	Yes	IV	(+)ssRNA	*Coronaviridae*	Mouse hepatitis	146, 177, 286, 287
					Infectious bronchitis	288
	Yes	V	(-)ssRNA	*Filoviridae*	Ebola	126
	Yes	V	(-)ssRNA	*Rhabdoviridae*	Vesicular stomatitis virus	16-18, 35, 49, 50, 72, 87-89, 115, 118, 289
					Infectious hematopoietic necrosis	36
					Rabies	109, 290, 291
	Yes	V	(-)ssRNA	*Bunyaviridae*	La Crosse	292
					Rift Valley Fever	15
	Yes	V	(-)ssRNA	*Orthomyxoviridae*	Influenza	2, 13, 37, 52, 70, 80, 90, 125, 131, 134, 141, 185
	Yes	V	(-)ssRNA	*Paramyxoviridae*	Newcastle disease	14
					Sendai	19, 22, 73, 100, 101, 189, 293-295
					Measles	73, 122, 148, 294, 295
					Mumps	296
					Human parainfluenza	51
	Yes	V	(-)ssRNA	*Pneumoviridae*	Human respiratory syncytial	74, 104
					Human metapneumovirus	55
	Yes	V	(-)ssRNA	*Areanaviridae*	Lymphocytic choriomeningitis	20, 297, 298

Appendix A: RNA viruses with known defective (interfering) particles

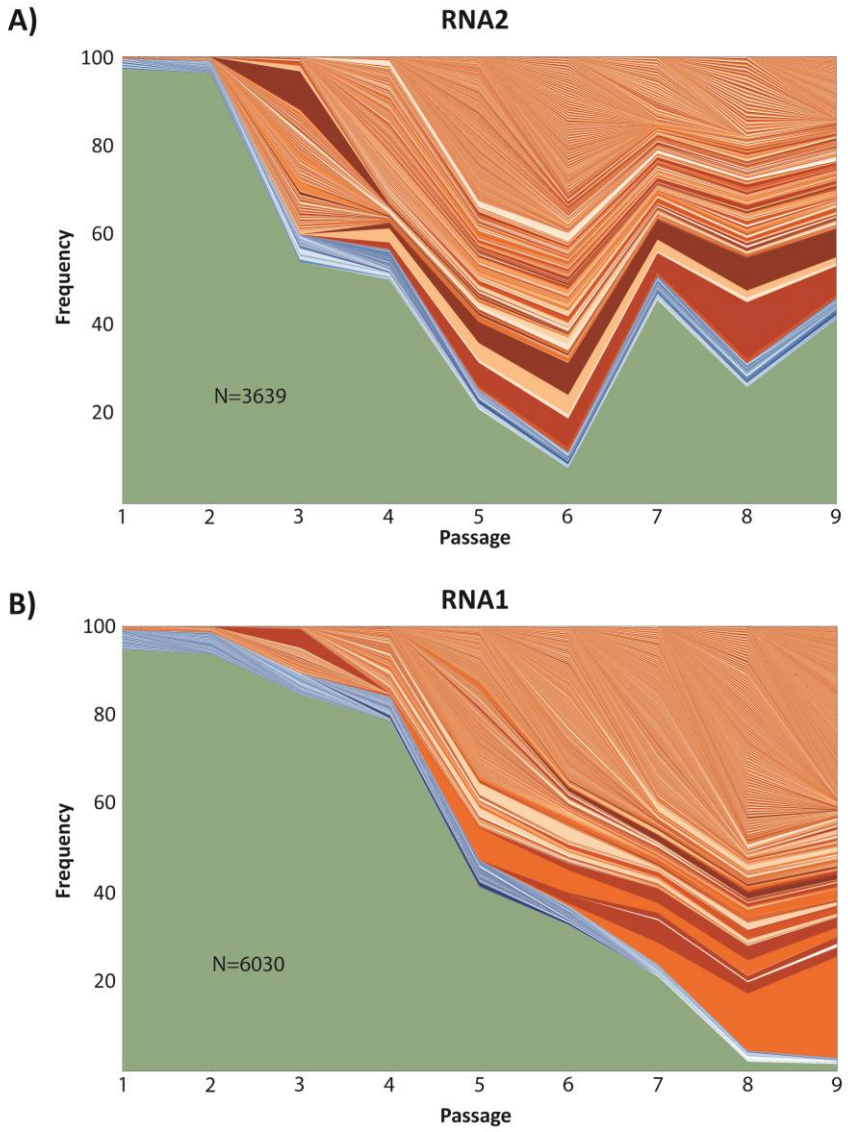
Tabulated list of families and common names of viruses with defective viral genomes. Publication list is not comprehensive and focuses on reviews, publications citied in this manuscript, and the first account of defective particles in that virus. (see Ref 34 for an explanation of the Baltimore classification).

APPENDIX B: CHAPTER 2 SUPPLEMENTAL DATA



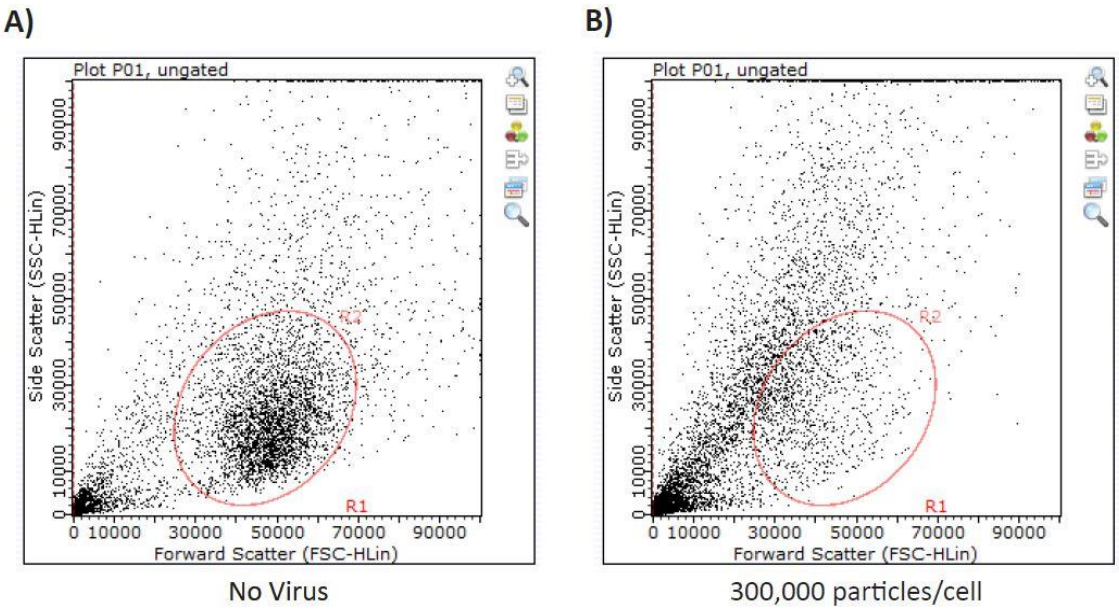
Appendix B.1: ClickSeq conservation maps of all passages and replicates.

Conservation maps similar to those illustrated in **Figure 2.7** are shown for every passage in every replicate including the original inoculum (P0).



Appendix B.2: Stacked-area plot of showing the pathways of FHV DI-RNA evolution.

Similar to **Figure 2.12**, except all species are included, including genomes represented by only one MinION nanopore read. The passage number is indicated on the x-axis and the stacked frequencies of each detected defective RNA is shown in the y-axis. Each non-contiguous color represents a specific genome characterized by MinION nanopore sequencing. Wild-type genomes are colored green, genomes with one deletion are colored in shades of blue, and genomes with two or more deletions are colored in shades of oranges (using the same color scheme as in **Figure 2.9B**).



Appendix B.3: Scatterplots showing live cell gating.

Screenshots of the InCyte software (of the Guava easyCyte HT flow cytometer) indicating gating used to count live cells. Cell counts were used to generate **Figure 2.13C.**

Supplemental Data Files: Data files can be accessed online following the provided links.

Appendix B.4: Raw virus recombination events data from *ViReMa* analysis of ClickSeq data.

Each passage in each replicate as well as the inoculum is shown. Output format, is given as "*DonorCoord_to_AcceptorCoord_#_Counts".* This is the raw data used to populate **Table 2.1** and **Table 2.2**. Download file at: https://doi.org/10.1371/journal.ppat.1006365.s009

Appendix B.5: Genomes characterized by MinION nanopore sequencing.

Table reports the annotated genomes (wild-type or defective) and the number of mapped reads in each passage. This the raw data used to populate the stacked-area plots in **Figure 2.9, Figure 2.12** and **Appendix B.2**. Download file at: https://doi.org/10.1371/journal.ppat.1006365.s010

Appendix B.6: Accession numbers for all raw data files

All raw Illumina data and demultiplexed MinION nanopore data passing quality filters (comprising 2D, template and complement strands) associated with this manuscript are available on the SRA NCBI archive with study number SRP094723 and BioProject number PRJNA352872.

APPENDIX C: COPYRIGHT PERMISSIONS

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Jun 01, 2018

This Agreement between Elizabeth Jaworski ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center

License Number	4358350575550
License date	May 29, 2018
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	ClickSeq: Replacing Fragmentation and Enzymatic Ligation with Click-Chemistry to Prevent Sequence Chimeras
Licensed Content Author	Elizabeth Jaworski, Andrew Routh
Licensed Content Date	Jan 1, 2018
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	<501
Author of this Springer Nature content	yes
Title	The broken genome: Merging cutting edge technologies for a molecular understanding of the Flock house virus defective interfering particle
Instructor name	Andrew Routh
Institution name	University of Texas Medical Branch
Expected presentation date	Aug 2018
Requestor Location	Elizabeth Jaworski 816 Walker St LA MARQUE, TX 77568 United States Attn: Elizabeth Jaworski
Billing Type	Invoice
Billing Address	Elizabeth Jaworski 816 Walker St LA MARQUE, TX 77568 United States Attn: Elizabeth Jaworski
Total	0.00 USD
Terms and Conditions	

Springer Customer Service Centr e GmbH (the Licensor) hereby grants you a non-exclusive, world-wide licence to reproduce the material and for the purpose and requirements specified in the attached copy of your order form, and for no other use, subject to the conditions below:

1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

   If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

2. Where print only permission has been granted for a fee, separate permission must be obtained for any additional electronic re-use.

3. Permission granted free of charge for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

4. A licence for 'post on a website' is valid for 12 months from the licence date. This licence does not cover use of full text articles on websites.

5. Where 'reuse in a dissertation/thesis' has been selected the following terms apply: Print rights for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

6. Permission granted for books and journals is granted for the lifetime of the first edition and does not apply to second and subsequent editions (except where the first edition permission was granted free of charge or for signatories to the STM Permissions Guidelines http://www.stm-assoc.org/copyright-legal-affairs/permissions/permissions-guidelines/), and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence.

7. Rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

8. The Licensor's permission must be acknowledged next to the licensed material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

9. Use of the material for incidental promotional use, minor editing privileges (this does not include cropping, adapting, omitting material or any other changes that affect the meaning, intention or moral rights of the author) and copies for the disabled are permitted under this licence.

10. Minor adaptations of single figures (changes of format, colour and style) do not require the Licensor's approval. However, the adaptation should be credited as shown in Appendix below.

REFERENCES

1. von Magnus P. *Studies on interference in experimental influenza*. Almqvist & Wiksell, 1947.

2. von Magnus P. Incomplete forms of influenza virus. *Adv Virus Res*. 1954; 2: 59-79.

3. Huang AS and Baltimore D. Defective viral particles and viral disease processes. *Nature*. 1970; 226: 325-7.

4. Li XH, Heaton LA, Morris TJ and Simon AE. Turnip crinkle virus defective interfering RNAs intensify viral symptoms and are generated de novo. *Proc Natl Acad Sci U S A*. 1989; 86: 9173-7.

5. Friedewald WF and Pickels EG. Centrifugation and ultrafiltration studies on allantoic fluid preparation of influenza virus. *J Exp Med*. 1944; 79: 301-17.

6. Pickels EG. Sedimentation in the angle centrifuge. *J Gen Physiol*. 1943; 26: 341-59.

7. Henle W. Studies on host-virus interactions in the chick embryo-influenza virus system; adsorption and recovery of seed virus. *J Exp Med*. 1949; 90: 1-11.

8. Hoyle L. The growth cycle of influenza virus A; a study of the relations between virus, soluble antigen and host cell in fertile eggs inoculated with influenza virus. *Br J Exp Pathol*. 1948; 29: 390-9.

9. von Magnus P. Propagation of the PR8 strain on influenza A virus in chick embryos. I The influence of various experimental conditions on virus multiplication. *Acta Pathol Microbiol Scand*. 1951; 28: 250-77.

10. von Magnus P. Propagation of the PR8 strain of influenza A virus in chick embryos. II. The formation of incomplete virus following inoculation of large doses of seed virus. *Acta Pathol Microbiol Scand*. 1951; 28: 278-93.

11. von Magnus P. Propagation of the PR8 strain of influenza A virus in chick embryos. III. Properties of the incomplete virus produced in serial passages of undiluted virus. *Acta Pathol Microbiol Scand*. 1951; 29: 157-81.

12. von Magnus P. Propagation of the PR8 strain of influenza A virus in chick embryos. IV. Studies on the factors involved in the formation of incomplete virus upon serial passage of undiluted virus. *Acta Pathol Microbiol Scand*. 1952; 30: 311-35.

13. Beale AJ and Finter NB. The infectivity of chorio-allantoic membrane influenza virus and incomplete influenza virus by the six-hour soluble antigen production test. *J Hyg (Lond)*. 1956; 54: 68-78.

14. GranoffF A. Noninfectious forms of Newcastle disease and influenza viruses; studies on noninfectious virus occurring within cells that are producing fully infectious virus. *Virology*. 1955; 1: 516-32.

15. MIMS CA. Rift Valley Fever virus in mice. IV. Incomplete virus; its production and properties. *Br J Exp Pathol*. 1956; 37: 129-43.

16. Cooper PD and Bellett AJ. A transmissible interfering component of vesicular stomatitis virus preparations. *J Gen Microbiol*. 1959; 21: 485-97.

17. Bellett AJ and Cooper PD. Some properties of the transmissible interfering component of vesicular stomatitis virus preparations. *J Gen Microbiol*. 1959; 21: 498-509.

18. Huang AS, Greenawalt JW and Wagner RR. Defective T particles of vesicular stomatitis virus. I. Preparation, morphology, and some biologic properties. *Virology*. 1966; 30: 161-72.

19. Sokol F, Neurath AR and Vilcek J. Formation of incomplete Sendai virus in embryonated eggs. *Acta Virol*. 1964; 8: 59-67.

20. Lehmann-Grube F, Slenczka W and Tees R. A persistent and inapparent infection of L cells with the virus of lymphocytic choriomeningitis. *J Gen Virol*. 1969; 5: 63-81.

21. Duesberg PH. The RNA of influenza virus. *Proc Natl Acad Sci U S A*. 1968; 59: 930-7.

22. Kingsbury DW, Portner A and Darlington RW. Properties of incomplete Sendai virions and subgenomic viral RNAs. *Virology*. 1970; 42: 857-71.

23. Huang AS. Defective interfering viruses. *Annu Rev Microbiol*. 1973; 27: 101-17.

24. Rapp F. Defective DNA animal viruses. *Annu Rev Microbiol*. 1969; 23: 293-316.

25. Patil BL and Dasgupta I. Defective Interfering DNAs of Plant Viruses. *Critical Reviews in Plant Sciences*. 2006; 25: 47-64.

26. Barrett ADT and Dimmock NJ. Defective Interfering Viruses and Infections of Animals. In: Clarke A, Compans RW, Cooper M, et al., (eds.). *Current Topics in Microbiology and Immunology*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1986, p. 55-84.

27. Manzoni TB and López CB. Defective (interfering) viral genomes re-explored: impact on antiviral immunity and virus persistence. *Future Virology*. 2018; 13: 493-503.

28. López CB. Defective viral genomes: critical danger signals of viral infections. *J Virol*. 2014; 88: 8720-3.

29. Palukaitis P. Chapter 50 - Satellite Viruses and Satellite Nucleic Acids. In: Hadidi A, Flores R, Randles JW and Palukaitis P, (eds.). *Viroids and Satellites*. Boston: Academic Press, 2017, p. 545-52.

30. Qiu X, Wu L, Huang H, et al. Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Appl Environ Microbiol*. 2001; 67: 880-7.

31. J. DN. The biological significance of defective interfering viruses. *Reviews in Medical Virology*. 1991; 1: 165-76.

32. Reczko E. Elektronenmikroskopische Untersuchungen am Virus der Stomatitis vesicularis. *Archiv für die gesamte Virusforschung*. 1961; 10: 588-605.

33. Cureton DK, Massol RH, Whelan SP and Kirchhausen T. The length of vesicular stomatitis virus particles dictates a need for actin assembly during clathrin-dependent endocytosis. *PLoS Pathog*. 2010; 6: e1001127.

34. Flint SJ, Racaniello VR, Rall GF, Skalka AM and Enquist LW. *Principles of virology*. 4th edition. ed. Washington, DC: ASM Press, 2015, p.volumes.

35. Odenwald WF, Arnheiter H, Dubois-Dalcq M and Lazzarini RA. Stereo images of vesicular stomatitis virus assembly. *J Virol*. 1986; 57: 922-32.

36. Drolet BS, Chiou PP, Heidel J and Leong JA. Detection of truncated virus particles in a persistent RNA virus infection in vivo. *J Virol*. 1995; 69: 2140-7.

37. Blough HA and Merlie JP. The lipids of incomplete influenza virus. *Virology*. 1970; 40: 685-92.

38. Prince AM, Huima-Byron T, Parker TS and Levine DM. Visualization of hepatitis C virions and putative defective interfering particles isolated from low-density lipoproteins. *J Viral Hepat*. 1996; 3: 11-7.

39. Bohn EM and Van Alstyne D. The generation of defective interfering Rubella virus particles. *Virology*. 1981; 111: 549-54.

40. Bruton CJ and Kennedy SI. Defective-interfering particles of Semliki Forest Virus: structural differences between standard virus and defective-interfering particles. *J Gen Virol*. 1976; 31: 383-95.

41. Nonoyama M, Watanabe Y and Graham AF. Defective virions of reovirus. *J Virol*. 1970; 6: 226-36.

42. Stace-Smith R. Purification and properties of tomato ringspot virus and an RNA-deficient component. *Virology*. 1966; 29: 240-7.

43. García-Arriaza J, Manrubia SC, Toja M, Domingo E and Escarmís C. Evolutionary transition toward defective RNAs that are infectious by complementation. *J Virol*. 2004; 78: 11678-85.

44. Jovel J and Schneemann A. Molecular characterization of Drosophila cells persistently infected with Flock House virus. *Virology*. 2011; 419: 43-53.

45. Routh A and Johnson JE. Discovery of functional genomic motifs in viruses with ViReMa-a Virus Recombination Mapper-for analysis of next-generation sequencing data. *Nucleic Acids Res*. 2014; 42: e11.

46. Pathak KB and Nagy PD. Defective Interfering RNAs: Foes of Viruses and Friends of Virologists. *Viruses*. 2009; 1: 895-919.

47. Monroe SS and Schlesinger S. Common and distinct regions of defective-interfering RNAs of Sindbis virus. *J Virol*. 1984; 49: 865-72.

48. Lehtovaara P, Söderlund H, Keränen S, Pettersson RF and Kääriäinen L. Extreme ends of the genome are conserved and rearranged in the defective interfering RNAs of Semliki Forest virus. *J Mol Biol*. 1982; 156: 731-48.

49. Li T and Pattnaik AK. Replication signals in the genome of vesicular stomatitis virus and its defective interfering particles: identification of a sequence element that enhances DI RNA replication. *Virology*. 1997; 232: 248-59.

50. Timm C, Akpinar F and Yin J. Quantitative characterization of defective virus emergence by deep sequencing. *J Virol*. 2014; 88: 2623-32.

51. Killip MJ, Young DF, Gatherer D, et al. Deep sequencing analysis of defective genomes of parainfluenza virus 5 and their role in interferon induction. *J Virol*. 2013; 87: 4798-807.

52. Saira K, Lin X, DePasse JV, et al. Sequence analysis of in vivo defective interfering-like RNA of influenza A H1N1 pandemic virus. *J Virol*. 2013; 87: 8064-74.

53. Rosskopf JJ, Upton JH, Rodarte L, et al. A 3' terminal stem-loop structure in Nodamura virus RNA2 forms an essential cis-acting signal for RNA replication. *Virus Res*. 2010; 150: 12-21.

54. Zhong W, Dasgupta R and Rueckert R. Evidence that the packaging signal for nodaviral RNA2 is a bulged stem-loop. *Proc Natl Acad Sci U S A*. 1992; 89: 11146-50.

55. van den Hoogen BG, van Boheemen S, de Rijck J, et al. Excessive production and extreme editing of human metapneumovirus defective interfering RNA is associated with type I IFN induction. *J Gen Virol*. 2014; 95: 1625-33.

56. SchÄFer W. Chapter VIII - The Comparative Chemistry of Infective Virus Particles and of other Virus-Specific Products : Animal Viruses**The survey of literature pertaining to this chapter was completed in October 1957. The author is greatly indebted to Dr. R. M. Franklin for the translation of the manuscript. In: Burnet FM and Stanley WM, (eds.). *General Virology*. Academic Press, 1959, p. 475-504.

57. Pogany J and Nagy PD. Authentic replication and recombination of Tomato bushy stunt virus RNA in a cell-free extract from yeast. *J Virol*. 2008; 82: 5967-80.

58. Wierzchoslawski R and Bujarski JJ. Efficient in vitro system of homologous recombination in brome mosaic bromovirus. *J Virol*. 2006; 80: 6182-7.

59. You S and Padmanabhan R. A novel in vitro replication system for Dengue virus. Initiation of RNA synthesis at the 3'-end of exogenous viral RNA templates requires 5'- and 3'-terminal complementary sequence motifs of the viral RNA. *J Biol Chem*. 1999; 274: 33714-22.

60. Kim MJ and Kao C. Factors regulating template switch in vitro by viral RNA-dependent RNA polymerases: implications for RNA-RNA recombination. *Proc Natl Acad Sci U S A*. 2001; 98: 4972-7.

61. Cheng CP and Nagy PD. Mechanism of RNA recombination in carmo- and tombusviruses: evidence for template switching by the RNA-dependent RNA polymerase in vitro. *J Virol*. 2003; 77: 12033-47.

62. Chuang C, Prasanth KR and Nagy PD. Coordinated function of cellular DEAD-box helicases in suppression of viral RNA recombination and maintenance of viral genome integrity. *PLoS Pathog*. 2015; 11: e1004680.

63. Monroe SS and Schlesinger S. RNAs from two independently isolated defective interfering particles of Sindbis virus contain a cellular tRNA sequence at their 5' ends. *Proc Natl Acad Sci U S A*. 1983; 80: 3279-83.

64. Eckerle LD, Albariño CG and Ball LA. Flock House virus subgenomic RNA3 is replicated and its replication correlates with transactivation of RNA2. *Virology*. 2003; 317: 95-108.

65. Wierzchoslawski R, Dzianott A, Kunimalayan S and Bujarski JJ. A transcriptionally active subgenomic promoter supports homologous crossovers in a plus-strand RNA virus. *J Virol*. 2003; 77: 6769-76.

66. Nagy PD, Pogany J and Simon AE. RNA elements required for RNA recombination function as replication enhancers in vitro and in vivo in a plus-strand RNA virus. *EMBO J*. 1999; 18: 5653-65.

67. Shapka N and Nagy PD. The AU-rich RNA recombination hot spot sequence of Brome mosaic virus is functional in tombusviruses: implications for the mechanism of RNA recombination. *J Virol*. 2004; 78: 2288-300.

68. DeStefano JJ, Bambara RA and Fay PJ. The mechanism of human immunodeficiency virus reverse transcriptase-catalyzed strand transfer from internal regions of heteropolymeric RNA templates. *J Biol Chem*. 1994; 269: 161-8.

69. Havelda Z, Dalmay T and Burgyán J. Secondary structure-dependent evolution of Cymbidium ringspot virus defective interfering RNA. *J Gen Virol*. 1997; 78 (Pt 6): 1227-34.

70. Jennings PA, Finch JT, Winter G and Robertson JS. Does the higher order structure of the influenza virus ribonucleoprotein guide sequence rearrangements in influenza viral RNA? *Cell*. 1983; 34: 619-27.

71. Lazzarini RA, Keene JD and Schubert M. The origins of defective interfering particles of the negative-strand RNA viruses. *Cell*. 1981; 26: 145-54.

72. Schubert M and Lazzarini RA. Structure and origin of a snapback defective interfering particle RNA of vesicular stomatitis virus. *J Virol*. 1981; 37: 661-72.

73. Kolakofsky D. Isolation and characterization of Sendai virus DI-RNAs. *Cell*. 1976; 8: 547-55.

74. Sun Y, Jain D, Koziol-White CJ, et al. Immunostimulatory Defective Viral Genomes from Respiratory Syncytial Virus Promote a Strong Innate Antiviral Response during Infection in Mice and Humans. *PLoS pathogens*. 2015; 11: e1005122.

75. Chetverin AB, Chetverina HV, Demidenko AA and Ugarov VI. Nonhomologous RNA recombination in a cell-free system: evidence for a transesterification mechanism guided by secondary structure. *Cell*. 1997; 88: 503-13.

76. Gmyl AP, Belousov EV, Maslova SV, Khitrina EV, Chetverin AB and Agol VI. Nonreplicative RNA recombination in poliovirus. *J Virol*. 1999; 73: 8958-65.

77. Gmyl AP, Korshenko SA, Belousov EV, Khitrina EV and Agol VI. Nonreplicative homologous RNA recombination: promiscuous joining of RNA pieces? *RNA*. 2003; 9: 1221-31.

78. Gallei A, Pankraz A, Thiel HJ and Becher P. RNA recombination in vivo in the absence of viral replication. *J Virol*. 2004; 78: 6271-81.

79. Galli A and Bukh J. Comparative analysis of the molecular mechanisms of recombination in hepatitis C virus. *Trends Microbiol*. 2014; 22: 354-64.

80. Dimmock NJ and Easton AJ. Defective interfering influenza virus RNAs: time to reevaluate their clinical potential as broad-spectrum antivirals? *J Virol*. 2014; 88: 5217-27.

81. Stark C and Kennedy SI. The generation and propagation of defective-interfering particles of Semliki Forest virus in different cell types. *Virology*. 1978; 89: 285-99.

82. Nagy PD, Pogany J and Lin JY. How yeast can be used as a genetic platform to explore virus-host interactions: from 'omics' to functional studies. *Trends Microbiol*. 2014; 22: 309-16.

83. Serviene E, Jiang Y, Cheng CP, Baker J and Nagy PD. Screening of the yeast yTHC collection identifies essential host factors affecting tombusvirus RNA recombination. *J Virol*. 2006; 80: 1231-41.

84. Serviene E, Shapka N, Cheng CP, et al. Genome-wide screen identifies host genes affecting viral RNA recombination. *Proc Natl Acad Sci U S A*. 2005; 102: 10545-50.

85. White KA and Morris TJ. Nonhomologous RNA recombination in tombusviruses: generation and evolution of defective interfering RNAs by stepwise deletions. *J Virol*. 1994; 68: 14-24.

86. Jaworski E and Routh A. Parallel ClickSeq and Nanopore sequencing elucidates the rapid evolution of defective-interfering RNAs in Flock House virus. *PLoS Pathog*. 2017; 13: e1006365.

87. Doyle M and Holland JJ. Prophylaxis and immunization in mice by use of virus-free defective T particles to protect against intracerebral infection by vesicular stomatitis virus. *Proc Natl Acad Sci U S A*. 1973; 70: 2105-8.

88. Holland JJ and Doyle M. Attempts to detect homologous autointerference in vivo with influenza virus and vesicular stomatitis virus. *Infect Immun*. 1973; 7: 526-31.

89. Rabinowitz SG, Dal Canto MC and Johnson TC. Infection of the central nervous system produced by mixtures of defective-interfering particles and wild-type vesicular stomatitis virus in mice. *J Infect Dis*. 1977; 136: 59-74.

90. Gamboa ET, Harter DH, Duffy PE and Hsu KC. Murine influenza virus encephalomyelitis. III. Effect of defective interfering virus particles. *Acta Neuropathol*. 1976; 34: 157-69.

91. Welsh RM, Lampert PW and Oldstone MB. Prevention of virus-induced cerebellar diseases by defective-interfering lymphocytic choriomeningitis virus. *J Infect Dis*. 1977; 136: 391-9.

92. Calain P and Roux L. Functional characterisation of the genomic and antigenomic promoters of Sendai virus. *Virology*. 1995; 212: 163-73.

93. Jones RW, Jackson AO and Morris TJ. Defective-interfering RNAs and elevated temperatures inhibit replication of tomato bushy stunt virus in inoculated protoplasts. *Virology*. 1990; 176: 539-45.

94. Hannon GJ. RNA interference. *Nature*. 2002; 418: 244-51.

95. Ding SW and Voinnet O. Antiviral immunity directed by small RNAs. *Cell*. 2007; 130: 413-26.

96. Silhavy D, Molnár A, Lucioli A, et al. A viral protein suppresses RNA silencing and binds silencing-generated, 21- to 25-nucleotide double-stranded RNAs. *EMBO J*. 2002; 21: 3070-80.

97. Havelda Z, Hornyik C, Válóczi A and Burgyán J. Defective interfering RNA hinders the activity of a tombusvirus-encoded posttranscriptional gene silencing suppressor. *J Virol*. 2005; 79: 450-7.

98. Poirier EZ, Goic B, Tomé-Poderti L, et al. Dicer-2-Dependent Generation of Viral DNA from Defective Genomes of RNA Viruses Modulates Antiviral Immunity in Insects. *Cell Host Microbe*. 2018; 23: 353-65.e8.

99. Loo YM and Gale M. Immune signaling by RIG-I-like receptors. *Immunity*. 2011; 34: 680-92.

100. Xu J, Mercado-López X, Grier JT, et al. Identification of a Natural Viral RNA Motif That Optimizes Sensing of Viral RNA by RIG-I. *MBio*. 2015; 6: e01265-15.

101. Baum A, Sachidanandam R and García-Sastre A. Preference of RIG-I for short viral RNA molecules in infected cells revealed by next-generation sequencing. *Proc Natl Acad Sci U S A*. 2010; 107: 16303-8.

102. Sanchez David RY, Combredet C, Sismeiro O, et al. Comparative analysis of viral RNA signatures on different RIG-I-like receptors. *Elife*. 2016; 5: e11275.

103. Boldogh I, Albrecht T and Porter DD. Persistent Viral Infections. In: th and Baron S, (eds.). *Medical Microbiology*. Galveston (TX): University of Texas Medical Branch at Galveston
The University of Texas Medical Branch at Galveston., 1996.

104. Treuhaft MW and Beem MO. Defective interfering particles of respiratory syncytial virus. *Infect Immun*. 1982; 37: 439-44.

105. Sikkel MB, Quint JK, Mallia P, Wedzicha JA and Johnston SL. Respiratory syncytial virus persistence in chronic obstructive pulmonary disease. *Pediatr Infect Dis J*. 2008; 27: S63-70.

106. Tan JJL, Balne PK, Leo YS, Tong L, Ng LFP and Agrawal R. Persistence of Zika virus in conjunctival fluid of convalescence patients. *Sci Rep*. 2017; 7: 11194.

107. Labadie K, Larcher T, Joubert C, et al. Chikungunya disease in nonhuman primates involves long-term viral persistence in macrophages. *J Clin Invest*. 2010; 120: 894-906.

108. Chughtai AA, Barnes M and Macintyre CR. Persistence of Ebola virus in various body fluids during convalescence: evidence and implications for disease transmission and control. *Epidemiol Infect*. 2016; 144: 1652-60.

109. Kawai A, Matsumoto S and Tanabe K. Characterization of rabies viruses recovered from persistently infected BHK cells. *Virology*. 1975; 67: 520-33.

110. Roux L and Waldvogel FA. Establishment of Sendai virus persistent infection: biochemical analysis of the early phase of a standard plus defective interfering virus infection of BHK cells. *Virology*. 1981; 112: 400-10.

111. Sekellick MJ and Marcus PI. Persistent infection. I Interferon-inducing defective-interfering particles as mediators of cell sparing: possible role in persistent infection by vesicular stomatitis virus. *Virology*. 1978; 85: 175-86.

112. Schmaljohn C and Blair CD. Persistent infection of cultured mammalian cells by Japanese encephalitis virus. *J Virol*. 1977; 24: 580-9.

113. Spandidos DA and Graham AF. Generation of defective virus after infection of newborn rats with reovirus. *J Virol*. 1976; 20: 234-47.

114. Atkinson T, Barrett AD, Mackenzie A and Dimmock NJ. Persistence of virulent Semliki Forest virus in mouse brain following co-inoculation with defective interfering particles. *J Gen Virol*. 1986; 67 (Pt 6): 1189-94.

115. Palma EL and Huang A. Cyclic production of vesicular stomatitis virus caused by defective interfering particles. *J Infect Dis*. 1974; 129: 402-10.

116. Cave DR, Hendrickson FM and Huang AS. Defective interfering virus particles modulate virulence. *J Virol*. 1985; 55: 366-73.

117. Thompson KA and Yin J. Population dynamics of an RNA virus and its defective interfering particles in passage cultures. *Virol J*. 2010; 7: 257.

118. DePolo NJ, Giachetti C and Holland JJ. 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*. 1987; 61: 454-64.

119. Brinton MA and Fernandez AV. A replication-efficient mutant of West Nile virus is insensitive to DI particle interference. *Virology*. 1983; 129: 107-15.

120. Aaskov J, Buzacott K, Thu HM, Lowry K and Holmes EC. Long-term transmission of defective RNA viruses in humans and Aedes mosquitoes. *Science*. 2006; 311: 236-8.

121. Nüesch JP, de Chastonay J and Siegl G. Detection of defective genomes in hepatitis A virus particles present in clinical specimens. *J Gen Virol*. 1989; 70 (Pt 12): 3475-80.

122. Baczko K, Liebert UG, Billeter M, Cattaneo R, Budka H and ter Meulen V. Expression of defective measles virus genes in brain tissues of patients with subacute sclerosing panencephalitis. *J Virol*. 1986; 59: 472-8.

123. Li D, Lott WB, Lowry K, Jones A, Thu HM and Aaskov J. Defective interfering viral particles in acute dengue infections. *PLoS One*. 2011; 6: e19447.

124. Becher P, Orlich M, König M and Thiel HJ. Nonhomologous RNA recombination in bovine viral diarrhea virus: molecular characterization of a variety of subgenomic RNAs isolated during an outbreak of fatal mucosal disease. *J Virol*. 1999; 73: 5646-53.

125. Vasilijevic J, Zamarreño N, Oliveros JC, et al. Reduced accumulation of defective viral genomes contributes to severe outcome in influenza virus infected patients. *PLoS Pathog*. 2017; 13: e1006650.

126. Calain P, Monroe MC and Nichol ST. Ebola virus defective interfering particles and persistent infection. *Virology*. 1999; 262: 114-28.

127. Ismail ID and Milner JJ. Isolation of Defective Interfering Particles of Sonchus Yellow Net Virus from Chronically Infected Plants. *Journal of General Virology*. 1988; 69: 999-1006.

128. Rochon DM. Rapid de novo generation of defective interfering RNA by cucumber necrosis virus mutants that do not express the 20-kDa nonstructural protein. *Proc Natl Acad Sci U S A*. 1991; 88: 11153-7.

129. Goic B, Vodovar N, Mondotte JA, et al. RNA-mediated interference and reverse transcription control the persistence of RNA viruses in the insect model Drosophila. *Nat Immunol*. 2013; 14: 396-403.

130. Bernkopf H. Study of infectivity and hemagglutination of influenza virus in deembryonated eggs. *J Immunol*. 1950; 65: 571-83.

131. McLain L, Armstrong SJ and Dimmock NJ. One defective interfering particle per cell prevents influenza virus-mediated cytopathology: an efficient assay system. *J Gen Virol*. 1988; 69 (Pt 6): 1415-9.

132. Dimmock NJ and Kennedy SI. Prevention of death in Semliki Forest virus-infected mice by administration of defective-interfering Semliki Forest virus. *J Gen Virol*. 1978; 39: 231-42.

133. Thomson M, White CL and Dimmock NJ. The genomic sequence of defective interfering Semliki Forest virus (SFV) determines its ability to be replicated in mouse brain and to protect against a lethal SFV infection in vivo. *Virology*. 1998; 241: 215-23.

134. Dimmock NJ, Rainsford EW, Scott PD and Marriott AC. Influenza virus protecting RNA: an effective prophylactic and therapeutic antiviral. *J Virol*. 2008; 82: 8570-8.

135. Haas BJ, Gevers D, Earl AM, et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res*. 2011; 21: 494-504.

136. Simon-Loriere E and Holmes EC. Why do RNA viruses recombine? *Nat Rev Microbiol*. 2011; 9: 617-26.

137. Kirkegaard K and Baltimore D. The mechanism of RNA recombination in poliovirus. *Cell*. 1986; 47: 433-43.

138. Palmenberg AC, Spiro D, Kuzmickas R, et al. Sequencing and analyses of all known human rhinovirus genomes reveal structure and evolution. *Science*. 2009; 324: 55-9.

139. Worobey M, Rambaut A and Holmes EC. Widespread intra-serotype recombination in natural populations of dengue virus. *Proc Natl Acad Sci U S A*. 1999; 96: 7352-7.

140. Cherkasova EA, Korotkova EA, Yakovenko ML, et al. Long-term circulation of vaccine-derived poliovirus that causes paralytic disease. *J Virol*. 2002; 76: 6791-9.

141. Smith CM, Scott PD, O'Callaghan C, Easton AJ and Dimmock NJ. A Defective Interfering Influenza RNA Inhibits Infectious Influenza Virus Replication in Human Respiratory Tract Cells: A Potential New Human Antiviral. *Viruses*. 2016; 8.

142. Roux L, Simon AE and Holland JJ. Effects of defective interfering viruses on virus replication and pathogenesis in vitro and in vivo. *Adv Virus Res*. 1991; 40: 181-211.

143. Routh A, Ordoukhanian P and Johnson JE. Nucleotide-resolution profiling of RNA recombination in the encapsidated genome of a eukaryotic RNA virus by next-generation sequencing. *J Mol Biol*. 2012; 424: 257-69.

144. Nagy PD and Bujarski JJ. Engineering of homologous recombination hotspots with AU-rich sequences in brome mosaic virus. *J Virol*. 1997; 71: 3799-810.

145. Runckel C, Westesson O, Andino R and Derisi JL. Identification and manipulation of the molecular determinants influencing poliovirus recombination. *PLoS Pathog*. 2013; 9: e1003164.

146. Makino S, Yokomori K and Lai MM. Analysis of efficiently packaged defective interfering RNAs of murine coronavirus: localization of a possible RNA-packaging signal. *J Virol*. 1990; 64: 6045-53.

147. Finzi D, Plaeger SF and Dieffenbach CW. Defective virus drives human immunodeficiency virus infection, persistence, and pathogenesis. *Clin Vaccine Immunol*. 2006; 13: 715-21.

148. Cattaneo R, Schmid A, Eschle D, Baczko K, ter Meulen V and Billeter MA. Biased hypermutation and other genetic changes in defective measles viruses in human brain infections. *Cell*. 1988; 55: 255-65.

149. Sugiyama K, Suzuki K, Nakazawa T, et al. Genetic analysis of hepatitis C virus with defective genome and its infectivity in vitro. *J Virol*. 2009; 83: 6922-8.

150. Pesko KN, Fitzpatrick KA, Ryan EM, et al. Internally deleted WNV genomes isolated from exotic birds in New Mexico: function in cells, mosquitoes, and mice. *Virology*. 2012; 427: 10-7.

151. Scotti PD, Dearing S and Mossop DW. Flock House virus: a nodavirus isolated from Costelytra zealandica (White) (Coleoptera: Scarabaeidae). *Arch Virol*. 1983; 75: 181-9.

152. Dasgupta R, Free HM, Zietlow SL, et al. Replication of flock house virus in three genera of medically important insects. *J Med Entomol*. 2007; 44: 102-10.

153. Odegard A, Banerjee M and Johnson JE. Flock house virus: a model system for understanding non-enveloped virus entry and membrane penetration. *Curr Top Microbiol Immunol*. 2010; 343: 1-22.

154. Li H, Li WX and Ding SW. Induction and suppression of RNA silencing by an animal virus. *Science*. 2002; 296: 1319-21.

155. Ball LA and Li Y. cis-acting requirements for the replication of flock house virus RNA 2. *J Virol*. 1993; 67: 3544-51.

156. Routh A, Domitrovic T and Johnson JE. Host RNAs, including transposons, are encapsidated by a eukaryotic single-stranded RNA virus. *Proc Natl Acad Sci U S A*. 2012; 109: 1907-12.

157. Routh A, Head SR, Ordoukhanian P and Johnson JE. ClickSeq: Fragmentation-Free Next-Generation Sequencing via Click Ligation of Adaptors to Stochastically Terminated 3'-Azido cDNAs. *J Mol Biol*. 2015; 427: 2610-6.

158. Mikheyev AS and Tin MM. A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol Resour*. 2014; 14: 1097-102.

159. Greninger AL, Naccache SN, Federman S, et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med*. 2015; 7: 99.

160. Quick J, Loman NJ, Duraffour S, et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature*. 2016; 530: 228-32.

161. Wang J, Moore NE, Deng YM, Eccles DA and Hall RJ. MinION nanopore sequencing of an influenza genome. *Frontiers in microbiology*. 2015; 6: 766.

162. Garalde DR, Snell EA, Jachimowicz D, et al. Highly parallel direct RNA sequencing on an array of nanopores. *bioRxiv*. 2016.

163. Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B and Akeson M. Improved data analysis for the MinION nanopore sequencer. *Nature methods*. 2015; 12: 351-6.

164. Routh A, Chang MW, Okulicz JF, Johnson JE and Torbett BE. CoVaMa: Co-Variation Mapper for disequilibrium analysis of mutant loci in viral populations using next-generation sequence data. *Methods*. 2015.

165. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal*. 2011; 17: 10-2.

166. Langmead B, Trapnell C, Pop M and Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009; 10: R25.

167. Loman NJ and Quinlan AR. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics*. 2014; 30: 3399-401.

168. Milne I, Bayer M, Cardle L, et al. Tablet--next generation sequence assembly visualization. *Bioinformatics*. 2010; 26: 401-2.

169. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009; 25: 2078-9.

170. Lindenbach BD. Measuring HCV infectivity produced in cell culture and in vivo. *Methods Mol Biol*. 2009; 510: 329-36.

171. Selling BH and Rueckert RR. Plaque assay for black beetle virus. *J Virol*. 1984; 51: 251-3.

172. van de Waterbeemd M, Fort KL, Boll D, et al. High-fidelity mass analysis unveils heterogeneity in intact ribosomal particles. *Nat Methods*. 2017; 14: 283-6.

173. Johnson KL and Ball LA. Replication of flock house virus RNAs from primary transcripts made in cells by RNA polymerase II. *J Virol*. 1997; 71: 3323-7.

174. Routh A, Domitrovic T and Johnson JE. Packaging host RNAs in small RNA viruses: An inevitable consequence of an error-prone polymerase? *Cell Cycle*. 2012; 11.

175. Kircher M, Sawyer S and Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res*. 2012; 40: e3.

176. Laurence M, Hatzis C and Brash DE. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS One*. 2014; 9: e97876.

177. Kim YN, Lai MM and Makino S. Generation and selection of coronavirus defective interfering RNA with large open reading frame by RNA recombination and possible editing. *Virology*. 1993; 194: 244-53.

178. Dasgupta R, Cheng LL, Bartholomay LC and Christensen BM. Flock house virus replicates and expresses green fluorescent protein in mosquitoes. *J Gen Virol*. 2003; 84: 1789-97.

179. Lindenbach BD, Sgro JY and Ahlquist P. Long-distance base pairing in flock house virus RNA1 regulates subgenomic RNA3 synthesis and RNA2 replication. *J Virol*. 2002; 76: 3905-19.

180. Ball LA. Requirements for the self-directed replication of flock house virus RNA 1. *J Virol*. 1995; 69: 720-7.

181. Eckerle LD and Ball LA. Replication of the RNA segments of a bipartite viral genome is coordinated by a transactivating subgenomic RNA. *Virology*. 2002; 296: 165-76.

182. Albarino CG, Eckerle LD and Ball LA. The cis-acting replication signal at the 3' end of Flock House virus RNA2 is RNA3-dependent. *Virology*. 2003; 311: 181-91.

183. Albarino CG, Price BD, Eckerle LD and Ball LA. Characterization and template properties of RNA dimers generated during flock house virus RNA replication. *Virology*. 2001; 289: 269-82.

184. Jain M, Olsen HE, Paten B and Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol*. 2016; 17: 239.

185. Frensing T, Heldt FS, Pflugmacher A, et al. Continuous influenza virus production in cell culture shows a periodic accumulation of defective interfering particles. *PLoS One*. 2013; 8: e72288.

186. Stephens PJ, Greenman CD, Fu B, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*. 2011; 144: 27-40.

187. Short JR, Speir JA, Gopal R, Pankratz LM, Lanman J and Schneemann A. Role of Mitochondrial Membrane Spherules in Flock House Virus Replication. *J Virol*. 2016; 90: 3676-83.

188. Li Y and Ball LA. Nonhomologous RNA recombination during negative-strand synthesis of flock house virus RNA. *J Virol*. 1993; 67: 3854-60.

189. Salinas Y and Roux L. Replication and packaging properties of short Paramyxovirus defective RNAs. *Virus Res*. 2005; 109: 125-32.

190. Schneemann A. The structural and functional role of RNA in icosahedral virus assembly. *Annu Rev Microbiol*. 2006; 60: 51-67.

191. Fisher AJ and Johnson JE. Ordered duplex RNA controls capsid architecture in an icosahedral animal virus. *Nature*. 1993; 361: 176-9.

192. Dong XF, Natarajan P, Tihova M, Johnson JE and Schneemann A. Particle polymorphism caused by deletion of a peptide molecular switch in a quasiequivalent icosahedral virus. *J Virol*. 1998; 72: 6024-33.

193. Schneemann A and Marshall D. Specific encapsidation of nodavirus RNAs is mediated through the C terminus of capsid precursor protein alpha. *J Virol*. 1998; 72: 8738-46.

194. Tang L, Johnson KN, Ball LA, Lin T, Yeager M and Johnson JE. The structure of pariacoto virus reveals a dodecahedral cage of duplex RNA. *Nat Struct Biol*. 2001; 8: 77-83.

195. Tihova M, Dryden KA, Le TV, et al. Nodavirus coat protein imposes dodecahedral RNA structure independent of nucleotide sequence and length. *J Virol*. 2004; 78: 2897-905.

196. Schneider CA, Rasband WS and Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods*. 2012; 9: 671-5.

197. Fort KL, van de Waterbeemd M, Boll D, et al. Expanding the structural analysis capabilities on an Orbitrap-based mass spectrometer for large macromolecular complexes. *Analyst*. 2017; 143: 100-5.

198. Hesketh EL, Saunders K, Fisher C, et al. The 3.3 Å structure of a plant geminivirus using cryo-EM. *Nature Communications*. 2018; 9: 2369.

199. Fernandez-Leiro R and Scheres SHW. A pipeline approach to single-particle processing in RELION. *Acta Crystallogr D Struct Biol*. 2017; 73: 496-502.

200. Zheng SQ, Palovcak E, Armache JP, Verba KA, Cheng Y and Agard DA. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat Methods*. 2017; 14: 331-2.

201. Zhang K. Gctf: Real-time CTF determination and correction. *J Struct Biol*. 2016; 193: 1-12.

202. Scheres SH. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol*. 2012; 180: 519-30.

203. Urnavicius L, Zhang K, Diamant AG, et al. The structure of the dynactin complex and its interaction with dynein. *Science*. 2015; 347: 1441-6.

204. Scheres SH and Chen S. Prevention of overfitting in cryo-EM structure determination. *Nat Methods*. 2012; 9: 853-4.

205. Leney AC and Heck AJ. Native Mass Spectrometry: What is in the Name? *J Am Soc Mass Spectrom*. 2017; 28: 5-13.

206. Venter PA, Marshall D and Schneemann A. Dual roles for an arginine-rich motif in specific genome recognition and localization of viral coat protein to RNA replication sites in flock house virus-infected cells. *J Virol*. 2009; 83: 2872-82.

207. Head SR, Komori HK, LaMere SA, et al. Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*. 2014; 56: 61-4, 6, 8, passim.

208. van Dijk EL, Jaszczyszyn Y and Thermes C. Library preparation methods for next-generation sequencing: tone down the bias. *Exp Cell Res*. 2014; 322: 12-20.

209. Birts CN, Sanzone AP, El-Sagheer AH, Blaydes JP, Brown T and Tavassoli A. Transcription of click-linked DNA in human cells. *Angew Chem Int Ed Engl*. 2014; 53: 2362-5.

210. Chen X, El-Sagheer AH and Brown T. Reverse transcription through a bulky triazole linkage in RNA: implications for RNA sequencing. *Chem Commun (Camb)*. 2014; 50: 7597-600.

211. Dallmann A, El-Sagheer AH, Dehmel L, et al. Structure and dynamics of triazole-linked DNA: biocompatibility explained. *Chemistry*. 2011; 17: 14714-7.

212. El-Sagheer AH and Brown T. Efficient RNA synthesis by in vitro transcription of a triazole-modified DNA template. *Chem Commun (Camb)*. 2011; 47: 12057-8.

213. El-Sagheer AH, Sanzone AP, Gao R, Tavassoli A and Brown T. Biocompatible artificial DNA linker that is read through by DNA polymerases and is functional in Escherichia coli. *Proc Natl Acad Sci U S A*. 2011; 108: 11338-43.

214. Fujino T, Yasumoto K, Yamazaki N, Hasome A, Sogawa K and Isobe H. Triazole-linked DNA as a primer surrogate in the synthesis of first-strand cDNA. *Chem Asian J*. 2011; 6: 2956-60.

215. Isobe H and Fujino T. Triazole-linked analogues of DNA and RNA ((TL)DNA and (TL)RNA): synthesis and functions. *Chem Rec*. 2014; 14: 41-51.

216. Isobe H, Fujino T, Yamazaki N, Guillot-Nieckowski M and Nakamura E. Triazole-linked analogue of deoxyribonucleic acid ((TL)DNA): design, synthesis, and double-strand formation with natural DNA. *Org Lett*. 2008; 10: 3729-32.

217. Qiu J, El-Sagheer AH and Brown T. Solid phase click ligation for the synthesis of very long oligonucleotides. *Chem Commun (Camb)*. 2013; 49: 6959-61.

218. Sanzone AP, El-Sagheer AH, Brown T and Tavassoli A. Assessing the biocompatibility of click-linked DNA in Escherichia coli. *Nucleic Acids Res*. 2012; 40: 10567-75.

219. Kolb HC, Finn MG and Sharpless KB. Click Chemistry: Diverse Chemical Function from a Few Good Reactions. *Angew Chem Int Ed Engl*. 2001; 40: 2004-21.

220. Baskin JM, Prescher JA, Laughlin ST, et al. Copper-free click chemistry for dynamic in vivo imaging. *Proc Natl Acad Sci U S A*. 2007; 104: 16793-7.

221. El-Sagheer AH and Brown T. Synthesis and polymerase chain reaction amplification of DNA strands containing an unnatural triazole linkage. *J Am Chem Soc*. 2009; 131: 3958-64.

222. Görzer I, Guelly C, Trajanoski S and Puchhammer-Stöckl E. The impact of PCR-generated recombination on diversity estimation of mixed viral populations by deep sequencing. *J Virol Methods*. 2010; 169: 248-52.

223. Meyerhans A, Vartanian JP and Wain-Hobson S. DNA recombination during PCR. *Nucleic Acids Res*. 1990; 18: 1687-91.

224. Hong V, Presolski SI, Ma C and Finn MG. Analysis and optimization of copper-catalyzed azide-alkyne cycloaddition for bioconjugation. *Angew Chem Int Ed Engl*. 2009; 48: 9879-83.

225. Abel GR, Calabrese ZA, Ayco J, Hein JE and Ye T. Measuring and Suppressing the Oxidative Damage to DNA During Cu(I)-Catalyzed Azide-Alkyne Cycloaddition. *Bioconjug Chem*. 2016; 27: 698-704.

226. Litovchick A, Dumelin CE, Habeshian S, et al. Encoded Library Synthesis Using Chemical Ligation and the Discovery of sEH Inhibitors from a 334-Million Member Library. *Sci Rep*. 2015; 5: 10916.

227. DeAngelis MM, Wang DG and Hawkins TL. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res*. 1995; 23: 4742-3.

228. Faircloth B and Glenn T. Speedbeads (AKA Serapure). 2016.

229. Rohland N and Reich D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res*. 2012; 22: 939-46.

230. Proudfoot NJ. Ending the message: poly(A) signals then and now. *Genes Dev*. 2011; 25: 1770-82.

231. Kempf BJ and Barton DJ. Picornavirus RNA polyadenylation by 3D(pol), the viral RNA-dependent RNA polymerase. *Virus Res*. 2015; 206: 3-11.

232. Poon LL, Pritlove DC, Fodor E and Brownlee GG. Direct evidence that the poly(A) tail of influenza A virus mRNA is synthesized by reiterative copying of a U track in the virion RNA template. *J Virol*. 1999; 73: 3473-6.

233. Wilusz J. Putting an 'End' to HIV mRNAs: capping and polyadenylation as potential therapeutic targets. *AIDS Res Ther*. 2013; 10: 31.

234. Sheets MD, Ogg SC and Wickens MP. Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res*. 1990; 18: 5799-805.

235. Chen F, MacDonald CC and Wilusz J. Cleavage site determinants in the mammalian polyadenylation signal. *Nucleic Acids Res*. 1995; 23: 2614-20.

236. Gil A and Proudfoot NJ. A sequence downstream of AAUAAA is required for rabbit beta-globin mRNA 3'-end formation. *Nature*. 1984; 312: 473-4.

237. Tian B and Manley JL. Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol*. 2017; 18: 18-30.

238. Di Giammartino DC, Nishida K and Manley JL. Mechanisms and consequences of alternative polyadenylation. *Mol Cell*. 2011; 43: 853-66.

239. Lianoglou S, Garg V, Yang JL, Leslie CS and Mayr C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev*. 2013; 27: 2380-96.

240. Ji Z, Lee JY, Pan Z, Jiang B and Tian B. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc Natl Acad Sci U S A*. 2009; 106: 7028-33.

241. Hollerer I, Curk T, Haase B, et al. The differential expression of alternatively polyadenylated transcripts is a common stress-induced response mechanism that modulates mammalian mRNA expression in a quantitative and qualitative fashion. *RNA*. 2016; 22: 1441-53.

242. Szkop KJ and Nobeli I. Untranslated Parts of Genes Interpreted: Making Heads or Tails of High-Throughput Transcriptomic Data via Computational Methods: Computational methods to discover and quantify isoforms with alternative untranslated regions. *Bioessays*. 2017; 39.

243. Kim D, Langmead B and Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 2015; 12: 357-60.

244. Elrod ND, Jaworski EA, Ji P, Wagner EJ and Routh A. Development of Poly(A)-ClickSeq as a tool enabling simultaneous genome-wide poly(A)-site identification and differential expression analysis. *Methods*. 2019; 155: 20-9.

245. Routh A, Ji P, Jaworski E, Xia Z, Li W and Wagner EJ. Poly(A)-ClickSeq: click-chemistry for next-generation 3′-end sequencing without RNA enrichment or fragmentation. *Nucleic Acids Res*. 2017; 45: e112.

246. Fuda NJ, Ardehali MB and Lis JT. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*. 2009; 461: 186-92.

247. Mayer A, Landry HM and Churchman LS. Pause & go: from the discovery of RNA polymerase pausing to its functional implications. *Curr Opin Cell Biol*. 2017; 46: 72-80.

248. Landick R. The regulatory roles and mechanism of transcriptional pausing. *Biochem Soc Trans*. 2006; 34: 1062-6.

249. Sims RJ, Belotserkovskaya R and Reinberg D. Elongation by RNA polymerase II: the short and long of it. *Genes Dev*. 2004; 18: 2437-68.

250. Liu X, Kraus WL and Bai X. Ready, pause, go: regulation of RNA polymerase II pausing and release by cellular signaling pathways. *Trends Biochem Sci*. 2015; 40: 516-25.

251. Core LJ, Waterfall JJ and Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*. 2008; 322: 1845-8.

252. Kwak H, Fuda NJ, Core LJ and Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*. 2013; 339: 950-3.

253. Tani H, Mizutani R, Salam KA, et al. Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res*. 2012; 22: 947-56.

254. Paulsen MT, Veloso A, Prasad J, et al. Coordinated regulation of synthesis and stability of RNA during the acute TNF-induced proinflammatory response. *Proc Natl Acad Sci U S A*. 2013; 110: 2240-5.

255. Veloso A, Kirkconnell KS, Magnuson B, et al. Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res*. 2014; 24: 896-905.

256. Fuchs G, Voichek Y, Benjamin S, Gilad S, Amit I and Oren M. 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. *Genome Biol*. 2014; 15: R69.

257. Churchman LS and Weissman JS. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*. 2011; 469: 368-73.

258. Nechaev S, Fargo DC, dos Santos G, Liu L, Gao Y and Adelman K. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science*. 2010; 327: 335-8.

259. Weber CM, Ramachandran S and Henikoff S. Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Mol Cell*. 2014; 53: 819-30.

260. Churchman LS and Weissman JS. Native elongating transcript sequencing (NET-seq). *Curr Protoc Mol Biol*. 2012; Chapter 4: Unit 4.14.1-7.

261. Pfeffer S, Lagos-Quintana M and Tuschl T. Cloning of small RNA molecules. *Curr Protoc Mol Biol*. 2005; Chapter 26: Unit 26.4.

262. Peck KM and Lauring AS. Complexities of Viral Mutation Rates. *J Virol*. 2018; 92.

263. Barrows NJ, Campos RK, Powell ST, et al. A Screen of FDA-Approved Drugs for Inhibitors of Zika Virus Infection. *Cell Host Microbe*. 2016; 20: 259-70.

264. Marshall D and Schneemann A. Specific packaging of nodaviral RNA2 requires the N-terminus of the capsid protein. *Virology*. 2001; 285: 165-75.

265. Venter PA, Krishna NK and Schneemann A. Capsid protein synthesis from replicating RNA directs specific packaging of the genome of a multipartite, positive-strand RNA virus. *J Virol*. 2005; 79: 6239-48.

266. Kennedy JC and Macdonald RD. Persistent infection with infectious pancreatic necrosis virus mediated by defective-interfering (DI) virus particles in a cell line showing strong interference but little DI replication. *J Gen Virol*. 1982; 58: 361-71.

267. Ehresmann DW and Schaffer FL. RNA synthesized in calicivirus-infected cells is atypical of picornaviruses. *J Virol*. 1977; 22: 572-6.

268. Cole CN, Smoler D, Wimmer E and Baltimore D. Defective interfering particles of poliovirus. I. Isolation and physical properties. *J Virol*. 1971; 7: 478-85.

269. McClure MA, Holland JJ and Perrault J. Generation of defective interfering particles in picornaviruses. *Virology*. 1980; 100: 408-18.

270. Ansardi DC, Porter DC and Morrow CD. Complementation of a poliovirus defective genome by a recombinant vaccinia virus which provides poliovirus P1 capsid precursor in trans. *J Virol*. 1993; 67: 3684-90.

271. Muslin C, Joffret ML, Pelletier I, Blondel B and Delpeyroux F. Evolution and Emergence of Enteroviruses through Intra- and Inter-species Recombination: Plasticity and Phenotypic Impact of Modular Genetic Exchanges in the 5' Untranslated Region. *PLoS Pathog*. 2015; 11: e1005266.

272. Hillman BI, Carrington JC and Morris TJ. A defective interfering RNA that contains a mosaic of a plant virus genome. *Cell*. 1987; 51: 427-33.

273. Rubino L, Burgyan J, Grieco F and Russo M. Sequence analysis of cymbidium ringspot virus satellite and defective interfering RNAs. *J Gen Virol*. 1990; 71 (Pt 8): 1655-60.

274. Damayanti TA, Nagano H, Mise K, Furusawa I and Okuno T. Brome mosaic virus defective RNAs generated during infection of barley plants. *J Gen Virol*. 1999; 80 (Pt 9): 2511-8.

275. Li M and Stoneking M. A new approach for detecting low-level mutations in next-generation sequence data. *Genome Biol*. 2012; 13: R34.

276. Brinton MA. Characterization of West Nile virus persistent infections in genetically resistant and susceptible mouse cells. I. Generation of defective nonplaquing virus particles. *Virology*. 1982; 116: 84-98.

277. Lancaster MU, Hodgetts SI, Mackenzie JS and Urosevic N. Characterization of defective viral RNA produced during persistent infection of Vero cells with Murray Valley encephalitis virus. *J Virol*. 1998; 72: 2474-82.

278. Fuller FJ and Marcus PI. Interferon induction by viruses. Sindbis virus: defective-interfering particles temperature-sensitive for interferon induction. *J Gen Virol*. 1980; 48: 391-4.

279. Frey TK and Hemphill ML. Generation of defective-interfering particles by rubella virus in Vero cells. *Virology*. 1988; 164: 22-9.

280. Rouzine IM and Weinberger LS. Design requirements for interfering particles to maintain coadaptive stability with HIV-1. *J Virol*. 2013; 87: 2081-93.

281. Chattopadhyay SK, Morse HC, Makino M, Ruscetti SK and Hartley JW. Defective virus is associated with induction of murine retrovirus-induced immunodeficiency syndrome. *Proc Natl Acad Sci U S A*. 1989; 86: 3862-6.

282. Aziz DC, Hanna Z and Jolicoeur P. Severe immunodeficiency disease induced by a defective murine leukaemia virus. *Nature*. 1989; 338: 505-8.

283. Voynow SL and Coffin JM. Truncated gag-related proteins are produced by large deletion mutants of Rous sarcoma virus and form virus particles. *J Virol*. 1985; 55: 79-85.

284. Bar-Joseph M and Mawassi M. The defective RNAs of Closteroviridae. *Front Microbiol*. 2013; 4: 132.

285. Hernandez C, Carette JE, Brown DJ and Bol JF. Serial passage of tobacco rattle virus under different selection conditions results in deletion of structural and nonstructural genes in RNA 2. *J Virol.* 1996; 70: 4933-40.

286. Makino S, Taguchi F and Fujiwara K. Defective interfering particles of mouse hepatitis virus. *Virology*. 1984; 133: 9-17.

287. Sabir JS, Lam TT, Ahmed MM, et al. Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. *Science*. 2016; 351: 81-4.

288. Dalton K, Casais R, Shaw K, et al. cis-acting sequences required for coronavirus infectious bronchitis virus defective-RNA replication and packaging. *J Virol*. 2001; 75: 125-33.

289. Hackett AJ. A possible morphologic basis for the autointerference phenomenon in vesicular stomatitis virus. *Virology*. 1964; 24: 51-9.

290. Wiktor TJ, Dietzschold B, Leamnson RN and Koprowski H. Induction and biological properties of defective interfering particles of rabies virus. *J Virol*. 1977; 21: 626-35.

291. Conzelmann KK, Cox JH and Thiel HJ. An L (polymerase)-deficient rabies virus defective interfering particle RNA is replicated and transcribed by heterologous helper virus L proteins. *Virology*. 1991; 184: 655-63.

292. Obijeski JF, McCauley J and Skehel JJ. Nucleotide sequences at the terminal of La Crosse virus RNAs. *Nucleic Acids Res*. 1980; 8: 2431-8.

293. Leppert M, Kort L and Kolakofsky D. Further characterization of Sendai virus DI-RNAs: a model for their generation. *Cell*. 1977; 12: 539-52.

294. Hall WW, Martin SJ and Gould E. Defective interfering particles produced during the replication of measles virus. *Med Microbiol Immunol*. 1974; 160: 155-64.

295. Pfaller CK, Mastorakos GM, Matchett WE, Ma X, Samuel CE and Cattaneo R. 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*. 2015; 89: 7735-47.

296. Cantell K. Mumps Virus*. In: Kenneth MSaMAL, (ed.). *Advances in Virus Research*. Academic Press, 1962, p. 123-64.

297. Welsh RM and Pfau CJ. Determinants of lymphocytic choriomeningitis interference. *J Gen Virol*. 1972; 14: 177-87.

298. Jacobson S and Pfau CJ. Viral pathogenesis and resistance to defective interfering particles. *Nature*. 1980; 283: 311-3.

299. Baltimore D. Expression of animal virus genomes. *Bacteriol Rev*. 1971; 35: 235-41.

VITA

Elizabeth Jaworski was born April 17th, 1991 to Thomasz and Agnieszka Jaworski. She attended Atlantic Community High School in Delray Beach Florida where she received an International Baccalaureate degree in 2009. She went on to attend the University of Florida. This is where her passion for science and research flourished as she worked in multiple research labs, co-authoring a total of three research publications. She graduated from the University of Florida in 2013 with a bachelor's degree in biology and a minor in entomology. Elizabeth went on to matriculate at the University of Texas Medical Branch in the fall of 2013. She joined the laboratory of Dr. Muge Kuyumcu-Martinez where she studied the role of alternative splicing in diabetic skeletal muscle, resulting in three publications. In February of 2016 she joined Dr. Andrew Routh to elucidate the mechanisms of defective interfering virus particle production. There she authored four publications and one methods book chapter. Her work on Next Generation Sequencing methodology development resulted in one patent. In June of 2018, she and Dr. Routh co-founded ClickSeq Technologies, LLC, a Next Generation Sequencing service company based upon the developed technology.

Education

University of Texas Medical Branch — Aug. 2013- Aug. 2018
Degree: Doctor of Philosophy
Department: Biochemistry and Molecular Biology

University of Florida — Aug. 2009- May 2013
Degree: Bachelor of Science
Major: Biology-Biotechnology
Minor: Entomology

Atlantic Community High School — Aug. 2005- May 2009
International Baccalaureate Program

Professional Experience

Co-Founder and CEO — June 2018 -
ClickSeq Technologies, LLC

Graduate Research Assistant — Mar. 2016 - July 2018
Biochemistry and Molecular Biology Department, UTMB
Dr. Andrew Routh

Graduate Research Assistant — Jan. 2014 - Mar. 2016
Biochemistry and Molecular Biology Department, UTMB
Dr. Muge Kuyumcu-Martinez

Graduate Research Assistant — Oct. 2013 - Dec. 2013
Neuroscience and Cell Biology Department, UTMB
Dr. Rakez Kayed

Research Assistant, — Mar. 2012 - July 2013
Floriculture Biotechnology and Genetics Laboratory
Environmental Horticulture Department, UF
Dr. David G. Clark and Dr. Thomas A. Colquhoun

Research Assistant, — May 2010 - Jan. 2012
Urban Entomology Laboratory
Entomology and Nematology Department, UF
Dr. Phillip Koehler

Teaching and Mentoring

Bench Tutorials Mentor Award: Best Mentor	UTMB/ Ball HS	Sept. 2017 - May. 2018
Personal Tutor	Varsity Tutors	Nov. 2015 - July. 2018
Bench Tutorials Mentor Award: Best Team Science	UTMB/ Ball HS	Sept. 2016 - May 2017
Small Group Facilitator Course: Biochemistry	UTMB	Sept. 2015 - Dec. 2015

Mentoring Rotating Medical Student	UTMB	May 2015 - July 2015
Bench Tutorials Mentor Award: Best Team Science	UTMB/ Ball HS	Sept. 2014- July 2015
Mentoring Rotating Medical Student	UTMB	May 2014- July 2014

Publications

Elrod, N.*, **Jaworski, E.A.***, Ji, P., Wagner, E., Routh, A., (2019) Development of Poly(A)-ClickSeq as a Tool Enabling Simultaneous Genome-wide Poly(A)-site identification and Differential Expression Analysis. *Methods*, Jan. 2019, doi: 10.1016/j.ymeth.2019.01.002 *Co-first authors

Jaworski, E.A., Routh, A. (2017) Parallel ClickSeq and Nanopore Sequencing elucidates the rapid evolution of Defective-Interfering RNAs in Flock House virus. *PLoS Pathogens,* May 2017, 13(5):e1006365

Routh, A., Ji, P., **Jaworski, E.A.**, Xia, Z., Li, W., Wagner, E. (2017) Poly(A)-ClickSeq: click-chemistry for next-generation 3'-end sequencing without RNA enrichment or fragmentation. *Nucleic Acids Research,* Apr. 2017, doi: 10.1093/nar/gkx286

Nutter C.A.*, **Jaworski E.A.***, Verma S.K., Perez-Carrsco Y., Kuyumcu-Martinez N.M. (2017) Developmentally regulated alternative splicing is perturbed in Type 1 diabetic skeletal muscle. *Muscle and Nerve,* Oct. 2017, doi:10.1002/mus.25599 *Co-first authors

Verma S.K., Deshmukh V., Nutter, C.A., **Jaworski, E.A.**, Jin, W., Wadhwa L., Abata, A., Ricci M., Lincoln J., Martin J.F., Yeo, G., Kuyumcu-Martinez M.N. (2016) Rbfox2 function in RNA metabolism is impaired in hypoplastic left heart syndrome patient hearts. *Scientific Reports,* Aug. 2016, 6(30896)

Nutter C.A., **Jaworski E.A.**, Verma S.K., Deshmukh V., Wang Q., Botvinnik O.B., Lozano M.J., Abass I.J., Ijaz T., Brasier A.R., Garg N.J., Wehrens, X.H.T, Yeo G.W., Kuyumcu-Martinez N.M. (2016) Dysregulation of RBFOX2 is an early event in cardiac pathogenesis of diabetes. *Cell Reports,* June 2016, 15(10), 2200-2213.

Langer K.M., Jones C.R., **Jaworski E.A.**, Rushing G.V., Kim J.Y., Clark, D.G., and Colquhoun T.A. (2014). PhDAHP1 is required for floral volatile benzenoid/phenylpropanoid biosynthesis in Petunia x hybrida cv 'Mitchell Diploid'. *Phytochemistry,* May 2014, 103, 22-31.

Schwieterman M.L., Colquhoun T.A., **Jaworski E.A**., Bartoshuk L.M., Gilbert J.L., et al. (2014) Strawberry Flavor: Diverse Chemical Compositions, a Seasonal Influence, and Effects on Sensory Perception. *PLoS ONE,* Feb. 2014, 9(2): e88446

Colquhoun, T.A., Schwieterman, M.L., Gilbert, J.L., **Jaworski, E.A.**, Langer, K.M., Jones, C.R., Rushing, G., Clark, D.G., and Folta, K.M. (2013). Light Modulation of Plant Flavor and Aroma Compounds in Select Fruits and Flowers. *Postharvest Biology Technology,* June 2013, 86, 37-44.

Book Chapters

Jaworski E.A., Routh, A. (2018) ClickSeq: Replacing fragmentation and enzymatic ligation with click-chemistry to prevent sequence chimeras. *Methods in Molecular Biology* Dec. 2017, 1712:71-85.

Patents

Poly(A)-ClickSeq: Click Chemistry for Next Generation 3' End Sequencing without RNA Enrichment or Fragmentation. Inventors: Routh, A., Wagner, E., **Jaworski, E.**, Ji, P. Patent: US 16/282,159. Filed: February 22nd, 2019

Permanent address: 816 Walker St., La Marque, TX, 77568

This dissertation was typed by Elizabeth A. Jaworski.