HIGH THROUGHPUT GENETICS AND CHARACTERIZATION OF AN RNA ARBOVIRUS, SINDBIS VIRUS, USING ACCURATE NEXT-GENERATION SEQUENCING OF VIRAL EVOLUTION AND RNA ENRICHMENT.

A Dissertation
Presented to the Faculty of the Weill Cornell Graduate School of Medical Sciences
In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Pradeep Morris Ambrose
May 2020
HIGH THROUGHPUT GENETICS AND CHARACTERIZATION OF AN RNA ARBOVIRUS, SINDBIS VIRUS, USING ACCURATE NEXT-GENERATION SEQUENCING OF VIRAL EVOLUTION AND RNA ENRICHMENT.

Pradeep Ambrose, Ph. D.
Cornell University 2020

The goal of these studies was to investigate Sindbis virus adaptation to various infection bottlenecks and utilize the dynamics of minor variants to study viral genetics in a high-throughput manner. During infection an RNA virus exists as a tremendously diverse population, and this genetic diversity underlies their ability to rapidly adapt to new conditions and cause disease. Since viruses evolve as populations, our understanding of viral evolution has historically been limited by the inability to characterize populations. Whilst new sequencing technologies provide sufficient depth to sequence full viral populations, their intrinsic base-calling error rate combined with mutations introduced during sample processing makes viral mutations and sequencing errors indistinguishable. Utilizing rolling reverse transcription, the novel CirSeq technique virtually eliminates sequencing errors by bioinformatically parsing tandem generated repeats, and for the first time allows a highly accurate mutational landscape profile of the whole viral population. In addition, a novel hybridization capture technique we developed allows us to maximize the sequencing coverage of desired viral RNA molecules. We used these technologies to map the mutational distribution of RNA virus populations and perform genetics are previously unseen scales. We sequenced serial
passages of the well-characterized Sindbis virus to yield novel information on genetic features crucial for viral replication. We analyzed how the starting \textit{in vitro} transcribed RNA population adapts to various bottlenecks encountered during electroporation and subsequent passaging, and during packaging and egress. Then we compared these data to previous studies of critical genome sites and expanded our study to new sites of interest. We posit that such unbiased high-throughput genetics pushes the envelope beyond the previous limits on discovery of viral functional elements. These techniques can be used to further characterize clinically relevant RNA viruses that are agents of current and recent epidemics, such as SARS-CoV-2 coronavirus, chikungunya virus, Zika virus, eastern equine encephalitis virus, dengue virus, and West Nile virus.
BIOGRAPHICAL SKETCH

Pradeep M. Ambrose is a member of the *Homo sapiens sapiens* subspecies who occasionally enjoys interactions with members of his own species and often enjoys observing the behavior of members of other species. He read a few short books and articles on viruses in his teenage years, for better or worse, and has "thought they were so cool" since. His time as a virologist-in-training during graduate school further confirmed these suspicions.
To my Grandparents;
Morris and Hannah Kasthuri Mathias, and Jebamoni and Victoria Ambrose.
ACKNOWLEDGMENTS

I would like to thank Charlie Rice, Peggy MacDonald, and all members of the Rice lab, Chris Mason, and all members of the Mason lab, Andrea Branch, and all members of the Branch lab - in particular, Arielle Klepper, who introduced me to the field of scientific research - and Basil Hanss and all members of the Hanss lab. I express my gratitude to members of my committee past and present; Jeremy Dittman, Scott Blanchard, Xin-Yun Huang, Randy Longman, and Jason Mezey. I thank Audrey Rivera, Randi Silver, Jason Banfelder, Alessio Accardi, Emre Aksay, Harel Weinstein, and all members of the Physiology, Biophysics and Systems Biology department family, it's been a delight and privilege to have been a part of. I thank Augustine Choi and everyone at the graduate school for their efforts toward all the behind-the-scenes stuff that needs to happen so I could just focus on classes and research without worry. I thank Hattie, Bruno and Lily for demanding attention and tummy-rubs without any care whatsoever as to what I was in the middle of, and I will cherish the time spent at my wall-adjacent bench in Room 507 in Charlie's lab and nearby. I also express my gratitude to my friends and family. Mohsan Saeed, Bill Schneider, Inna Ricardo-Lax and Mariel Bartley generously provided time and effort in the editing of this thesis dissertation.

Many, many people have impacted my life to enable me to be where I am and do the things I have been able to. A couple pages in a document's front matter cannot adequately convey the gratitude that I have; several of these people and their impacts are described throughout this thesis dissertation, and I hope I have done some justice to their grace and kindness. As such I have felt that it is rather disingenuous for me to use the pronoun "I"
when describing successes, at any scale, as they were all collaborative endeavors, and I have tried to consciously minimize such use. I use the pronoun freely, however, when describing relative failures.

The New York Yankees just had several magical seasons, rivaling the excitement of the late 1990's team, that were only cut short by teams which were later discovered to be engaged in institutionalized cheating. During a podcast hosted by veteran pitcher C.C. Sabathia and sportscaster Ryan Ruocco, the current manager of the team, Aaron Boone, credited this success to their 'clubhouse environment'. They all explained further, "no one is arguing about needing the underlying talent, but having a good clubhouse environment is what makes you want to be at your best every single day". They contrasted this with their own experiences under poor managers and in suboptimal environments, where "you still try, but you end up just grinding through the day and not caring all that much". That, for them, is the primary determinant of success.

The Rice Lab, Mason Lab, and the Physiology, Biophysics and Systems Biology department at Weill Cornell Medicine, all of which I am but a small part, have created a wonderful 'science-house' environment that has made me happy to try to be at my personal best every day, for as far as I can remember. Every great year was followed by an even better and happier one, in which my curiosity of the biological world only grew. During my time here so far, I cannot think of examples of being in a suboptimal environment and just grinding through the day that I can share on a podcast to be relatable. For that I am ever so lucky, and I would be the luckiest person alive if I can continue to have such experiences, even if only periodically, throughout the rest of my life.
# TABLE OF CONTENTS

Biographical sketch ........................................................................................................ iii
Dedication.......................................................................................................................... iv
Acknowledgements ......................................................................................................... v
Table of contents ............................................................................................................ vii
List of figures ................................................................................................................ xi
List of tables .................................................................................................................. xxi
List of abbreviations ..................................................................................................... xxiv
Preface ............................................................................................................................ xxvii

## Chapter I: Introduction: RNA Virus Evolution and Sindbis Virus Genetics

1.1. RNA Virus Evolution ................................................................................................. 1
1.2. Significance ................................................................................................................ 10
1.3. Discovery of Sindbis Virus ....................................................................................... 11
1.4. Inception of Sindbis Genetics Using Temperature-Sensitive Mutants .......... 27
1.5 Discovery of Components of Sindbis Virus Virions ......................................... 38
1.6. Identification of Sindbis Viral Components During Intracellular Infection .......... 44
1.7 Sequencing of Sindbis virus and ts Mutant Lesions and Elucidation of Nonstructural Protein Functions ................................................................. 54
1.8. Discovery of Viral Translational and Transcriptional Control ....................... 66
1.9. Discovery of Viral RNA Sequence Elements ....................................................... 94
1.10. Further Discovery Using Phenotypic Screens and Concluding Remarks ........ 121
1.11. Chapter I References ............................................................................................ 126
Chapter II: Methods: How to Purify a Virus

2.1. Introduction ........................................................................................................................................ 144

2.2. Quantitative PCR Assays for Sindbis virus, Zika virus, and Human Cellular RNA ........................................................................................................................................ 151

2.3. Introduction to Oligonucleotide Based Hybridization Capture ........................................ 158

2.4. Extraction of RNA from Infected Cells: A Short Protocol .................................................. 159

2.5. Enrichment of Target RNA from Total RNA: A Short Protocol ................................ 164

2.6. General Principles of Oligonucleotide Based Hybridization Capture ..... 173

2.7. Hybridization Oligonucleotide Probe Design for Sindbis Genomic Viral RNA ........................................................................................................................................ 182

2.8. Water and Low Heat Effectively Elutes Target RNA ....................................................... 190

2.9. Efficacy of Direct vs Indirect Binding and Identification of Interference During Ethanol Precipitation ........................................................................................................................................ 203

2.10. Solving Interference During Ethanol Precipitation ................................................................ 213

2.11. Importance of DNA:RNA Molar Ratio and Cost Comparison ........................................ 222

2.12. The Biotinylated Probe is Selective but the Streptavidin Beads Non-Specifically Bind Input RNA ........................................................................................................................................ 231

2.13. Vortexing Selectively Removes Non-Specifically Bound RNA ...................................... 238

2.14. Two Rounds of Capture Yields Final Pool of 95% Sindbis Viral RNA ........................................................................................................................................ 247

2.15. Total RNA Input into Reaction Affects Percentage Viral RNA Captured ........................................................................................................................................ 263

2.16. Lessons from Zika: Effect of Mismatches and Annealing Conditions on Binding ........................................................................................................................................ 268

2.17. The Necessity of a Combined Phenol:Chloroform and
LIST OF FIGURES

Chapter One.

1.1. Evolution of Viral Populations................................................................. 1
1.2. Fitness Effect Distribution of Random Mutations................................. 4
1.3. Polymerase and Detection Error Rates and Error Rate Correction Method (CirSeq)......................................................................................... 6
1.4. The Sindbis Life Cycle and Genome Composition................................. 8
1.5. Field Virus Discovery in 1950s................................................................. 16
1.6. Map of Seroprevalence Survey and Discovery of Sindbis Virus and Other Agents of Recent Epidemics................................................................. 25
1.7. Temperature Dependence of Sindbis Virus Strains................................. 29
1.8. Heat Stability and Complementation of Sindbis ts Mutants.................. 33
1.9. Electron Microscopy of Sindbis Virus..................................................... 39
1.10. Resolution of the Structural Proteins of the Sindbis Virion................ 41
1.11. Intracellular Alphavirus RNAs and Structural Proteins....................... 46
1.12. Intracellular Alphavirus Double-Stranded RNAs................................. 50
1.13. Translation of Nonstructural Proteins from Alphavirus Genomic RNA........................................................................................................ 53
1.15. Map of nsP2 ts Mutants and Active Site Residues............................... 63
1.16. Translational Control in Alphaviruses Using Opal Stop Codon.............. 68
1.17. Identity of Nonstructural Protein Cleavage Sites................................. 72
1.18. Growth and Translation Characteristics of Nonstructural Protein Cleavage Mutants .......................... 74
1.19. Effect of nsP Cleavage Mutations on Ratio of Genomic and Subgenomic RNA During Infection ............................................. 78
1.20. Effect of nsP Cleavage Mutants on Minus Strand Shutoff ......................................................................................... 84
1.21. Effect of nsP Cleavage Mutants on Double Stranded Viral RNA Accumulation ................................................................. 87
1.22. Summary of Sindbis Nonstructural Protein Cleavage Mechanism .......................................................................................... 90
1.23. Alphavirus Conserved RNA Elements Identified Using Multiple Sequence Alignments ................................................................. 97
1.24. Weird Plaque Phenotypes of Sindbis Virus 51 Nucleotide CSE Mutants ...................................................................................... 104
1.25. RNA SHAPE Secondary Structure of Sindbis Virus 51 Nucleotide CSE ......................................................................................... 106
1.26. Binding of Sindbis RNA Packaging Signal Region to Homologous and Heterologous Capsid Proteins ............................................. 109
1.27. Alphavirus RNA Packaging Signal Mutagenesis Reveals Motifs Critical for Packaging Function ............................................. 111
1.28. RNA SHAPE Secondary Structure of the Sindbis Virus Packaging Signal ..................................................................................... 115
1.29. Sindbis Subgenomic:Genomic RNA Ratio in Virions and Multinucleation Theory of Packaging ............................................. 118
1.30. Effect of Viral Mutations on Cellular Transcriptional and Translational Shutoff ................................................................. 123
# Chapter Two

2.1. Viral RNA Enrichment Efficiency of Oligonucleotide Target Capture and Polyethylene Glycol Precipitation

2.2. Quantitative PCR Standard Curves of Zika, Sindbis and Cellular RNA

2.3. Quantitative PCR Program Details for One Step Assay

2.4. Oligonucleotide Based Hybridization Capture Schematic

2.5. Electroelution and Ribodepletion Results

2.6. Oligo 7 Software Predicted Secondary Structure Free Hybridization Probes

2.7. Chemical Structure of a Biotin Tetraethylene Glycol Oligonucleotide

2.8. Majority of Target RNA is Eluted in Water and 75°C

2.9. DNA Oligos Are Eluted by Mild and Harsh Conditions

2.10. Non-Inferiority of Prepared Annealing Buffer for Oligo Capture

2.11. Comparison of Direct vs Indirect Biotinylated DNA : Streptavidin : RNA Binding Protocols

2.12. Low DNA:RNA Molar Ratio Condition Yields Higher RNA Retrieval Than Loss

2.13. Influence of DNA:RNA Molar Ratio on RNA Retrieval in Elution


2.15. Influence of Biotinylated DNA Oligonucleotides on RNA
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethanol Precipitation</td>
<td>218</td>
</tr>
<tr>
<td>2.16. Influence of 0.22 um Cellulose Acetate Filter on RNA</td>
<td>220</td>
</tr>
<tr>
<td>Ethanol Precipitation</td>
<td>222</td>
</tr>
<tr>
<td>2.17. DNA Oligo : Viral RNA Molar Ratio Determines Capture Success of 5 ug <em>in vitro</em> Transcribed Sindbis RNA</td>
<td>223</td>
</tr>
<tr>
<td>2.18. Agilent SureSelectXT Small Sample Full Kit (2016)</td>
<td>226</td>
</tr>
<tr>
<td>2.19. Large Volume Purchase of Agilent SureSelectXT Baits (2019)</td>
<td>228</td>
</tr>
<tr>
<td>2.20. A Proportion of Input RNA Non-Specifically Binds to Streptavidin Beads</td>
<td>232</td>
</tr>
<tr>
<td>2.22. Comparison of Vortexing-Based and Heating-Based Steps on Non-Specific Binding</td>
<td>239</td>
</tr>
<tr>
<td>2.23. Vortexing and DNase Digestion Does Not Affect Captured Sindbis RNA</td>
<td>241</td>
</tr>
<tr>
<td>2.24. Oligo Capture of Infected Intracellular RNA With Vortexing Steps</td>
<td>243</td>
</tr>
<tr>
<td>2.25. A Second Round of Oligo Capture Further Clears Non-Specific Binding</td>
<td>248</td>
</tr>
<tr>
<td>2.26. Combined Two Round Capture Removes 99.9% of Cellular RNA</td>
<td>250</td>
</tr>
<tr>
<td>2.27. Retention and Loss of Viral and Cellular RNA Per Round of Capture</td>
<td>254</td>
</tr>
<tr>
<td>2.28. Two Rounds of Capture Enriches Sindbis Viral RNA to 95% of Total RNA Pool</td>
<td>258</td>
</tr>
</tbody>
</table>
2.29. Next Generation Sequencing Confirms Composition of RNA Pool as 95% Sindbis Virus ................................................................. 261
2.30. Total RNA Input Per Reaction Determines Target RNA Capture Efficiency .......................................................................................... 267
2.31. Zika in vitro Transcribed RNA Capture with and without Mismatches ................................................................................................. 270
2.32 Greater Hybridization Times Increase Binding of Both Viral and Nonspecific RNA ............................................................................... 272
2.33. Prior RNA Column Purification Step is Critical for Successful Oligo Capture .................................................................................... 277
2.34. Bioinformatically Predicted Hybridization Oligos Are Highly Variable in Capture Efficiency .............................................................. 284
2.35. Two Rounds of Capture on Zika Infected Intracellular RNA Enriches Viral RNA from ~2% to ~99% of Total RNA Pool ................................................................. 286
2.36. Higher Total RNA Input in Reaction Increases Total RNA Retrieved but Reduces Viral RNA Ratio .............................................................. 290
2.37. Next Generation Sequencing on Enriched Sample Produces an Order of Magnitude more Zika Reads than Cellular RNA Reads ........................................................................................................... 293
2.38. Oligo Capture Enrichment Enables CirSeq Revealing That Zika Viral RNA is a Diverse Population ........................................................................ 295
2.39. Infection Growth Curve of Sindbis Virus Toto1101 in Huh 7.5 Cells ............................................................................................................. 297
2.40. Polyethylene Glycol Brands Behave Differently During Virion Concentration ............................................................................................... 300
2.41. Polyethylene Glycol Brands Vary in Recovery of Arbovirus Virions .......................................................... 304

Chapter Three.

3.1 The CirSeq Method ................................................................................. 315
3.2. SP6 in vitro transcribed RNA is a diverse population ..................... 317
3.3. Specific Mutation Rates of SP6 in vitro transcription reaction .......................................................... 318
3.4. RNA Synthesis Rates Are Dependent on Reaction Conditions .......................................................... 319
3.5. The Diversity of SP6 in vitro Transcribed RNA is Affected by Reaction Conditions .......................................................... 320
3.6. Dynamics of Viral RNA Population During Electroporation ............ 324
3.7. Removal of Input in vitro Transcribed RNA in Cell Culture Media .......................................................... 325
3.8. Experiment Schematic ........................................................................... 328
3.9. Total Mutation Frequency of Viral Populations Determined by CirSeq .......................................................... 329
3.10. Mean Mutation Frequency in Polymerase Active Site and Mutations to Stop Codons .......................................................... 331
3.11. Tolerability of Amino Acid Substitutions Across the Whole Genome ........................................................................... 334
3.12. Relative Enrichment of All Possible Single Nucleotide Substitutions ........................................................................... 335
3.13. Proportion of Mutation Frequency in Coding and
Noncoding Regions of the Viral Genome .......................................................... 337

3.14. Mutation Frequency Distribution by Nucleotide Position
in Codons ............................................................................................................. 339

3.15. Number of Insertions and Deletions Per Genome
Copied .................................................................................................................... 344

3.16. The Frequency of Nucleotides Inserted or Deleted .................................. 345

3.17. Size Dependence of Generated Insertions and
Deletions ............................................................................................................. 346

3.18. Lengths of Insertions and Deletions Across Viral
Genome .............................................................................................................. 350

3.19. CirSeq Sequencing Depth .......................................................................... 353

3.20. Viral Infection Population Dynamics Reveals Amino
Acid Substitution Tolerances at P1/2 Cleavage Site ........................................ 357

3.21. Amino Acid and Alternative Stop Codon Substitution
Tolerances at P3/4 Opal Stop Codon .................................................................. 362

3.22. Counts of Viral Variants in CirSeq Datasets ............................................. 364

3.23. Highly Accurate Next-Generation Sequencing Allows
High Throughput, Whole Genome Virus Genetics ........................................... 369

3.24. Viral Infection Population Dynamics Reveal Silent
Mutations Adjacent to Nonstructural Polyprotein Start Codon
Are Highly Detrimental ...................................................................................... 373

3.25. Low Tolerance for Silent Mutations in Codons Within
Sindbis Subgenomic Promoter .......................................................................... 377

3.26. Codon Context of G158A Silent Mutant Within The 51-
Nucleotide Conserved Sequence Element Copromoter .................................. 380

3.27. Estimates of in vivo Mutation Rates of Sindbis Virus
Polymerase..................................................................................................................385
3.28. Classical Genetics Based Model to Estimate Relative Fitness..................................................386
3.29. Relative Fitness Distribution for All Single Nucleotide Variants of Sindbis Virus..................................................387
3.30. Relative Fitness of Variants Highlight Structure-Function Relationships..................................................389
3.31. Schematic of Oligonucleotide Hybridization Based Capture Method..................................................396
3.32. Oligo-Capture of Infected Cell RNA Yields Pure Sindbis Viral RNA..................................................397
3.33. Intracellular Viral RNA Enrichment and Sequencing Reveals Selection During Packaging and Egress..................................................399
3.34. Sequencing Coverage for Hybridization-Enriched Sindbis Intracellular RNA passages 2 to 5..................................................400
3.35. Selection During Packaging and Egress Occurs in Packaging Signal and Across Whole Genome..................................................403
3.36. The Most Adaptive and Detrimental Variants Detected During Packaging and Egress Lie In nsP2 Gene Region..................................................404
3.37. Codon Mutants of Codons Containing Highly Detrimental Synonymous Mutations Within the Packaging Signal Region..................................................408
3.38. Mean Enrichment of Binned Sindbis Genome Regions During Packaging and Egress..................................................413
3.39. Intracellular vs Extracellular Fitness Map of Nucleocapsid Substitutions Reveals Regions Experiencing
Strong Selection During Packaging and Egress .......................................................... 416
3.40. Nonsynonymous Substitution Fitness Map of nsP2 ........................................ 425
3.41. Nonsynonymous Substitution Fitness Map of nsP3 ........................................ 427
3.42. Nonsynonymous Substitution Fitness Map of nsP4 ........................................ 429
3.43. Nonsynonymous Substitution Fitness Map of Capsid Protein ............................... 435
3.44. Nonsynonymous Substitution Fitness Map of E1-E2 Envelope Heterodimer ............... 438
3.45. Nonsynonymous Substitution Fitness Map of E3 and 6K Proteins ............................... 440

Chapter Four.
4.1. CirSeq Bioinformatics Pipeline Summary ............................................................. 463
4.2. PushpaScripts Database and Filter Functions ....................................................... 473
4.3. PushpaScripts Generated Default Outputs ............................................................ 477
4.4. Builder and Sindbis Program Inputs ....................................................................... 482
4.5. PushpaScripts Program Files .................................................................................. 487
4.6. Process Program Module CirSeq Data Flowchart ................................................... 490
4.7. The CircSeqLinuxHD Operating System ............................................................ 498

Chapter Five.
5.1. Harnessing in vitro Transcribed RNA Viral Variants to Identify Drug Resistance .......................................................... 504
5.2. Sindbis Single Molecule RNA Immuno-FISH Allows Single Cell Imaging of Viral RNA Species .......................................................... 508
5.3. Study of Various RNA Modifications Benefit from RNA
Enrichment ...........................................................................................................511

5.4. Potentially Interesting Comparisons of Viral Populations Using CirSeq ...........................................................................................................520

5.5. Transgene Directed Evolution using Sindbis and RNA Pol II Proofreading Fidelity Measurements ...........................................................................................................528

5.6. SARS Coronavirus 2 Virions ...........................................................................................................536

5.7. Exponential Growth of People Testing Positive for SARS-CoV-2 ...........................................................................................................538

5.8. Estimated Coronavirus Mutation Rates in Comparison with RNA Viruses and DNA Replication ...........................................................................................................541

5.9. Several Drug Candidates in Trials for Coronavirus Disease 2019 ...........................................................................................................544
# LIST OF TABLES

**Chapter One.**

1.1. A Filterable Agent ........................................................................................................... 18

1.2. Experimental Infection of Several Vertebrate Species via Mosquito and Injection ................................................................. 20

1.3. Specimens Examined and Isolations of Live Sindbis Virus. .................................................................................................. 21

1.4. Survey of Humans and Animals for Evidence of Prior Sindbis Virus Infection. .............................................................................. 23

1.5. Mutagenesis and RNA Synthesis of Generated Temperature-Sensitive Mutants of Sindbis Virus. ......................................................... 31

1.6. Temperature-Sensitive Mutants of Various Alphaviruses ........................................................................................................ 37

1.7. Genomic Lesions in Temperature-Sensitive Mutants of Sindbis Virus .................................................................................................. 59

1.8. Growth and Temperature Sensitivity of Nonstructural Protein Cleavage Mutants ........................................................................ 73

1.9. Translation Products of Nonstructural Protein Cleavage Mutants .................................................................................................. 76

1.10. Effect of Transexpression of nsP Cleavage Mutants on Ratio of Genomic and Subgenomic RNA ................................................................................. 81

1.11. Phenotypic Properties of Sindbis Virus 51 Nucleotide CSE Mutants .......................................................................................... 101

1.12. Growth Properties of Chimeric Alphaviruses with Heterologous Packaging Signals and Capsid Proteins ................................................ 113
1.13. Sindbis Virus Capsid Protein Domains

Chapter Two.

2.1. Cost Comparison of Oligonucleotide Capture RNA Enrichment Method

2.2. Cost Comparison of Polyethylene Glycol Based Virus Concentration

2.3. Preparation of 2x Bind & Wash Buffer

2.4. Preparation of 10x Annealing Buffer

2.5. Hybridization Oligos Aligned to Human and Mosquito Genome and Transcriptome

2.6. Typical Elution Efficiencies of Biotin-Streptavidin Interaction in Different Elution Conditions

2.7. Harsh Elution Comparison Reveals High Efficiency of Mild Conditions

2.8. Loss of Viral RNA in Oligo Capture Solutions During Column-Based RNA Purification

2.9. Ethanol Precipitation Efficiencies of Unfiltered RNA Solutions

2.10. Cost Comparison of Baits for Our Protocol Versus Common Kit

2.11. Combined Two Round Capture Results on Infected Intracellular RNA

2.12. Combined Capture Results on High Input Infected Intracellular RNA

2.13. Effect of Hybridization Times on Viral and Nonspecific
RNA Capture .................................................................................................................. 274

Chapter Three.
3.1. The Most De-Enriched Synonymous Mutants Within The 10th percentile of all De-enriched Single Nucleotide Mutations During Infection .................................................................................................................. 443
3.2. Human Codon Usage of Selected Sindbis Synonymous Variants ........................................................................................................................................................................... 450
3.3. The 25 Most Significantly De-Enriched Synonymous Mutants of All Single Nucleotide Mutations During Packaging and Egress .................................................................................................................. 451

Chapter Four.
4.1. CirSeqScripts Insertion and Deletion Analysis Tables ........................................................................................................................................................................... 468
4.2. CirSeqScripts Mutation Counts and Rates Tables ........................................................................................................................................................................... 470
4.3. Directory of Functions and Required Inputs ........................................................................................................................................................................... 488
4.4. CPU Processor Single Thread Benchmarks ........................................................................................................................................................................... 495

Chapter Five.
5.1. Several Attenuated Vaccine Strains of Clinically Relevant Arboviruses ........................................................................................................................................................................... 525
5.2. Brief Comparison of Common Current Next-Generation Sequencing Platforms (2020) ........................................................................................................................................................................... 532
**LIST OF ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(+)RNA</td>
<td>plus strand RNA, serves as messenger</td>
</tr>
<tr>
<td>(-)RNA</td>
<td>minus strand RNA, complementary to (+)RNA</td>
</tr>
<tr>
<td>26S</td>
<td>alphavirus subgenomic RNA, sediments at 26 Svedbergs</td>
</tr>
<tr>
<td>3' UTR</td>
<td>three prime untranslated region</td>
</tr>
<tr>
<td>42S, 45S, 49S</td>
<td>alphavirus genomic RNA, sediments at ~45 Svedbergs</td>
</tr>
<tr>
<td>5' UTR</td>
<td>five prime untranslated region</td>
</tr>
<tr>
<td>aa</td>
<td>amino acid</td>
</tr>
<tr>
<td>AGO2</td>
<td>protein argonaute 2 (RNAi pathway component)</td>
</tr>
<tr>
<td>AR339</td>
<td>original SINV wild isolate strain</td>
</tr>
<tr>
<td>BHK</td>
<td>baby hamster kidney cell line</td>
</tr>
<tr>
<td>BSL</td>
<td>biosafety level (facility requirement)</td>
</tr>
<tr>
<td>C6/36</td>
<td>an <em>Aedes albopictus</em> cell line</td>
</tr>
<tr>
<td>C7/10</td>
<td>an <em>Aedes albopictus</em> cell line</td>
</tr>
<tr>
<td>CEF</td>
<td>chicken embryo fibroblast cells</td>
</tr>
<tr>
<td>CF</td>
<td>complement fixation assay</td>
</tr>
<tr>
<td>CHIKV</td>
<td>chikungunya virus</td>
</tr>
<tr>
<td>COVID-19</td>
<td>coronavirus disease 2019, caused by SARS-CoV-2</td>
</tr>
<tr>
<td>CPM</td>
<td>radioactivity counts per minute</td>
</tr>
<tr>
<td>CSE</td>
<td>conserved sequence element</td>
</tr>
<tr>
<td>DENV</td>
<td>dengue virus</td>
</tr>
<tr>
<td>DI</td>
<td>defective interfering virus RNA or particle</td>
</tr>
<tr>
<td>dsRNA</td>
<td>double stranded RNA</td>
</tr>
<tr>
<td>EEEV</td>
<td>eastern equine encephalitis virus</td>
</tr>
<tr>
<td>GFP</td>
<td>green fluorescent protein</td>
</tr>
<tr>
<td>gRNA</td>
<td>genome strand of an RNA virus</td>
</tr>
<tr>
<td>h, hrs</td>
<td>hours</td>
</tr>
<tr>
<td>HBV</td>
<td>hepatitis B virus</td>
</tr>
<tr>
<td>HCV</td>
<td>hepatitis C virus</td>
</tr>
<tr>
<td>HI</td>
<td>hemagglutination inhibition assay</td>
</tr>
<tr>
<td>HJV</td>
<td><em>Highlands J</em> virus</td>
</tr>
<tr>
<td>HR</td>
<td>heat resistant SINV strain</td>
</tr>
<tr>
<td>Huh-7.5</td>
<td>human hepatoma cell line</td>
</tr>
<tr>
<td>IFN</td>
<td>interferon</td>
</tr>
<tr>
<td>ISG</td>
<td>interferon-stimulated gene</td>
</tr>
</tbody>
</table>
ivt ................................................................. in vitro transcribed
KFDV ......................................................... Kyasanur Forest disease virus
LD50 .................................................................. median lethal dose
LP ............................................................. large plaque producing viral variant (relative to WT)
MOI .............................................................. multiplicity of infection
NGS .................................................................... next-generation sequencing
NIAID .......... National Institute of Allergy and Infectious Diseases, at NIH
NIH ............................................................. National Institutes of Health (Bethesda, Maryland)
NIH 3T3 ........................................................... NIH swiss mouse embryo cell line
nsP ................................................................... alphavirus nonstructural protein
nt ....................................................................... nucleotide
ONNV .............................................................. O'Nyong Nyong virus
OS ..................................................................... operating system
p.e. ...................................................................... post-electroporation
p.i. ....................................................................... post-infection
P1234 .............................................................. alphavirus nonstructural polyprotein
PCR ................................................................. polymerase chain reaction
PFU ..................................................................... plaque forming units
PolyI:C ............................................................ hybridized polyinosinic and polycytidylic acids
PS ...................................................................... alphavirus packaging signal
qPCR ................................................................... quantitative PCR
rDNA......................................................... ribosomal DNA, from which rRNA is transcribed
RdRp .............................................................. RNA-dependent RNA polymerase
RNA- ............................................................ ts mutant that produces < 10% RNA as WT
RNA+ ............................................................ ts mutant that produces > 60% RNA as WT
rRNA ................................................................... ribosomal RNA
rNTP .............................................................. ribonucleoside tri-phosphate, monomer of RNA
RRV ................................................................. Ross River virus
RT ........................................ reverse transcriptase / room temperature (context-dependent)
RT-PCR ........................................................ PCR of RNA reverse transcribed into DNA
Rxn ................................................................. reaction
SARS-CoV-2 ........................................ sudden acute respiratory syndrome coronavirus 2
SD ...................................................................... standard deviation
SEM .................................................................... standard error of the mean
SFV ................................................................. Semliki Forest virus
sgRNA .................................... alphavirus subgenomic RNA, also called 26S
SHAPE ................................ selective 2’ hydroxyl acylation analyzed by primer extension
SINV ............................................................... Sindbis virus
smFISH................................. single molecule fluorescent in situ hybridization
smFRET ......................... single molecule fluorescence resonance energy transfer
SP ................................... small plaque producing viral variant (relative to WT)
ssRNA ........................................................... single stranded RNA
STAT-1 ..................... signal transducer and activator of transcription 1 (an ISG)
Toto1101 ................................................. a SINV infectious clone strain
Ts ........................................ temperature sensitive
VEEV ........................................... Venezuelan equine encephalitis virus
VEGAS .................................. viral evolution of genetically actuating sequences
Vero .......................................................... an African green monkey cell line
WEEV ........................................ western equine encephalitis virus
WNV ......................................................... West Nile virus
WT ........................................ wild type or parental strain (context-dependent)
YFV .......................................................... yellow fever virus
ZAP ...................................................... zinc finger antiviral protein (an ISG)
ZCS .......................................................... ZIKV Cambodian strain
ZIKV .......................................................... Zika virus
ZPR .......................................................... ZIKV Puerto Rican strain
Chapter I contains a brief introduction to RNA virus evolution, followed by a description of the goals of this project. Then it contains a fairly significant foray into the history of Sindbis virology, within the context of alphavirology and arbovirology. The chapter is intended to place what is known about the features of alphavirus genomes in context. Chapter II describes the methods used in this project. Moreover, it is written with the intention of enabling the reader to purify any viral RNA (or any RNA species) or packaged virions of interest with high efficiency.

Chapter III details the results and relevant discussions from our efforts to sequence the evolution of virus populations with high accuracy, and harness that information for high-throughput genetics. It contains what I was predominantly, but not solely, up to during graduate school. Chapter IV describes the bioinformatics tools developed and used for this project, and is intended to equip an interested reader with the ability to perform all the analyses shown in the previous chapter.

Chapter V contains musings on potentially interesting future directions. It ends with a brief discussion of interesting experiments that could be performed on coronaviruses, which at the time of writing is the cause of a global pandemic. The first appendix is meant as a companion to a comprehensive 1994 review on alphaviruses by Drs. Jim and Ellen Strauss, and the second appendix is an expanded acknowledgements section. Looking back on my experience of writing this document, I have unofficially titled this thesis dissertation "Sindbis Virus RNA: A Love Story".
Chapter I. Introduction: RNA Virus Evolution and Sindbis Virus Genetics.

1.1. RNA Virus Evolution.

To successfully replicate, RNA viruses must contend with diverse physiological environments, defense mechanisms, and anatomical restrictions within a cell and highly variable environments in cell-to-cell transmission [1]. In addition to contributing to the aforementioned abilities, it has been long reported that the difficulty in developing drugs and vaccines for viral diseases is due to their rapid evolution and adaptation [2]. Relative to all other phylogenies of life, most RNA viruses possess both extremely high mutation rates and very small genomes, usually leading to mutations introduced every time they replicate [3]. These mutations create a pool of viral genome variants, sometimes referred to as a ‘quasispecies’, in which subpopulations of mutants with differing fitness will facilitate adaptation to new conditions (Figure 1.1) [1].

**Figure 1.1. Evolution of Viral Populations.** Viral genomes are represented as horizontal lines and mutations as different colored symbols on the lines. Due to
its low fidelity polymerase, an RNA virus exists and evolves as a dynamic population of mutants. Constant evolution in the mutant spectrum can still yield the same consensus sequence. The consensus sequence can change when a selective advantage drives a mutation to dominance, but will still mask the true diversity of the adapting population. Adapted from [4].

In order to manipulate viral genomes and study key aspects of the viral life cycle, virologists have traditionally worked only with the consensus sequence, which lists the most frequent nucleotide at each position [5]. However, over the past forty years it has become increasingly evident that the consensus sequence alone masks huge amounts of genetic diversity. This diversity was first experimentally shown using the Q beta RNA bacteriophage, where it was determined that the virus cannot be described as a homogenous population that carries the same sequence, but rather as a large number of genetically related but distinct sequences [6]. A single nucleotide mutation in a virus can dramatically change its fitness in a given environment, such as conferring interferon or antiviral drug resistance or lethality [1]. The viral population heterogeneity thus creates a pool of individuals with diverse fitness values, where beneficial subpopulations can enable adaptation to changing environments and selective pressures [4]. Beyond the consensus sequence, Sanger sequencing can only detect variants present above a frequency of 20% in the population, and traditional deep sequencing can only detect variants above a 1% frequency.
The nucleotide incorporation errors introduced by the low-fidelity polymerases of RNA viruses is primarily responsible for their high mutation rates, though they can also be influenced by recombination, post-transcriptional editing and genetic drift. [2]. It was shown in poliovirus that introduced mutations that increased RNA polymerase fidelity, or addition of mutagens that decreased polymerase fidelity, both resulted in drastically reduced fitness and pathogenicity [1]. This suggests that the low fidelity of the RNA polymerase is evolutionarily determined and is within an optimal range for viral survival. The spectrum of mutants generated in any given environment is determined by the balance between mutation rates and selection acting on those mutants. Evidence suggests that RNA viruses have a highly dynamic population structure, where selection constantly drives the population to new sequence spaces even in a relatively constant environment, supporting the Red Queen hypothesis; “it takes all the running one can do just to stay in the same place” [7, 8]. When strong selective pressures are imposed, the virus population rapidly evolves to a new sequence space using its extreme population heterogeneity [9]. Major events in RNA virus biology, such as their capacity to change their cell tropism or host range or to overcome selective constraints such as immune responses and antiviral agents, originate via present and arising variants in the population’s mutant spectrum [4]. Thus, distinct viral subpopulations emerge when selective pressures are applied, revealing functional adaptive strategies.
The enormous human disease burden and economic distresses caused by RNA viruses highlight the importance of understanding their adaptation patterns. As a necessary corollary, a better understanding of viral evolution processes can have, and indeed has had, major implications for public and economic health [4]. Additionally, the mutant spectrum generated by low fidelity polymerases and the explosive replication of RNA viruses provides a rapid system in which natural selection acts and evolutionary hypotheses can be tested [3]. Fitness describes the ability of an organism to succeed in a given environment, and in virology is usually defined as replicative ability, which is quantitated via growth assays [10]. The fitness distributions of viral variants can be determined by performing growth assays on constructed mutants and comparing replicative ability to the wild-type virus (Figure 1.2) [5, 11, 12].
Figure 1.2. Fitness Effect Distribution of Random Mutations. The frequency of fitness values, defined as the ratio of the titer of the mutant to the titer of the wild type, for 48 random mutations inserted into vesicular stomatitis virus. Fitness of one indicated a neutral mutation, fitness below one indicates that the mutant is less fit than the wild type, and a fitness value of zero indicates lethality; no virus with that mutation was recovered. Adapted from [11, 13].

However, the analyses of processes that underlie the adaptation of RNA viruses to selective pressures suffer when the sample sizes are small and consequently biased [3]. This has historically been the case, as sequencing technologies that were in common use over the past few decades can only identify the consensus sequence or very high frequency variants, leaving most of the population diversity a mystery. Next generation sequencing platforms now provide the depth necessary to sequence entire genomes and transcriptional profiles of higher order organisms, and entire populations of RNA viruses [14, 15]. A current problem of deep sequencing techniques, however, is that their error rate is in the same range as RNA viruses, making viral mutations and intrinsic errors of sample preparation and deep sequencing indistinguishable (Figure 1.3A). Circular sequencing, which uses rolling reverse transcription of viral RNA fragments (CirSeq) can be used to overcome this limitation [16]. The physically linked tandem repeat cDNA can be bioinformatically parsed using the fact that processing errors will likely be present in only one of the repeats (Figure 1.3B). This optimization reduces the error probability from $10^{-4}$ to $10^{-12}$, several orders of magnitude below the intrinsic mutation rate of RNA viruses. This enables the detection of very low
frequency variants that were previously indistinguishable and allows the accurate characterization of complete RNA mutational populations [5].

![Mutation rate graph](image)

**Figure 1.3. Polymerase and Detection Error Rates and Error Rate Correction Method (CirSeq).** A) The ranges of error rates per nucleotide for viruses with different genomic architectures, humans, and next-generation sequencing platforms. Current platforms provide immense depth but most sequencing and preparation errors are indistinguishable from intrinsic RNA virus mutations. Adapted from [2]. B) CirSeq overcomes this limitation by using circularized genomic fragments that serve as templates for rolling-circle replication, producing tandem repeats which are aligned to generate a consensus. Green symbols represent true genetic variation, other colors represent random sequencing error. Adapted from [5].
The construction of full-length cDNA clones corresponding to the entire genome of a virus is a crucial development in the field of virology. Most studies are performed with these clones, which have allowed the identification of genomic regions essential for various aspects of the viral life cycle and genetic markers that facilitate adaptation [17]. A typical laboratory infection establishment procedure for positive-stranded RNA viruses involves an *in vitro* RNA transcription of a viral infectious clone, which is electroporated into a cell type of interest, which allows the viral RNA to surpass the entry steps normally facilitated by the virion (Figure 1.4). This leads to virus replication and production and release of infectious virions by the cells. The mature virions are then serially passaged in cells using a desired ratio of infectious virus particles to target cells, termed multiplicity of infection (MOI). However, the evolutionary dynamics of RNA virus infection establishment remain poorly understood. Selective pressures faced by variant genomes during packaging and exit from the cell are also similarly poorly understood. There is also evidence that synthetically produced RNA polymerases used *in vitro* introduce errors into the transcribed RNA population when transcribing the cDNA clone [18–20]. If the *in vitro* transcribed RNA population is itself diverse, perhaps their dynamics upon encountering selective pressures during cellular infection can yield useful information.
Figure 1.4. The Sindbis Life Cycle and Genome Composition. A) Life cycle, B) Genome composition. After entry, the viral genomic ssRNA is translated into the nonstructural polyprotein, which is cleaved into the nonstructural proteins which form the replication complex: nsP1, nsP2, nsP3 & nsP4. This complex
transcribes new viral RNA’s and the subgenomic RNA which is translated into a polyprotein and cleaved into the various structural proteins required for assembly: Capsid, E1, E2, E3 & 6K. Capsid binding to viral RNA and glycoprotein recruitment facilitates budding of the virion from the cell membrane. Adapted from [21, 22].

Historically, studies of RNA virus infectious clones have led to the identification and characterization of various domains and genetic markers of functional significance (Figure 1.4). Such studies typically compare the sequences of different strains of a virus or different viruses within a family highlighting their macroevolution, or evolution on the scale of separated gene pools [23–26]. Conserved regions, which indicate macroevolutionary selective constraints, are then subject to mutagenesis studies and functional characterization [17–19, 27]. However, the vast scale of time and space that belies macroevolution limits robust functional characterization of selection pressures and viral adaptation responses.

Microevolutionary patterns, which refer to changing variant frequencies within a population, can be characterized with unprecedented resolution using accurate full population sequencing (CirSeq). Genomic regions that vary in microevolutionary constraints when a particular selection pressure is imposed should reflect viral adaptation strategies. Microevolutionarily conserved regions that are not macroevolutionarily conserved may reveal new genetic features of importance to a specific virus (in a specific host context) that were
missed in previous studies. Viral polymerases continuously generate mutants upon which selection acts, allowing the assessment of the dynamics of every variant and robust characterization of adaptation mechanisms [13]. The main goal of this project is to harness the dynamics of viral mutants during adaptation to various selection pressures for high-throughput genetics. Arboviruses (arthropod-borne viruses), particularly Sindbis virus (SINV), represent one of the best-studied groups of animal viruses, and provide a vast literature background upon which to compare, test and expand our analyses [17].

1.2. Significance

One fascinating aspect of viral infection is the tremendous speed by which viral evolution takes place. This allows viruses to be one step ahead of the host immune system, spread between hosts, and poses a constant problem to antiviral therapy. Additionally, arthropod-borne RNA (arbo)viruses, which are predominantly single-stranded, positive-sense RNA viruses, cycle between vertebrate and arthropod hosts, rapidly adapting to heterogeneous host defenses. Arboviruses such as chikungunya virus, Zika virus, dengue virus, and West Nile virus cause clinically significant diseases for which there are no effective treatments [28]. Arboviruses pose a serious threat to human health in many areas of the world, which was highlighted by the 1999 outbreak of West Nile meningoencephalitis in New York, the 2008 global outbreak of dengue
fever, the 2013 global outbreak of chikungunya fever, and the 2016 Zika fever epidemic in the Americas [29–33]. Elucidation of viral genetic mechanisms will lead to a better understanding of cell biology and molecular virology, and consequently, pathogenesis. Efforts to understand these adaptive mechanisms may lead to new and improved therapeutics to combat various viral infections critical to human health.

Additionally, undergoing evolution as populations of heterogeneous genetic variants in response to selective pressures is not exclusive to RNA viruses, but also applies to many biological areas including other types of viruses, microbes, and cancer [34, 35]. Such mechanisms are thought to underlie bacterial development of antibiotic resistance and the rise of chemotherapy resistance in tumor cells. Therefore, identifying RNA virus adaptation mechanisms has broad implications across diverse fields beyond infectious diseases, and the techniques and approaches proposed for the RNA viruses above will be a stepping stone in the quest to understand the interplay between selection and adaptation.

1.3. Discovery of Sindbis Virus.

Herein I detail the discovery of Sindbis virus, within the context of a period of arbovirus discovery. Global surveillance efforts in the 1940’s to the 1960’s yielded the discovery of many new viruses, spearheaded by the United States
military energized by World War II, and the Rockefeller Foundation, then stationed at what is now Rockefeller University in New York City and the Forrestal Campus of Princeton University. What initially started as surveillance for yellow fever transformed into "a shotgun approach at what may be out there" [36]. Dr. Tom Monath, prior director of the Division of Vector-Borne Viral Diseases at the Centers for Disease Control (CDC), told me over lunch that over 600 individual viruses were discovered by the program. Known arboviruses increased steadily from 5 prior to 1930, to 34 in 1950, to 204 in 1966, to 440 in 1981 [37]. Newly discovered arboviruses in this era include recent global epidemic agents such as West Nile virus and Zika virus, and will surely include agents of future epidemics, especially with increasing air travel and as mosquito and tick species expand to new territories. The details of the discovery and classification of Sindbis virus serves as a nice example of how one goes from not knowing anything to knowing something about a virus, as the tools used are largely the same for the hundreds of newly discovered arboviruses, and often involved the same researchers as well. I direct the reader to Dr. Charles Calisher's excellent book *Lifting the Impenetrable Veil: From Yellow Fever to Ebola Hemorrhagic Fever & SARS* for a comprehensive treatment of arbovirus discovery, and to the film *Story of Kyasanur Forest Disease* by Dr. Telford Work for an intriguing movie of an example (Figure 1.5) [38, 39].
These 'shotgun approach' viral surveillance efforts involved testing for the presence of an infectious agent from (i) blood of healthy humans and a variety of animals, (ii) blood of humans and animals with illness (Figure 1.5 A), and various organs of dead humans and animals, and (iii) arthropods collected from the wild such as mosquitoes and ticks (Figure 1.5 B). In the first case, the blood was typically stored and tested for the presence of antibodies, indicating a prior infection, after the discovery of a virus. In the second case, sentinel monkeys, i.e., monkeys in cages whose temperature was monitored daily and blood drawn in instances of fever or visible illness, were frequently employed in these studies. Zika virus, for example, was discovered in a rhesus macaque caged in the Zika forest neighboring Lake Victoria in Uganda, from blood drawn when macaque number 766 exhibited a 104°F fever (prototype MR766 strain; [Sentinel] Macaque Rhesus Number 766) [40, 41]. There are several recorded instances of experimental difficulties due to sentinel monkeys being eaten by predators [42]. The third case yielded the discovery of Sindbis virus, as described by Richard Taylor; "The virus was first encountered in August 1952, in a group of 63 Culex pipiens and/or Culex univittatus mosquitoes captured by light-trap in the Sindbis health district, situated some 30 kilometers north of Cairo... upon arrival at the laboratory, the arthropods were pooled according to classification (usually to species) and source (Figure 1.5 C). Each pool was triturated in a porcelain mortar with pestle and suspended [in] 10 per cent inactivated normal rabbit serum in 0.85 per cent
NaCl. The suspension was centrifuged for 20 minutes at 2000 rpm and the supernate withdrawn for inoculation and storage" [43]. The discovery experiments were performed at U.S. Naval Medical Research Unit #3 (NAMRU-3), an extant military emerging diseases research facility in Cairo, Egypt (Figure 1.5 D). Drs. Richard Taylor and U.S. Navy Captain Herbert Hurlbut (PhD, Cornell University, 1940) [44] originally isolated the virus [45, 46], and Telford Work, James Kingston and Thomas Frothingham assisted with its characterization [43] - the book by Charles Calisher mentioned above contains more details on the interesting lives of these globetrotting virus discoverers.

The following procedure, described for Sindbis virus using the above inoculum, applies to most viruses discovered in this era using the respective biofluid or organ homogenate; "Infant mice one to three days of age were routinely employed for virus isolation.. each mouse of a litter.. received a total of 0.05 ml of the suspension, 0.02 ml intracranially and 0.03 ml subcutaneously (Figure 1.5 E).. the mice were observed daily for a period of three to four weeks and any dying or becoming manifestly ill (Figure 1.5 F) were autopsied and a 10 per cent suspension of the brains prepared (Figure 1.5 G) and cultured for bacteria. If the brain suspension was bacteriologically sterile, it was passed to a new litter of mice. If the passage of bacteria-free brain suspension consistently produced paralysis or death of the mice the
presence of a virus was suspected and the problem of identification was then presented" [43]. Viral titers, which nowadays almost always refer to quantitation using cell culture methods, here refers to the LD50 (median lethal dose) of infant mice infections. Stocks of the fourth to seventh infant mouse brain infection passages were named as the AR339 strain, now considered the original 'wild' strain of Sindbis virus. It is important to remember that in almost no case is the originally isolated virus still available from this era, and what is referred to as the original strain is usually virus passaged anywhere from a couple to hundreds of times in infant mouse brains and other sources, which could have allowed the accumulation of viral mutations. The first clues, including the fact that the virus caused death in infant but not in adult mice, produced skeletal muscle inflammation and was resistant to overnight 20% ether treatment, led the discoverers (Figure 1.5 H) to erroneously classify it as a 'Coxsackie-like' virus [45, 46]. The next set of clues allowed the proper classification of the virus.
Figure 1.5. Field Virus Discovery in 1950s. Viruses were often discovered from sick animals (A) or arthropod collections in forests (B) which were then sorted by entomologists (C) and pooled for testing. Sindbis virus was similarly discovered from a pool of *Culex* mosquitoes at NAMRU-3 (D), a US Navy installation in Cairo, Egypt. Human or animal samples or pooled arthropod extracts are injected into the brains of infant mice (E) and paralysis or death of injected mice (F) would indicate the potential presence of a virus. Infant mouse brains would then be homogenized and resuspended (G) to make the virus stock which can then be passaged in infant mouse brains as above or be subject to other analysis. The production of inactivated vaccines in this era involved simply adding formaldehyde to the mouse brain suspension and refrigerating it for a couple of weeks, followed by efficacy testing in humans [47]. A-C and E-G are still images from a movie *Story of Kyasanur Forest Disease* made by Telford Work [39], one of the discoverers of Sindbis virus, on his discovery of a virus in India using the same tools [48]. More information on Telford Work, pictured here with his pet leopard Pushpa (H), and other virus discoverers and their efforts can be found in the book *Lifting the Impenetrable Veil* by Charles Calisher [38].
I thank Tom Monath for procuring and shipping a copy of the movie to me - please contact me at pmambrose@gmail.com or the Rice laboratory if the reader is interested in a copy.

The infected infant mouse brain suspension was passed through filters of varying pore sizes also to confirm that the transmissible disease agent is a virus and determine the approximate size of the viral particle (Table 1.1). This size estimate placed Sindbis virus closer to other alphaviruses and flaviviruses than to bunyaviruses discovered in the area at the time. U.S. NAMRU-3 had a large flock of chicken hens on campus, which conveniently allowed for a variety of chicken infection experiments [43]. The virus suspension was found to be highly pathogenic to embryonated chicken eggs, and cause the death of embryos within 1 to 3 days. Additionally, the virus quickly induced cell death in chicken embryo fibroblasts (CEFs) - outgrowths of fibroblastic cells from minced chick embryo tissue on glass cell culture plates [49]. The virus was not cytopathic to HeLa cells, which was commonly used for growing poliovirus at the time. Consequently, CEFs would become the primary cell culture system for the study of Sindbis virus until the 1980's, whereupon it was largely superseded by mammalian cell cultures. Young chicks were found to be easily infected by the bite of between one and three infected mosquitoes and produce blood titers high enough to re-infect naive biting mosquitoes. Thus a potential wild transmission cycle between young birds and Culex mosquitoes was established. However, none of the young chicks or adult chickens that were proven to be infected showed any symptoms or died, and pathological
effects were exclusive to chicken embryos.

**Table 1.1. A Filterable Agent.** Filtration experiments both confirmed that the agent killing infant mice upon injection was not a bacteria and likely viral in nature, and gave a size estimate of the virus particles which can then be compared to other known viruses for initial classification. * A titer of $<10^{-2}$ means that no virus was recovered from the undiluted filtrate. † from The Rockefeller Foundation Virus Laboratories, New York, USA. ‡ from A. Gallenkamp & Company, London, England. From [43].

<table>
<thead>
<tr>
<th>Membrane pore size</th>
<th>Source of membrane</th>
<th>Titer of virus$^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>N. Y.†</td>
<td>$10^{-7.4}$</td>
</tr>
<tr>
<td>95.8</td>
<td>N. Y.</td>
<td>$&gt;10^{-4.5}$</td>
</tr>
<tr>
<td>79.4</td>
<td>N. Y.</td>
<td>$&lt;10^{-3}$</td>
</tr>
<tr>
<td>45</td>
<td>N. Y.</td>
<td>$&lt;10^{-8}$</td>
</tr>
<tr>
<td>300</td>
<td>N. Y.</td>
<td>$10^{-7.4}$</td>
</tr>
<tr>
<td>100</td>
<td>England‡</td>
<td>$&gt;10^{-4.5}$</td>
</tr>
<tr>
<td>77</td>
<td>England</td>
<td>$&lt;10^{-1}$</td>
</tr>
<tr>
<td>30</td>
<td>England</td>
<td>$&lt;10^{-2}$</td>
</tr>
</tbody>
</table>

The rapid death of CEFs allowed for the development of a cell culture based neutralization assay, which was uncommon [49]. Neutralization assays were typically conducted by mixing blood containing potential antibodies with dilutions of the virus stock and seeing if the LD50 of intracerebrally injected infant mice was lower than that of the virus stock alone, due to binding and inactivation by these antibodies [40, 41]. For Sindbis virus, the neutralization potency of antibodies was instead quantified as the difference between the dilution of virus stock and the dilution of virus mixed with antibody-containing blood that caused visible CEF cell death. Testing a variety of known antibodies provided the final clues that Sindbis virus is likely a new discovery; "Sindbis
virus is not neutralized by immune sera of any of the following viruses: Eastern, Western and Venezuelan equine encephalomyelitis viruses, Semliki forest, Japanese B, Murray valley, Ilheus, Anopheles A and B, Bunyamwera, West Nile, Ntaya, Zika, Dengue 2, Encephalomyocarditis, Bwamba fever, St. Louis, and Russian [Spring and Summer]" [43]. Finally, hemagglutination inhibition (HI) and complement fixation (CF) tests developed by Dr. Jordi Casals aided classification of the new virus. Hemagglutination refers simply to the visible clumping of red blood cells (RBCs), most often extracted from chicken or geese in these studies, due to the recognition and binding of receptors on the surface of RBCs to some part of the virus [37, 50]. HI refers to when the addition of sera prevents RBC clumping presumably due to antibodies in the sera preferentially binding the virus and blocking recognition by RBC receptors; "[HI] Group A [arboviruses] comprises Eastern, Western and Venezuelan equine encephalomyelitis viruses, and Semliki forest virus. It is into this group that Sindbis virus falls. Although HI and CF reveal an antigenic component common to the group, Sindbis virus is distinguishable from other members of the group by quantitative titration, and it is easily differentiated by means of the neutralization test" [43, 51]. Group A arboviruses are now called alphaviruses, and members of the group which are agents of multiple recent epidemics, such as Chikungunya, were yet to be classified [29]. A variety of vertebrates in addition to chickens were then infected by both mosquito bites and injection, including grivet and rhesus
monkeys, sheep and several bird species (Table 1.2). Neutralization tests showed that all animals except for sheep produced antibodies against Sindbis virus.

Table 1.2. Experimental Infection of Several Vertebrate Species via Mosquito and Injection. Several vertebrates were infected using bites from multiple infected mosquitoes or subcutaneous injections, and virus titer was assayed using LD50 of infant mice. The 4th to 7th infant mouse passage of the AR339 Sindbis strain was used for these inoculations. The presence of antibodies against Sindbis virus was then checked over a month after infection using neutralization assays in CEF cells. - = undetected, ND = not done, D = died. From [43].

<table>
<thead>
<tr>
<th>Animal species</th>
<th>Infected by</th>
<th>Virus titer in blood (Log) Day following infection</th>
<th>Neutralization test</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Griet monkey P. (Cercopithecus aethiops) 14</td>
<td>Bite 13 moeq.</td>
<td>0.5 1.5 1.0 1.5 — — 4.8 41</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>Rhesus monkey P. (Macaca mulatta) 15</td>
<td>Bite 9 moeq.</td>
<td>1.8 4.5 2.5 2.5 — — 4.5 30-60</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>Chicken (27 days) 220</td>
<td>Bite 7 moeq.</td>
<td>&gt;2.5 &gt;2.5 &lt;1.0 ND ND &gt;2.0 22</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>Chicken (27 days) 221</td>
<td>Bite 5 moeq.</td>
<td>&gt;2.5 &gt;2.5 &lt;1.0 ND ND &gt;2.0 22</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>Chicken (27 days) 223</td>
<td>Bite 4 moeq.</td>
<td>&gt;2.5 &gt;2.5 1.5 ND ND &gt;2.0 22</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>Chicken (young adult) 317</td>
<td>Bite 12 moeq.</td>
<td>2.5 2.5 1.6 &lt;1.0 ND ND &gt;2.0 20</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Chicken (young adult) 318</td>
<td>Bite 2 moeq.</td>
<td>&lt;1.0 0.8 2.5 1.5 ND ND &gt;2.0 30</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>Zebra (Buff-banked) (Bubalaea ibis) 323</td>
<td>Inoc. s.e. 10^4</td>
<td>3.5 3.5 2.5 2.5 2.5 2.5 D</td>
<td>247</td>
<td></td>
</tr>
<tr>
<td>Culex (Biology) (Cercops cuneiformis) 247</td>
<td>Inoc. s.e. 10^4</td>
<td>? &lt;1.0 1.4 &lt;1.0 ND ND &gt;2.0 35</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Culex (Biology) (Cercops cuneiformis) 247</td>
<td>Inoc. s.e. 10^4</td>
<td>2.5 2.5 2.5 2.5 2.5 2.5 D</td>
<td>267</td>
<td></td>
</tr>
<tr>
<td>Culex (Biology) (Cercops cuneiformis) 247</td>
<td>Inoc. s.e. 10^4</td>
<td>? ? ? 1.8 &lt;1.0 ND ND &gt;2.0 35</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>Sheep 144</td>
<td>Bite 5 moeq.</td>
<td>— — — — — — 55</td>
<td>35</td>
<td></td>
</tr>
</tbody>
</table>

These antibodies provided a tool that can be used to detect Sindbis virus in various samples, whereupon the samples were injected into baby mice and passaged if the mice experienced paralysis or death, and then subjected to neutralization tests using CEFs. Sindbis virus was thus detected in Culex mosquitoes 12 times in three consecutive summers and once in a juvenile hooded crow which "was shot and whether or not it was sick [was] unknown" (Table 1.3) [43]. It is important to realize that this study did not occur in a vacuum; "besides the 13 isolations of Sindbis virus, the inoculation of the
vertebrate and arthropod specimens listed in Table [1.3] has yielded **145 other virus strains** - 71 from human bloods, 9 from birds, and the remainder from mosquitoes and ticks. Many of these strains have not been identified and are still under study. Of the identified strains, West Nile, Coxsackie, and Sandfly fever viruses have been isolated from human blood, and West Nile from a crow and two pigeons and from 21 pools of mosquitoes" [43, 52].

**Table 1.3. Specimens Examined and Isolations of Live Sindbis Virus.** The total number of vertebrate blood samples and sorted homogenized pools of arthropods examined for Sindbis virus using infant mouse brain passaging and neutralization tests with Sindbis virus antibodies (1951-1954). Human blood samples were extracted during a fever episode or visible illness. 12 pools of *Culex* mosquitoes over three consecutive summers and one juvenile hooded crow tested positive for the virus. The total samples also yielded 145 other viruses including West Nile virus, Coxsackie virus and sandfly fever virus. From [43].

<table>
<thead>
<tr>
<th>Nature of specimen</th>
<th>No. of specimens</th>
<th>No. of arthropod pools inoculated</th>
<th>No. of Sindbis virus isolations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vertebrates: (Blood specimens)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human ..................</td>
<td>3,218</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Hooded crow ..........</td>
<td>159</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buff-backed heron ..</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pigeon ...............</td>
<td>34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chicken ..............</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total vertebrates</strong></td>
<td><strong>3,479</strong></td>
<td></td>
<td><strong>1</strong></td>
</tr>
<tr>
<td><strong>Hemophagous arthropods:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mosquitoes ............</td>
<td>51,939</td>
<td>1,004</td>
<td>12</td>
</tr>
<tr>
<td>Ticks ..................</td>
<td>5,849</td>
<td>322</td>
<td></td>
</tr>
<tr>
<td>Fleas ..................</td>
<td>3,272</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>Lice ....................</td>
<td>3,648</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>Mites ...................</td>
<td>6,887</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>Sandflies (<em>Phlebotomus</em> sp.)</td>
<td>123</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td><strong>Total hemophagous arthropods</strong></td>
<td><strong>71,718</strong></td>
<td><strong>1,514</strong></td>
<td><strong>12</strong></td>
</tr>
</tbody>
</table>
Finally, the prevalence of prior infections was surveyed by conducting neutralization tests on blood samples from humans and various animals, with the assumption that the presence of antibodies against Sindbis virus indicated prior infection and clearance (Table 1.4). These surveys indicated that the prevalence of prior Sindbis virus infection ranged from 10% to 67% in native adults in the Nile Delta region, and 18% in Anglo-Egyptian Sudan to the south, with a lower prevalence in children (Table 1.4A and Figure 1.6). Similarly, 31% of domestic mammals and 7% of birds tested in the Nile Delta region had antibodies to Sindbis virus (Table 1.4B). Thus Sindbis viral infection is fairly prevalent in the region, which is indicative of an established natural cycle between mosquitoes and vertebrates.
Table 1.4. Survey of Humans and Animals for Evidence of Prior Sindbis Virus Infection. A) Results of neutralization tests on human blood samples to detect the presence of antibodies against Sindbis virus, on natives organized by localities (Figure 1.6) and age group, and on foreigners who were all adults. * British soldiers were stationed in the Suez Canal near zone 7, and the only positive test was on a soldier who had been stations for over 18 months. † All of the Americans were NAMRU-3 staff in zone 9 and had been in the region for under 2 years at the time of testing. B) Similarly, neutralization test results on blood samples from domestic mammals and various birds in the Nile Delta region. From [43].
### A)

<table>
<thead>
<tr>
<th>Map No.</th>
<th>Locality</th>
<th>Less than 15 Yrs.</th>
<th>15 Yrs. and over</th>
<th>All ages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No. tested</td>
<td>No. Pos.</td>
<td>% Pos.</td>
</tr>
<tr>
<td></td>
<td><em>Nile Delta</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Idku</td>
<td>8</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>2</td>
<td>Baltim</td>
<td>10</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>Mansala</td>
<td>30</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>Kafr Mahallet Dawood</td>
<td>6</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>5</td>
<td>Gabares</td>
<td>1</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>6</td>
<td>Shiwa</td>
<td>21</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>Tel El Kebir</td>
<td>15</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>Sindbis area</td>
<td>48</td>
<td>9</td>
<td>19</td>
</tr>
<tr>
<td>9</td>
<td>Cairo area</td>
<td>4</td>
<td>0</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Total Nile Delta</td>
<td>143</td>
<td>20</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td><em>Sudan</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Khartoum</td>
<td>18</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>Nahud</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>Malakal</td>
<td>17</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>Mayen Mission</td>
<td>8</td>
<td>2</td>
<td>—</td>
</tr>
<tr>
<td>14</td>
<td>Juba</td>
<td>16</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Total Sudan</td>
<td>70</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td><em>Foreigners</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>British soldiers*</td>
<td>34</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Americans†</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Total all tests</td>
<td>213</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### B)

<table>
<thead>
<tr>
<th>Source</th>
<th>No. tested</th>
<th>No. pos.</th>
<th>% pos.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domestic quadrupeds</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cow</td>
<td>18</td>
<td>6</td>
<td>33</td>
</tr>
<tr>
<td>Gamoose (water buffalo)</td>
<td>59</td>
<td>24</td>
<td>41</td>
</tr>
<tr>
<td>Donkey</td>
<td>12</td>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td>Goat</td>
<td>26</td>
<td>14</td>
<td>54</td>
</tr>
<tr>
<td>Sheep</td>
<td>27</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Horse</td>
<td>26</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>Total quadrupeds</td>
<td>168</td>
<td>52</td>
<td>31</td>
</tr>
<tr>
<td>Avian species</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chicken</td>
<td>17</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Crow</td>
<td>15</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>Dove</td>
<td>15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Heron</td>
<td>29</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Pigeon</td>
<td>10</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Sparrow</td>
<td>39</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Total avian species</td>
<td>125</td>
<td>9</td>
<td>7</td>
</tr>
</tbody>
</table>
Figure 1.6. Map of Seroprevalence Survey and Discovery of Sindbis Virus and Other Agents of Recent Epidemics. A map of the Nile Delta and Anglo-Egyptian Sudan showing the places where human bloods were collected to test for Sindbis virus antibodies, indicating prior infection, using neutralization assays. The zoomed panel on the left represents the highlighted area around the Sindbis virus discovery site. The numbered zones correspond to the left column in Table 1.4A. Prevalence in neighboring regions of French Equatorial
Africa and Belgian Congo at the time was not tested, perhaps due to political inconvenience. The discovery of viruses that caused several recent global epidemics is also indicated, with the year of publication of the discovery report [40, 53, 54]. * Ebola virus is not an arbovirus and was not discovered through the arbovirus surveillance programs detailed above, nonetheless it did cause global fear during the epidemic that started in February 2014.

The live virus was never isolated from over 3000 blood samples taken from humans during a fever or visible illness in this study, and thus the pathology of the discovered virus was still unclear; "Nothing is known concerning the symptomatology of the infection in nature, either in man or in lower animals. It is as yet a virus without a disease. Except in infant mice and possibly a crow and a heron, experimental laboratory infection has not produced signs of disease" [43]. In the decades since, Sindbis virus has been found globally in North and South Africa, Europe, Asia and Australia, and was discovered to be the cause of several epidemics, including Ockelbo disease in Sweden, Pogosta disease in Finland and Karelian fever in Russia [55]. Clinical manifestations are usually benign, and severe acute symptoms include fever, rashes, joint pain in the ankle, knee, fingers, and wrist, and occasionally diffuse muscle aches. In the laboratory, as related to me by several older alphavirologists during their visits to New York, mouth pipetting accidents can cause canker sores for a couple of days [56]. During an epidemic in the early 1980's, 39% of people tested in Eastern Finland had antibodies to Sindbis virus - equivalent to the human seroprevalence in the Sindbis and Cairo areas in the early 1950's (Table 1.4A and Figure 1.6) - and there was a 1:17 ratio of
seroprevalence in Eastern Finland was highest in black grouse and wood grouse birds and approached 65%, indicating an established mosquito-bird transmission cycle. Thus Sindbis virus is a globetrotting infectious agent, not unlike its discoverers. However, genetic studies on the virus began before a link to disease had been established, and therefore the rise of these epidemics was not the driving factor behind the study of the virus.

1.4. Inception of Sindbis Genetics Using Temperature-Sensitive Mutants.

Some of the first studies of Sindbis virus after its discovery involved the generation of temperature sensitive (ts) mutants in the mid-1960s and subsequent analysis by Dr. Elmer Pfefferkorn and his graduate student, Boyce Burge, at Harvard University [58, 59]. Temperature differences provided a phenotypic difference that could be easily assayed at the time - Elmer Pfefferkorn had done his PhD thesis on bacteriophages, and ts mutants "have been of great value in understanding the physiology and genetics of bacteriophage T4" [60]. Consequently, studies of ts mutants of poliovirus, polioma virus, Newcastle disease virus, pseudorabies virus and a variety of bacteriophages were already underway in the virology field. These ts mutants have the common characteristic that due to a protein defect, they are unable to grow at a high, nonpermissive temperature but grow normally at a low, permissive temperature, in contrast to the parental strain that grows equally
well at both temperatures. Consequently, they can be easily assayed and differentiated from the parental strain by comparing plaque sizes of infections conducted in incubators set to different temperatures. Importantly, ts mutant stocks can be grown to high yields at the permissive temperature and subsequently its defect can be studied at the nonpermissive temperature [61].

To increase the effective assay range, a forward evolution experiment was conducted using Sindbis virus to generate a heat-resistant (HR) strain; "a mutant was selected by multiple cycles of heating at 60°C in complete medium and regrowing the survivors to provide another stock for heating.. blind selection through multiple cycles yielded a mutant (HR) that was substantially more heat stable than the wild type" [60] (Figure 1.7A). The HR strain can better resist heating to 60°C for 5 minutes, which is a prominently featured assay in these studies, but does not infect at temperatures much higher than 40°C, likely due to host cell problems at these higher temperatures (Figure 1.7B). The development of this HR strain, which has multiple differences from the strain originally isolated in the wild, is important to the field because most subsequent studies were conducted on this strain or on infectious clones based on this strain. Thus, unless explicitly stated to be derived from the AR339 strain, most strains referred to as "wild type" in Sindbis biology (including the Toto1101 strain used in our studies) are based on the HR strain.
Figure 1.7. Temperature Dependence of Sindbis Virus Strains. A) Heat inactivation of original Sindbis virus isolate (wild) and heat-resistant strain (HR) when incubated at 60°C. B) Virus yield as a function of incubation temperature of HR strain and two ts mutants. Infected CEFs were incubated for 8 hours at the specified temperatures and assayed. Virus titers between the HR strain and ts mutants are equivalent at 27°C and have a 4-log difference at 39°C. Adapted from [60].

The HR strain was then used as the parental strain to generate a variety of ts mutants. Cloned stocks of the HR strain were mutagenized with either nitrous acid, nitrosoguanidine or ethylmethane sulfonate, compounds that were previously shown to be effective in generating ts mutants in other viruses. In CEFs, these mutagenized stocks were grown at 27°C, and the virus in the agar over resulting plaques was grown at both 27°C and 39°C. Viruses that produced many plaques at 27°C and no plaques at 39°C were then plaque purified as ts mutant stocks (Table 1.5). Using these mutagenizing compounds increased the rate of ts mutant production from less than 0.3% to
3% of all plaques, allowing the isolation of about 40 $ts$ mutants. These compounds, however, tend to induce specific nucleotide mutation biases, which was realized upon DNA sequencing several decades later. Viral RNA synthesis at the nonpermissive temperature was tested by assaying radiolabeled uridine incorporation in the presence of actinomycin D, which inhibits cellular RNA synthesis but permits viral RNA synthesis. $ts$ mutants that do not produce RNA at 39°C were categorized as RNA minus (RNA-) mutants. $ts$ mutants that synthesized a substantial fraction of RNA as the parental HR strain were categorized as RNA plus (RNA+) mutants, and were thought to be defective in virus production for reasons other than the ability to produce functional RNA. Upon testing for reversion, about half of the $ts$ mutants with reversion frequencies above the average were discarded due to their inconvenience, in a rather unfortunate event. The remaining 23 $ts$ mutant isolates, labeled as $ts1$ through $ts24$ (skipping $ts3$ for reasons unknown) (Table 1.5) have formed the foundations of Sindbis genetics in the subsequent decades, and mutations present in these isolates represent many sites that have since been identified as critical to the viral life cycle [17, 61]. Perhaps a similar number of sites of equal or greater importance were present in the ~20 isolates that were sent to a biohazardous waste incinerator. They, of course, could not have predicted tools such as DNA sequencing and infectious clones that would be developed several decades later, which would have allowed the convenient identification and study of these mutants with high reversion
frequencies.

Table 1.5. Mutagenesis and RNA Synthesis of Generated Temperature-Sensitive Mutants of Sindbis Virus.  
a: Nitrous acid (HNO₂), ethylmethane sulfonate (EMS) and nitrosoguanidine (NTG) were used as mutagens to increase the frequency of ts mutant production.  
b: Reversion frequency is the ratio of plaque titer at non-permissive temperature (revertants) to titer at permissive temperature. ts mutants with reversion frequencies above these values were discarded.  
c: RNA synthesis, as a percent of that of the parental HR strain, measured using uridine-C¹⁴ incorporation between 2 and 4 hrs post-infection. Uninfected cells incorporate 2% uridine-C¹⁴ as HR strain infection, and represents the limit of detection.  
d: Infectious RNA titers were determined via plaques for a few mutants, ND = Not Done. RNA phenotype indicates whether the mutant synthesizes RNA at the non-permissive temperature.  
e: Mutants, named CL for conditional-lethal in previous publications, are renamed ts for temperature-sensitive. Adapted from [60].

<table>
<thead>
<tr>
<th>Mutant number</th>
<th>Mutagen used in isolation</th>
<th>Approximate reversion frequency</th>
<th>Stimulation of RNA synthesis at 39°C in the presence of actinomycin D²</th>
<th>Synthesis of infectious RNA at 39°C (% of HR value)</th>
<th>RNA phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>ts-1</td>
<td>HNO₂</td>
<td>1 × 10⁻⁴</td>
<td>&lt;1</td>
<td>0.2</td>
<td>-</td>
</tr>
<tr>
<td>ts-2</td>
<td>HNO₂</td>
<td>5 × 10⁻⁶</td>
<td>74</td>
<td>100</td>
<td>+</td>
</tr>
<tr>
<td>ts-4</td>
<td>HNO₂</td>
<td>2 × 10⁻⁴</td>
<td>&lt;1</td>
<td>0.05</td>
<td>-</td>
</tr>
<tr>
<td>ts-5</td>
<td>EMS</td>
<td>5 × 10⁻⁴</td>
<td>30</td>
<td>44</td>
<td>+</td>
</tr>
<tr>
<td>ts-6</td>
<td>NTG</td>
<td>1 × 10⁻⁶</td>
<td>&lt;1</td>
<td>ND</td>
<td>-</td>
</tr>
<tr>
<td>ts-7</td>
<td>NTG</td>
<td>&lt;2 × 10⁻⁶</td>
<td>&lt;1</td>
<td>ND</td>
<td>-</td>
</tr>
<tr>
<td>ts-8</td>
<td>NTG</td>
<td>1 × 10⁻⁶</td>
<td>&lt;1</td>
<td>ND</td>
<td>-</td>
</tr>
<tr>
<td>ts-9</td>
<td>NTG</td>
<td>&lt;2 × 10⁻⁶</td>
<td>33</td>
<td>47</td>
<td>+</td>
</tr>
<tr>
<td>ts-10</td>
<td>NTG</td>
<td>5 × 10⁻⁶</td>
<td>22</td>
<td>ND</td>
<td>+</td>
</tr>
<tr>
<td>ts-11</td>
<td>NTG</td>
<td>5 × 10⁻⁶</td>
<td>&lt;1</td>
<td>1.0</td>
<td>-</td>
</tr>
<tr>
<td>ts-12</td>
<td>NTG</td>
<td>4 × 10⁻⁶</td>
<td>&lt;1</td>
<td>ND</td>
<td>-</td>
</tr>
<tr>
<td>ts-13</td>
<td>NTG</td>
<td>1 × 10⁻⁴</td>
<td>39</td>
<td>ND</td>
<td>+</td>
</tr>
<tr>
<td>ts-14</td>
<td>NTG</td>
<td>1 × 10⁻⁴</td>
<td>&lt;1</td>
<td>ND</td>
<td>-</td>
</tr>
<tr>
<td>ts-15</td>
<td>NTG</td>
<td>&lt;2 × 10⁻⁶</td>
<td>&lt;1</td>
<td>ND</td>
<td>-</td>
</tr>
<tr>
<td>ts-16</td>
<td>NTG</td>
<td>2 × 10⁻⁶</td>
<td>&lt;1</td>
<td>ND</td>
<td>-</td>
</tr>
<tr>
<td>ts-17</td>
<td>NTG</td>
<td>1 × 10⁻⁴</td>
<td>&lt;1</td>
<td>2.0</td>
<td>-</td>
</tr>
<tr>
<td>ts-18</td>
<td>HNO₂</td>
<td>1 × 10⁻³</td>
<td>&lt;1</td>
<td>ND</td>
<td>-</td>
</tr>
<tr>
<td>ts-19</td>
<td>NTG</td>
<td>2 × 10⁻³</td>
<td>&lt;1</td>
<td>ND</td>
<td>-</td>
</tr>
<tr>
<td>ts-20</td>
<td>NTG</td>
<td>2 × 10⁻³</td>
<td>62</td>
<td>57</td>
<td>+</td>
</tr>
<tr>
<td>ts-21</td>
<td>NTG</td>
<td>2 × 10⁻⁵</td>
<td>&lt;1</td>
<td>ND</td>
<td>-</td>
</tr>
<tr>
<td>ts-22</td>
<td>NTG</td>
<td>2 × 10⁻⁶</td>
<td>&lt;1</td>
<td>ND</td>
<td>-</td>
</tr>
<tr>
<td>ts-23</td>
<td>HNO₂</td>
<td>5 × 10⁻⁶</td>
<td>43</td>
<td>47</td>
<td>+</td>
</tr>
<tr>
<td>ts-24</td>
<td>NTG</td>
<td>&lt;2 × 10⁻⁶</td>
<td>&lt;1</td>
<td>ND</td>
<td>-</td>
</tr>
</tbody>
</table>
After RNA synthesis categorization, these ts mutants were tested for degradation at 60°C relative to the original wild isolate and the parental HR strain (Figure 1.8A). Interestingly, the heat stability of RNA- mutants are clustered near the parental HR strain, and most RNA+ mutants are much less heat stable. They predicted that this was likely due to RNA+ mutants having defects in structural proteins that are incorporated and consequently alter the structure of the finished viral particle. In contrast, RNA- mutants are likely primarily defective in non-structural proteins involved in RNA synthesis, and the virus particles themselves should not be significantly different from the parental strain [60]. Burge and Pfefferkorn also theorized that since "two-thirds of the mutants are RNA-.. the cistrons yielding RNA+ mutants should outnumber those yielding RNA- mutants". If 'cistrons' in this context refers to relative gene size, then this prediction was validated after the development of DNA sequencing, as approximately two-thirds of the Sindbis genome encodes the non-structural polyprotein and one-third encodes the structural polyprotein. Burge and Pfefferkorn then began the process of categorizing these ts mutants in non-overlapping complementation groups, which "is defined as a set of mutants that do not complement each other but that do complement all other mutants in the population" [62]. Complementation level was detected by mixing two ts mutants and assaying the coinfection for an additive effect relative to individual infection by each ts mutant, which could indicate that the
defect in each complementing mutant was in a different gene (Figure 1.8A).

As a rule, all RNA+ mutants complement all RNA- mutants, and complementation groups should represent finer granulations on the genome. These complementation groups began the process of identifying the individual viral proteins and their functions [61].

![Figure 1.8. Heat Stability and Complementation of Sindbis ts Mutants.](image)

Figure 1.8. Heat Stability and Complementation of Sindbis ts Mutants. Heat inactivation of virus strains upon incubation at 60°C. A) The original Sindbis isolate (Wild), heat-resistant strain (HR) and RNA+ ts mutants are represented by solid lines. RNA- ts mutants are represented by dashed lines. B) Heat inactivation of ts5 and ts10 in individual infection and the progeny resulting from complementary coinfection, in comparison with HR strain. From [60, 62].

These ts mutants ushered the era of Sindbis genetics, and largely contributed to the authors' stated goal in developing conditional lethal mutants; "the hope that each complementation group with correspond to a specific biochemical
defect in the sequence of infection, the identification of which will lead to a more precise understanding of discrete events in viral multiplication" [62]. Another interesting prediction by the authors is that "the complementation between RNA- mutants is especially interesting because it implies that two cistrons are required for the synthesis of viral RNA. Since there is no agreement at present as to the number of enzymes required for viral RNA synthesis and since it is not certain that bacterial and animal RNA viruses will employ identical modes of RNA replication, our observation is open to several interpretations. Two enzymes may be essential, one to form a double-stranded "replicative form" from the input strand, and one to produce new progeny strands" [62]. This prediction also turned out to be largely correct, and the mechanism is even more elegant as explained below, with the 'replicative form' enzyme and 'progeny strand' enzyme being uncleaved and cleaved forms of the same polyprotein.

The development of these ts mutants by Elmer Pfefferkorn's laboratory provided the first genetic tools for the alphavirus field and led to Sindbis virus becoming the best-studied member of the group. In the subsequent decades, alphavirologists completed the complementation group map for the remaining ts mutants, and the development of Sindbis infectious clones and DNA sequencing allowed for precise characterization of ts mutations. These studies, described below, identified the genes in the viral genome, and many
sites within those genes that are critical for productive infections. The availability of many genetic tools and the number of studies using these tools established Sindbis virus as the 'prototype' alphavirus - the specific alphavirus in which most of the mechanisms that pertain to all alphaviruses were first elucidated and then expanded to other members of the group.

However, Sindbis virus is neither the first alphavirus discovered nor the most clinically relevant, so how did it become the 'prototype' alphavirus? How did one as-of-yet "virus without a disease" out of the hundreds of viruses newly discovered during arbovirus global surveillance efforts become one of the best studied? Elmer Pfefferkorn moved to Dartmouth University in the late 60's and focused on teaching medical school students for the next four decades, and won so many teaching awards that "a rule that students couldn't elect a previous winner until three years had passed [was established] to give other faculty a chance to win" [63]. In a 2008 interview in Dartmouth Medicine Magazine, he describes a conversation with Dr. John Enders, who had developed the measles vaccine and received the 1954 Nobel Prize in Physiology or Medicine for developing a poliovirus cell culture system. Upon completing his graduate studies on bacteriophages at Harvard University in 1959, Pfefferkorn recalls, "I went to one of the greatest virologists at Harvard — John Enders, and I told Dr. Enders that I intended to be a virologist and I was going to work on poliovirus. Wise John Enders said to me that there
[were] already too many people working on poliovirus. He suggested that it would be much wiser to work on a virus that no one knew anything about. All [Sindbis virus] had was a name, nothing was known about it. [John Enders] dug a sample out of his freezer and said that this would be an appropriate way to begin a career in virology. And he was dead right. It rapidly became a model system for studying animal viruses. Laboratories all over the world began working on this virus and building on the experiments we did" [63].

In the following years several additional catalogs of ts mutants were constructed using various mutagens for Sindbis virus and also for several other alphaviruses, starting with Semliki Forest virus (Table 1.6). Researchers soon realized that the same complementation groups with similar phenotypes arose on each of these sets, regardless of the alphavirus, mutagen or selection procedure, indicating that similar proteins or functions were likely being targeted. Perhaps several of the high-reversion mutants that were discarded by Burge and Pfefferkorn were recreated in subsequent catalogs [61]. Due to the head start, the first 23 Sindbis ts mutants generated by Burge and Pfefferkorn remain the best characterized, followed by ts mutants of Semliki Forest virus, which was the primary virus studied in most alphavirus laboratories located in Europe. Also perhaps due to the head start, Sindbis virus and Semliki Forest virus remain the best characterized alphaviruses, followed by chikungunya, eastern equine encephalitis, Venezuelan equine
encephalitis, and western equine encephalitis viruses [64].

Table 1.6. Temperature-Sensitive Mutants of Various Alphaviruses. ts mutants of Sindbis virus HR and AR339 strains [60, 65, 66], Semliki Forest virus [67, 68], western equine encephalitis virus [69], and eastern equine encephalitis virus [70] were generated using various mutagens in the late 1960's and 1970's, forming the foundation of alphavirus. Total number of ts mutants is tallied for catalogs generated using multiple mutagens. a: Nitrous acid (HNO₂), ethylmethane sulfonate (EMS), nitrosoguanidine (NNG), 5-azacytidine (AzaC), 5-fluorouracil (5-FU) and hydroxylamine (HA) were used as mutagens to increase the frequency of ts mutant production. b: The RNA± phenotype was somewhat arbitrarily defined as between 10% and 60% of RNA synthesis as the parental wild type virus at the non-permissive temperature, and the RNA+ phenotype above 60% and the RNA- phenotype under 10%. Adapted from [61].

<table>
<thead>
<tr>
<th>Virus</th>
<th>Selection procedure</th>
<th>Mutagen⁴</th>
<th>Number of mutants</th>
<th>Phenotype⁵</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sindbis virus, HR strain</td>
<td>Plaque efficiency at 30°C versus 40°C</td>
<td>HNO₂</td>
<td>5</td>
<td>2 RNA⁺, 3 RNA⁻, 1 RNA⁻</td>
<td>Burge and Pfefferkorn (1966a)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NNG</td>
<td>17</td>
<td>4 RNA⁺, 13 RNA⁻, 1 RNA⁻</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>EMS</td>
<td>1</td>
<td>1 RNA⁺, 1 RNA⁻, 1 RNA⁻</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>23</td>
<td>7 RNA⁺, 16 RNA⁻, 1 RNA⁻</td>
<td></td>
</tr>
<tr>
<td>Sindbis virus, HR strain</td>
<td>Plaque efficiency at 30°C versus 40°C</td>
<td>HNO₂</td>
<td>17</td>
<td>7 RNA⁺, 5 RNA⁻, 5 RNA⁻</td>
<td>Straus et al. (1976)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NNG</td>
<td>24</td>
<td>15 RNA⁺, 2 RNA⁻, 7 RNA⁻</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AzaC</td>
<td>35</td>
<td>12 RNA⁺, 16 RNA⁻, 7 RNA⁻</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5-FU</td>
<td>2</td>
<td>2 RNA⁺, 1 RNA⁻, 2 RNA⁻</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>78</td>
<td>34 RNA⁺, 25 RNA⁻, 19 RNA⁻</td>
<td></td>
</tr>
<tr>
<td>Sindbis virus, HR strain</td>
<td>Plaque efficiency at 30°C versus 37°C</td>
<td>AzaC</td>
<td>32</td>
<td>7 RNA⁺, 23 RNA⁻, 2 RNA⁻</td>
<td>Straus (unpublished observations)</td>
</tr>
<tr>
<td>Sindbis virus, AR339 strain</td>
<td>Plaque enlargement technique</td>
<td>HNO₂</td>
<td>20</td>
<td>15 RNA⁺, 5 RNA⁻</td>
<td>Atkins et al. (1974a)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NNG</td>
<td>19</td>
<td>17 RNA⁺, 2 RNA⁻</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5-FU</td>
<td>15</td>
<td>4 RNA⁺, 9 RNA⁻</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>EMS</td>
<td>23</td>
<td>3 RNA⁺, 20 RNA⁻</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>HA</td>
<td>27</td>
<td>5 RNA⁺, 14 RNA⁻</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>104</td>
<td>12 RNA⁺, 75 RNA⁻, 17 RNA⁻</td>
<td></td>
</tr>
<tr>
<td>Semliki Forest virus, strain 25639</td>
<td>Plaque efficiency at 29°C versus 38.5°C</td>
<td>5-FU</td>
<td>11</td>
<td>1 RNA⁺, 8 RNA⁻, 2 RNA⁻</td>
<td>Tan et al. (1969)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HA</td>
<td>13</td>
<td>4 RNA⁺, 3 RNA⁻, 6 RNA⁻</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>NNG</td>
<td>14</td>
<td>11 RNA⁺, 2 RNA⁻, 1 RNA⁻</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>38</td>
<td>16 RNA⁺, 13 RNA⁻, 9 RNA⁻</td>
<td></td>
</tr>
<tr>
<td>Semliki Forest virus</td>
<td>Cytopathic effect at 27°C versus 30°C</td>
<td>NNG</td>
<td>16</td>
<td>6 RNA⁺, 8 RNA⁻, 2 RNA⁻</td>
<td>Keränen and Kääriäinen (1974)</td>
</tr>
<tr>
<td>Western equine encephalitis virus</td>
<td>Cytopathic effect at 37°C versus 41.5°C</td>
<td>NNG</td>
<td>10</td>
<td>1 RNA⁺, 8 RNA⁻, 1 RNA⁻</td>
<td>Hashimoto and Simizu (1978)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HNO₂</td>
<td>3</td>
<td>2 RNA⁺, 1 RNA⁻</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td>1 RNA⁺, 10 RNA⁻, 2 RNA⁻</td>
<td></td>
</tr>
<tr>
<td>Eastern equine encephalitis virus</td>
<td>Plaque efficiency at 30°C versus 42°C</td>
<td>HNO₂</td>
<td>1</td>
<td>1 RNA⁺, 1 RNA⁻</td>
<td>Zebovitz and Brown (1970)</td>
</tr>
</tbody>
</table>
1.5. Discovery of Components of Sindbis Virus Virions.

The first studies on Sindbis virus in Elmer Pfefferkorn's laboratory found that Sindbis virus particles contained 61% protein, 27% lipids, 6.5% carbohydrates and 5.5% RNA, indicating that it was likely an enveloped virus with glycoproteins - proteins with covalently linked sugars [71]. Various electron microscope studies between the mid 1960s and early 1970s confirmed that the ~50 nm virus particle contains an electron-dense protein nucleocapsid, surrounded by a lipid bilayer (envelope) with 7 nm 'spike' proteins attached (Figure 1.9 A) [64]. These 'spikes' were essential for infectivity as treatment with proteases created smoothened the particles and made them non-infectious (Figure 1.9 B) [72]. Since viruses are obligate intracellular parasites, they contain an intracellular replication phase and an extracellular particle phase. The ease of isolation of the extracellular virus particles (i.e., virions) led to the preferential characterization of its components in the 1960s and 1970s: “Production of α-virus particles on a milligram scale for morphologic, chemical, and physical studies is a relatively simple task. The virus is released from the infected cells by budding through the plasma membrane (Figure 1.9 D) and can then be recovered from the culture medium. The viral envelope can be disrupted with mild detergents and the nucleocapsid (Figure 1.9 C) separated from the solubilized envelope components by sucrose gradient centrifugation. Treatment of the virion with strong detergents, such as sodium dodecyl sulfate [SDS], causes a total
disintegration of the particle, making it possible to [separate and] study the proteins and RNA” [64].

Figure 1.9. Electron Microscopy of Sindbis Virus. Negative staining of the (A) full Sindbis virion, (B) the virion with the envelope glycoprotein projections (spikes) removed by treatment with bromelain protease, and (C) the nucleocapsids with the envelope removed by treatment with detergents. The arrow indicates a particle into which the negative strain has penetrated clearly showing a bright membrane layer. D) Freeze-etching of Sindbis virions budding from the plasma membrane of infected CEFs. Bars represent 100 nm. (A,B) was stained by Richard Compans at Rockefeller University, (C) by Carl-Henrik von Bonsdorff and Stephen Harrison at Harvard University, and (D) by Dennis Brown in Elmer Pfefferkorn's laboratory [72, 73]. Adapted from [64, 74].

Regular SDS polyacrylamide gel electrophoresis (SDS-PAGE) only showed a single band in the nucleocapsid fraction and in the envelope fraction, and thus
initially the Sindbis virions were thought to contain one capsid protein (C) and one envelope protein (E), representing the spikes. The envelope fraction caused hemagglutination, indicating that at least a fraction of the envelope proteins were glycosylated. Dr. James (Jim) Strauss and Boyce Burge, in the laboratory of Dr. James (Jim) Darnell at Albert Einstein College of Medicine in 1968, found that a mole of envelope proteins on average contained 17 mannose and galactose, 19 glucosamine, 1-2 sialic acid and 2 fucose monosaccharide molecules when the virus is grown in vertebrate cells [75]. These sialic acid molecules on envelope proteins are generally thought to bind to sialic acid receptors on the surface of red blood cells to induce hemagglutination, but selective removal of sialic acid sugars did not prevent infectivity or hemagglutination, and thus the mechanism of red blood cell binding is unclear [64]. Drs. Milton Schlesinger and Sondra Schlesinger at Washington University in St. Louis, along with Boyce Burge who was then at MIT, in 1972 elegantly showed that there were two distinct sugar-containing envelope proteins in the virions using a discontinuous electrophoresis method, which were named E1 and E2 (Figure 1.10) [76]. Further characterization revealed that the hemagglutinating activity exclusively resided on E1 [77]. Treatments with cross-linking and proteolytic agents showed that in the virion, E1 and E2 form dimers and cross through the envelope with a hydrophobic, transmembrane region which then interacts with the nucleocapsid, and these findings were later confirmed with X-ray Crystallography [17, 64].
Figure 1.10. Resolution of the Structural Proteins of the Sindbis Virion. Radiolabeling of amino acids and using a discontinuous gel method which involved a stacking gel to increase resolution allowed for separation of two envelope proteins from the virion. Radiolabeling of glucosamine showed that both envelope proteins are glycosylated, unlike in several other virus families, and that the capsid protein was not. Adapted from [76].

The first evidence that the alphavirus genome was composed of RNA came from studies in the late 1950s using western equine encephalitis virus, which was then expanded to other alphaviruses [78]. The aqueous phase after mixing heated phenol with purified virions induced death when injected into the allantoic cavity of embryonated chicken eggs. This infectivity was removed when the RNase was added to this aqueous phase, but was unaffected by the
addition of DNase. This was one of the first studies in the animal virology field that showed that viral RNA itself can be infectious, without any parental protein component, and indicated that it must serve as a messenger template for a polymerase [64]. Initially the Sindbis RNA genome was thought to be segmented due to RNase contamination in experiments, but later studies with careful purification of virions consistently yielded one gel band and confirmed that Sindbis contains an unsegmented RNA genome [79]. Biochemical analyses yielded asymmetric base compositions of 29% adenine, 20% uracil, 26% guanine and 25% cytosine, indicating that the genome was single-stranded [71]. The virion, after the envelope is removed, is susceptible to low concentrations of RNase which added further evidence that the genome is single-stranded and also provided the first clues that it was plus stranded, according to David Baltimore's classification; "the 'plus' strand has the base sequence of messenger RNA and the 'minus' strand is complementary to it" [74]. Minus-stranded RNA viral genomes in virions are typically covered in proteins and consequently less susceptible to RNases. Subsequent identification in the genome of a polyA sequence by Drs. Robert Johnston and Henry Bose in 1972 and a 5' inverted 7-methylguanosine cap structure by Dr. Victor Stollar's laboratory in 1976 added further evidence that the virion RNA is a plus strand [80, 81]. The envelope prevents susceptibility to RNases, indicating that perhaps one role of the envelope is to protect the genome from enzymatic activity.
The above observations allowed the assignment of the three RNA+ ts mutant complementation groups to the three structural proteins in the virion well before the advent of DNA sequencing. Group D mutants do not hemagglutinate at the non-permissive temperature (39°C), and thus contain defects in E1. In certain mutants the infected cells themselves do not hemagglutinate at 39°C, indicating at E1 is not present in the cell membrane at this temperature. Nucleocapsids are made but freely float in the cytoplasm at 39°C, indicating that insertion of E1 into the plasma membrane is a prerequisite for localization of nucleocapsids at the plasma membrane. Additionally, processed E2 is absent at 39°C, indicating that functional E1 is required for the processing of E2 from its precursor. Group E mutants similarly contain a processing defect in E2 but hemagglutinate and localize nucleocapsids normally at 39°C, indicating that their E1 protein is normal. Thus by the process of elimination, group E mutants contain a defect in E2 [61].

Group C mutants generally do not form nucleocapsids at 39°C, and thus are defective in the capsid protein. Some mutants do produce capsid proteins of altered mobility in electrophoresis gels. The mutants that do not produce nucleocapsids at 39°C were found to accumulate a large protein within the cell that was named the ts2 protein after the mutant in which it was first discovered.
These mutants hemagglutinate weakly, indicating that little E1 is at the plasma membrane, and this was the first indication that the ts2 protein is a precursor for the structural proteins, and was later renamed as the structural polyprotein upon confirmation. Complementation of group C mutants with either group D or E mutants reduced the amount of this precursor and increased hemagglutination, providing the first evidence that a functional capsid molecule contains protease activity [61]. Intracellular protein gels in initial studies were not able to resolve the precursor protein [74]. The availability and study of ts mutants greatly contributed to understanding various aspects of these structural proteins [61, 64, 74].

1.6. Identification of Sindbis Viral Components During Intracellular Infection.

The majority of alphavirus virions added to CEFs are adsorbed into the cell within the first half hour. Analysis of virions that cannot be adsorbed into cells revealed that there are about 10,000 receptor sites on the surface of cells, the identity of which are still unknown. Upon adsorption, CEFs at 37°C begin to reveal severe cytopathic effects and the majority die within a day. Translation of cellular proteins is potently inhibited within 3-5 hours post-infection (p.i.), and viral proteins are easier to resolve after this time by analyzing the incorporation of radiolabeled amino acids. Equilibrium nucleotide radiolabeling experiments revealed that plus strand RNA production enters exponential phase between 2 and 5 hours p.i. after which it remains fairly constant, and
values can reach 200,000 per cell by 8 hours p.i., with a depletion of cellular nucleotide pools in parallel. Each cell can release up to 20,000 progeny virions, which represents up to 10% of the total phospholipid content of the plasma membrane. These processes are slower and the maxima are lower in alphavirus infection of cultured mosquito cells at 28°C, and invertebrate cells generally do not die from infection [64].

The discovery in 1961 that actinomycin D - an antibiotic and chemotherapeutic drug extracted from *Streptomyces parvullus* bacteria - shut off cellular RNA production but allowed viral RNA production to continue uninhibited enabled researchers to resolve viral RNA on electrophoresis gels [82]. The addition of radiolabeled RNA components, such as tritiated uridine, upon actinomycin D treatment would only label newly produced viral RNA allowing subsequent analysis. Three forms of RNA was first discovered in Semliki Forest virus infected cells in 1967; (i) a genomic RNA with a sedimentation coefficient of 42S similar to virion RNA, (ii) a subgenomic RNA with a sedimentation coefficient of 26S, and (iii) a heterogenous RNA species that sediments between 18S and 22S and was assumed to be the double stranded RNA (dsRNA) intermediate due to RNase resistance ([Figure 1.11 A] [83]). Hybridization experiments showed that the 26S RNA is identical to a third of the genomic RNA, and hence subgenomic, and has the same plus-strand polarity, which was confirmed by the discovery of a 5' cap and a polyA
The 42S genomic RNA was infectious to chick embryos and the 26S subgenomic RNA was not (Figure 1.11 A).

**Figure 1.11. Intracellular Alphavirus RNAs and Structural Proteins.** CEF cells were treated with actinomycin D, labeled with $^{14}$C-uridine and infected with Semliki Forest virus for 7 hours. Extracted cellular RNA was combined with chick ribosomal RNA and sedimented through a sucrose gradient, and assayed for total radioactivity (●) and RNase-resistant radioactivity (○) (left y-axis), infectivity in check embryos (△) and optical density at 260 nm (—). B) BHK cells were infected with Sindbis virus for 8 hours and labeled with $^{35}$S-methionine for 10 mins and chased for 20 mins (30' lane) or labeled for 150 mins (150' lane). Portion above the arrow was exposed for 5 days, and the portion below for 19 days. B represents the uncleaved structural protein precursor, 9.8K represents E3, and 4.2K represents the 6K protein. Adapted from [83, 86].

The subgenomic RNA was determined to be the messenger for the structural polyprotein as only the 26S RNA produced high amounts of structural proteins.
in cell-free wheat germ extract translation systems. These were produced as a 1:1 molar ratio of capsid and a polyprotein with all envelope proteins in rabbit reticulocyte translation systems, giving the first evidence that the capsid protein contains an autoprotease. Further processing into the various individual envelope proteins only occurred upon the addition of endoplasmic reticulum membrane fractions, indicating that cellular proteases were responsible for the remaining cleavage events [64]. Additional evidence further confirmed that the structural proteins are made as a polyprotein and successively cleaved during cellular infection; (i) trypsin digests of the polyprotein in ts2 produce the same fragments as found in all the structural proteins combined, (ii) radiolabeled formylmethionine, which is exclusively used for initiation, was only incorporated into one tryptic fragment indicating a single translation initiation event, and (iii) sequential radiolabeling always labeled the proteins in the same order [64].

A new protein with a higher molecular weight than E2 was identified in infection of baby hamster kidney (BHK) cells, which was difficult to detect in infections of CEFs, likely due to differences in the rates of cleavage events by cellular proteases. Maps using similar trypsin digests showed that this protein contained E2 and was named precursor to E2 (PE2) [87]. The cleavage of PE2, in addition to E2, produces the E3 protein which was initially found in the virions of Semliki Forest virus, but remains intracellular in Sindbis virus [88].
The search for E3 during Sindbis virus infection of BHK cells, by Drs. William Welch and Bartholomew Sefton at the Salk Institute, yielded the discovery of another small intracellular cleavage product that arose when mature E1 was processed, which was 4,200 daltons in size (Figure 1.11 B) [86]. Welch and Sefton then searched for a similar protein in Semliki Forest virus infection and discovered one 6,000 daltons in size. Hereafter all analogous proteins in alphaviruses, which vary considerably in size, have been named the '6K protein' [89]. The 6K name is retained even today, unlike the other viral proteins in which the original size-based nomenclature was changed upon the discovery of a location or function, perhaps because a precise function for 6K still has not been assigned and the entire protein can be deleted with no severe effect in cell culture infection [90, 91]. Individual amino acid mutations in 6K do decrease viral fitness, and the virus 'ejects' inserted transgenes such as fluorescent proteins within a few passages with ease, therefore the protein likely serves a function critical to the virus life cycle in nature. More recently, bioinformatic analysis of conserved regions within 6K in 2008 by Andrew Firth et. al. at University College Cork in Ireland led to the discovery of a frameshift translation product, the transframe (TF) protein, and its functions are still being elucidated [92]. Contemporary information can be found in the 2017 review, "Disentangling the frames, the State of Research on the Alphavirus 6K and TF Proteins", by Dr. Tuli Mukhopadhyay's laboratory at Indiana University in Bloomington [93]. Thus all alphavirus structural proteins aside from TF were
discovered and fairly well characterized, and ts mutant complementation
groups were assigned, prior to sequencing the virus. Direct protein sequencing
efforts began on fragments of these proteins, but was largely superseded by
DNA sequencing [94].

Short radiolabeling pulses showed that the heterogenous mixture of dsRNA
that sedimented around 20S contained a dsRNA core and multiple nascent
single strands that were being actively transcribed [64]. Pre-treatment of the
cells with interferon blocked the formation of these dsRNA cores. Studies of
this pool of dsRNA by Daniel Simmons in 1972, in the newly started laboratory
of Jim Strauss at Caltech, revealed a novel transcriptional mechanism utilized
by alphaviruses (Figure 1.12 C) [95]. RNA that was radiolabeled and purified
from Sindbis virus infection of BHK cells, treated with RNase and then run on
a polyacrylamide-agarose gel yielded three distinct bands of dsRNA (under
the assumption that all single stranded RNA would have been degraded by
the RNase) (Figure 1.12 B). These dsRNA species that remain after RNase
treatment were named as replicative forms (RF) I, II and III. RFI had twice the
molecular weight as the genomic RNA, and was thus determined to be the
replicative intermediate (RI) of genomic RNA (Figure 1.12 C, RIa). RFII and
RFIII had the same molar ratio and were thus considered to come from the
same replicative intermediate, which must produce subgenomic RNA since
RFIII had twice its molecular weight (Figure 1.12 C, RIb). Short radiolabeling
pulses only showed RFI and RFIII, indicating that the genomic and subgenomic RNA's are made in far excess of the RNA that is in RFII (Figure 1.12 A). This suggested that in the subgenomic RNA intermediate, perhaps the polymerase stalls at the end of RFII, and an open single-stranded region in the minus strand (that is accessible to experimental RNase cleavage) allows for polymerase binding and preferential initiation of nascent subgenomic RNA in RFIII (Figure 1.12 C). "Therefore, there must be an active and direct control mechanism which activates the synthesis of only part of the viral genome", claim Simmons and Strauss predicting the subgenomic promoter that would be later revealed by DNA sequencing [95].

**Figure 1.12. Intracellular Alphavirus Double-Stranded RNAs.** A & B) BHK cells were infected with Sindbis virus for 8 hours and pulsed with 5-3H-uridine for 1 min (A) or 2 mins (B). RNAs were extracted, treated with mild RNase (10 μg/ml) for 15 mins at 37°C, sedimented in a sucrose gradient and assayed for
radioactivity. Short labeling pulses show that RFIs (dsRNAs of genomic RNA) and RFIIIs (dsRNAs of subgenomic RNA) are produced at a much higher rate than RFII. C) Schematic of the production of genomic (42S) and subgenomic (26S) from two replicative intermediates, RIₐ and RIₕ respectively, which were revealed by the production of three replicative forms upon mild RNase treatment. Circles represent viral polymerases. Adapted from [64, 95].

Importantly, the convenience of being able to use electrophoresis gels of RNase-resistant molecules to determine the relative amounts of genomic and subgenomic RNAs, which in turn represented the production of the RNA and protein components of the virion respectively, enabled the study of viral transcriptional regulation mechanisms and the potential involvement of viral proteins. The amount of minus strand RNA involved in its primary function of transcribing the two plus strand RNAs can be estimated by its necessitated inclusion in these replicative forms. Carefully controlling for RNase degradation of single strands, the ratios of the plus strand RNAs involved in transcription complexes versus performing another function, such as translation or encapsidation for genomic RNA or translation for subgenomic RNA, can be estimated by the relative amount of their respective double-stranded and single-stranded forms. Studies of various RNA- ts mutants showed disregulation of this ratio of genomic to subgenomic RNAs at the non-permissive temperature (39°C). When Sindbis ts4 is first grown at the permissive temperature (28°C) and then shifted up to 39°C, the synthesis of the subgenomic RNA stops and radiolabels shift from RFII and RFIII to RFI, indicating that minus strands that were templates for subgenomic RNA were
shifted to being templates for the genomic strand. This effect is reversible, and such temperature shift-up and shift-down experiments showed that the synthesis of subgenomic RNA and consequently the production of structural proteins was a late function in infection [61].

The early functions, which were presumed to include the production of a viral polymerase that creates the first minus strands from the plus strands that entered the cell, were more difficult to assay as nonstructural proteins were not easily visible in gels at the time [64]. These were found in the mid-1970's by first detecting the precursor polyprotein during temperature shift-up experiments using Semliki Forest virus ts1 in Dr. Levi Kääriäinen's laboratory at the University of Helsinki, Finland (Figure 1.13), and by using Sindbis virus ts21 and ts24 in the Schlesinger laboratory [96, 97]. Trypsin digest maps confirmed that this protein was different than the structural polyprotein, and the formation of this protein was accompanied by a large increase in the ratio of genomic to subgenomic RNA in both viruses, yielding the first evidence that the processing of the nonstructural polyprotein may affect these RNA ratios. Various radiolabeling strategies showed that upon temperature shift-down, these proteins were processed in steps into four individual polyproteins, with the fourth protein deduced from the subtraction of the size of the third protein from its precursor; "The existence of the fourth nonstructural protein has been difficult to show in either virus, due to its migration at the position of PE2.. the
[difference] strongly suggests that a fourth nonstructural protein exists, and thus we should be able to find the whole nonstructural protein in the future" [64].

Figure 1.13. Translation of Nonstructural Proteins from Alphavirus Genomic RNA. A) The 5' segment of genomic RNA is translated into a polyprotein (I), which is eventually cleaved into four fully processed nonstructural proteins (II) through various detected cleavage intermediates (III). Numbers correspond to sizes detected using Semliki Forest virus. B) PAGE gel of Semliki Forest virus ts1 infection - which overproduces the nonstructural proteins - showing viral structural proteins (C, E1 and p62 which represents PE2), processed nonstructural proteins (ns70, ns72 and ns86), and nonstructural cleavage intermediates (ns155 and ns135). Ns135 is cleaved to ns72 and ns60, which is obscured by p62. Adapted from [64].

This nicely corresponded to the existence of four RNA-complementation groups, however the overlap of phenotypes made an assignment to the nonstructural proteins rather nebulous, and required the use of DNA sequencing. ts4 from group A and ts6 from group F both did not form RNase-
resistant dsRNA at the 39°C. ts11 from group B formed a large molecular weight precursor similar to ts21 and ts24 from group A at 39°C, albeit at a smaller molecular weight, representing an overlap of polyprotein processing deficiencies [61]. All RNA- ts mutants analyzed showed a continuum of altered genomic:subgenomic ratios at 39°C, from much lower to much higher than the parental strain, which defied neat categorization into complementation groups [65]. In 1980, on the precipice of the era of DNA sequencing, Jim and Ellen Strauss stated that; "Genetic analysis of the alphaviruses has been an important tool for studying viral replication, and many of the details of the virus replication cycle have been obtained through the use of [ts] mutants. This is particularly true of the late events in the virus life cycle, production of the structural proteins and virus budding, which have used RNA+ groups. It is to be expected that a more complete characterization of the RNA- mutants will lead to a better understanding of the early events in the viral replication cycle and of the RNA transcription events. Yet.. these studies.. have not revealed.. genetic control elements, such as repressors, operators, and promoters.. it is possible that the complete sequencing of the RNA of an alphavirus will reveal complexities of genetic organization which have been hitherto undetectable" [61].

1.7. Sequencing of Sindbis virus and ts Mutant Lesions and Elucidation of Nonstructural Protein Functions.

The subgenomic RNA of the Sindbis HR strain was first sequenced in 1981,
followed by the genomic RNA of the HR strain in 1984, by Ellen Strauss, Jim Strauss and (future) Dr. Charles (Charlie) Rice, a graduate student in the Strauss laboratory at the time [98, 99]. The viral RNA was incubated with avian myeloblastosis virus reverse transcriptase to produce a cDNA-RNA hybrid, the RNA was hydrolyzed with NaOH, and the cDNA was digested with a variety of restriction enzymes. The digested cDNA fragments were then sequenced using the Maxam and Gilbert method and assembled together, with multiple coverage of the genome provided by overlap between the restriction digest fragments, somewhat analogous to contemporary next-generation sequencing assembly methods [100]. The start sites of the four nonstructural proteins (nsP's) were confirmed in comparison with direct protein sequencing of N-terminal fragments and named in series from the 5' end of the genome; nsP1, nsP2, nsP3 and nsP4 [99, 101]. The characterization of ts mutant lesions began by similarly reverse transcribing cDNA from the RNA of various HR strain ts mutants, using "the oldest stocks in our possession, which had not been passed since 1971, as seed stocks in this work", to minimize reversions or further adaptive mutations arising in the population [102]. The cDNAs were similarly digested and the fragments with the suspected ts lesion were inserted into the Escherichia coli plasmid vector pBR322, and the DNA region was sequenced upon bacterial amplification [102–105].

Instead of such individual fragments, the insertion of a cDNA copy of the entire
virus genome into this plasmid, from which infectious RNA can be transcribed as had been done with several bacteriophages and plant and animal viruses, would provide the field with numerous advantages. Infectious clones are DNA plasmids that contain the virus sequence that can (i) be amplified with high accuracy, (ii) be in vitro transcribed into the viral RNA and transfected into a cell to produce infectious virus, (iii) manipulated using recombinant biology tools that exist for DNA and not for RNA, and (iv) be stored and transported much more easily than RNA or virions. According to Dr. Henry Huang, "the idea of making an infectious cDNA clone of [Sindbis virus] was conceived in 1980. It took a while to get it to work" [106]. The first Sindbis infectious clone was made in 1987 by Charlie Rice, Robin Levin and Henry Huang, then at Washington University in St. Louis, along with Jim Strauss [107]. Viral RNA with the best infectious performance was derived from a mix of cDNA fragments from the HR strain and HR strain small plaque variant (HRsp), which includes the Toto1101 strain used in our next-generation sequencing studies (chapter 3). Select features of the pToto1101 plasmid include the bacterial origin of replication and an ampicillin resistance cassette for efficient amplification in E. Coli, the virus sequence with an SP6 promoter at the 5' end for efficient in vitro RNA transcription, and a Xho1 'runoff' site at the 3' end of the viral sequence such that the plasmid can be linearized and the SP6 enzyme will fall off upon transcribing only the virus sequence. Specific mutations can be introduced into the virus and characterized for the first time;
"cDNA clones may be mutagenized by any of a number of [recombinant DNA] methods to generate transcripts with novel mutations, i.e., lethal, viable or conditional, which can be studied both in vivo and in vitro. The characterization of these mutants, in conjunction with the traditionally derived mutants, will further our understanding of the molecular biology of Sindbis virus" [107].

Figure 1.14. Mapping of ts2 Lesion Using cDNA Infectious Clone. Comparison of proteins made during infection by passaged virus and infectious clones at 30°C ('30' lanes, permissive for ts2) and 40°C ('40' lanes, non-permissive for ts2). CEF cells were pulse-labeled for 1 hr with 35S-methionine at either 6 or 7 hrs p.i. and lysates were resolved on a SDS-PAGE gel. Y-axis: labeled structural proteins, ts2 protein represents the uncleaved structural polyprotein precursor. X-axis: HRsp and ts2 represent infections by the original passaged virus stocks, and the remainder are infections by virus that originated from transfected RNA in vitro transcribed from cDNA infectious clones. 10 and
1000 represent Toto series infectious clones combined from HR and HRsp strain sequences, and ts2.1 represents the Toto1000 infectious clone with the inserted ts2 lesion. Adapted from [107].

The authors then inserted the ts2 lesion, which prevents cleavage of the capsid protein and accumulates the structural polyprotein at the non-permissive temperature; "To demonstrate the utility of this approach for mapping ts mutations, we replaced the capsid sequences of [the infectious clone] with cDNA containing the ts2 mutation. The virus stock derived from these infectious transcripts was clearly temperature sensitive [and produces] a prominent species of about 130 kDa characteristic of ts2 and with greatly diminished quantities of the cleaved structural proteins" (Figure 1.14) [107]. Sequencing the infectious clone of the ts2 plasmid and a revertant to the wild type confirmed that mapped mutation alone induced the cleavage-defective phenotype. The majority of the Sindbis virus ts mutants originally isolated by Burge and Pfefferkorn were similarly mapped via efforts by the Strauss laboratory and by the laboratory of Drs. Dorothea Sawicki and Stanley Sawicki at the University of Toledo College of Medicine and Life Sciences (Table 1.7) [108–111].
Table 1.7. Genomic Lesions in Temperature-Sensitive Mutants of Sindbis Virus. Sequence characteristics of all mapped mutations in Sindbis virus HR strain ts mutants originally isolated by Burge and Pfefferkorn, aligned to Toto1101 strain. ts2 and ts5 were independent isolations of the same mutation. Suppressor: ts9 failed to complement with any mutant and suppressed their yields, n.t.: not tested, n.s.: not sequenced as of the writing of this chapter. * denotes ts mutants with multiple nucleotide mutations, and mutations in black were shown to contribute to ts phenotype (these mutations either induced the ts phenotype when inserted in isolation into an infectious clone or assayed, and/or a reversion at that position removed the ts phenotype); e.g. ts7 contains three amino acid mutations in two separate proteins, two of which in isolation confer a ts phenotype, ts24 contains three amino acid mutations in three proteins, but only two in isolation confer a ts phenotype. ‡: Detected mutations which did not confer ts phenotype (in grey). † ts10 contains both ts phenotype-conferring nucleotide mutations within the same codon resulting in one amino acid mutation, and revertants contained a single G10590A mutation to an arginine codon. # The two ts phenotype-conferring mutations in ts23 were not tested in isolation, unlike all other mutations shown to confer a ts phenotype, but both are thought to be contributors since revertants always had back-mutations in both sites. § Additional mutations in ts23 that were found in a passaged strain in 1981 but not in a frozen strain from 1969, indicating that they are likely spontaneous mutations that arose during passaging. Note that whereas the mutations which are major contributors to the ts phenotype have been mapped, in no case has the entire genome been sequenced, so additional mutations which may be minor contributors or may have spontaneously arisen and do not contribute to the ts phenotype may exist in these mutants. Sequence data compiled from [102–104, 108–111].
<table>
<thead>
<tr>
<th>Mutant</th>
<th>Comp. Group</th>
<th>Detected Mutations</th>
<th>Protein</th>
<th>Nucleotide Mutation</th>
<th>Amino Acid Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ts1</td>
<td>n.t.</td>
<td>n.s.</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ts2</td>
<td>C</td>
<td>1</td>
<td>Capsid</td>
<td>C8298 → G</td>
<td>Pro-218 → Ser</td>
</tr>
<tr>
<td>ts4</td>
<td>A</td>
<td>1</td>
<td>nsP3</td>
<td>C4903 → U</td>
<td>Ala-268 → Val</td>
</tr>
<tr>
<td>ts5</td>
<td>C</td>
<td>1</td>
<td>Capsid</td>
<td>C8298 → G</td>
<td>Pro-218 → Ser</td>
</tr>
<tr>
<td>ts6</td>
<td>F</td>
<td>1</td>
<td>nsP4</td>
<td>G6226 → A</td>
<td>Gly-153 → Glu</td>
</tr>
<tr>
<td>ts7*</td>
<td>G</td>
<td>3</td>
<td>nsP2</td>
<td>G3243 → A</td>
<td>Asp-522 → Asn</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>nsP3</td>
<td>U5035 → A</td>
<td>Phe-312 → Ser</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A4752 → G</td>
<td>Lys-218 → Glu†</td>
</tr>
<tr>
<td>ts8</td>
<td>n.t.</td>
<td>n.s.</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ts9</td>
<td>Suppressor</td>
<td>n.s.</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ts10*</td>
<td>D</td>
<td>8</td>
<td>E1</td>
<td>10590 + 10591 AAG → GGG</td>
<td>Lys-176 → Gly†</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E1</td>
<td>U10469 → A</td>
<td>Silent: Ile-135‡</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E1</td>
<td>A10599 → G</td>
<td>Ile-179 → Val‡</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E1</td>
<td>A10692 → G</td>
<td>Ser-210 → Gly‡</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E1</td>
<td>U10773 → G</td>
<td>Ser-237 → Ala‡</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E2</td>
<td>C9010 → U</td>
<td>Ala-127 → Val‡</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E2</td>
<td>C9809 → U</td>
<td>Silent: Arg-393‡</td>
</tr>
<tr>
<td>ts11</td>
<td>B</td>
<td>1</td>
<td>nsP1</td>
<td>G1101 → A</td>
<td>Ala-348 → Thr</td>
</tr>
<tr>
<td>ts12</td>
<td>n.t.</td>
<td>n.s.</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ts13</td>
<td>C</td>
<td>1</td>
<td>Capsid</td>
<td>A8059 → U</td>
<td>Lys-138 → Ile</td>
</tr>
<tr>
<td>ts14</td>
<td>A</td>
<td>1</td>
<td>nsP2</td>
<td>G2166 → A</td>
<td>Glu-163 → Lys</td>
</tr>
<tr>
<td>ts15</td>
<td>A</td>
<td>n.s.</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ts16</td>
<td>A</td>
<td>1</td>
<td>nsP2</td>
<td>G2502 → A</td>
<td>Val-275 → Ile</td>
</tr>
<tr>
<td>ts17</td>
<td>A</td>
<td>1</td>
<td>nsP2</td>
<td>G3228 → A</td>
<td>Ala-517 → Thr</td>
</tr>
<tr>
<td>ts18</td>
<td>G</td>
<td>1</td>
<td>nsP2</td>
<td>G3885 → A</td>
<td>Cys-304 → Tyr</td>
</tr>
<tr>
<td>ts19</td>
<td>A</td>
<td>1</td>
<td>nsP2</td>
<td>U2926 → C</td>
<td>Leu-416 → Ser</td>
</tr>
<tr>
<td>ts20*</td>
<td>E</td>
<td>2</td>
<td>E2</td>
<td>A9502 → U</td>
<td>His-291 → Leu</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E2</td>
<td>G9890 → U</td>
<td>Silent: Ser-420‡</td>
</tr>
<tr>
<td>ts21</td>
<td>A</td>
<td>1</td>
<td>nsP2</td>
<td>G2590 → A</td>
<td>Ala-106 → Thr⁷</td>
</tr>
<tr>
<td>ts22</td>
<td>n.t.</td>
<td>n.s.</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ts23*</td>
<td>D</td>
<td>5</td>
<td>E1</td>
<td>G10380 → A</td>
<td>Ala-106 → Thr⁷</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E1</td>
<td>G10864 → A</td>
<td>Arg-267 → Gln⁸</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E1</td>
<td>A10174 → U</td>
<td>Glu-37 → Val⁸</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E1</td>
<td>A11193 → U</td>
<td>Asn-377 → Tyr⁸</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E2</td>
<td>C9442 → U</td>
<td>Ala-271 → Val⁸</td>
</tr>
<tr>
<td>ts24*</td>
<td>A</td>
<td>3</td>
<td>nsP2</td>
<td>G3885 → A</td>
<td>Gly-736 → Ser</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>nsP4</td>
<td>C6339 → A</td>
<td>Gln-191 → Lys</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>nsP3</td>
<td>U4756 → C</td>
<td>Val-219 → Ala‡</td>
</tr>
</tbody>
</table>
These combined ts mutant mapping efforts revealed that mutations in any of the four nonstructural proteins can induce a RNA-phenotype, indicating that all four proteins are critical for RNA replication. However, the four RNA-complementation groups did not map nicely to the four nonstructural proteins. Complementation groups B and F could be clearly assigned. The group B mutant, ts11, has a mutation in nsP1. Ts11 ceases minus-strand RNA synthesis at the non-permissive temperature which does not reactivate during shift-down, indicating that nsP1 is important for minus-strand synthesis [112]. Separate efforts in the Stollar laboratory mapped low methionine resistance mutations to nsP1, and studies in the Kääriäinen laboratory revealed that nsP1 expressed in vitro contains methyltransferase and guanyltransferase activities, both necessary functions for capping and thus implicating nsP1 in capping viral RNAs [113, 114]. Group F mutants including ts6 are defective in viral RNA elongation, and all map to nsP4, indicating that nsP4 is the viral RNA-dependent RNA polymerase [108, 109]. It also contains homology to other viral polymerases and a GDD (Gly-Asp-Asp) nucleotide-binding domain conserved in RNA polymerases, providing additional evidence that nsP4 is the alphavirus polymerase [115].

The majority of ts mutations mapped to nsP2, and have varied phenotypes (Figure 1.15). Both groups A and G mapped to mutations in nsP2, and thus
represent intragenic complementation rather than individual proteins, as had been originally thought. Many mutants of both groups, including *ts*17 and *ts*24 of group A and *ts*18 of group G, fail to process the nonstructural polyprotein into individual proteins at the non-permissive temperature, indicating that nsP2 is its protease. Several mutants (*ts*17, *ts*18, *ts*21, *ts*24) are defective in the synthesis of subgenomic RNA upon shift-up (from permissive temperature, 30°C, to non-permissive temperature, 40°C) suggesting that functional nsP2 is essential for subgenomic RNA initiation. Additionally, several mutants (*ts*17, *ts*24) do not properly shut down minus-strand synthesis at 40°C, and interestingly will resume minus-strand synthesis if the temperature is shifted up after normal cessation at 30°C [116]. Of the revertants, which form normal plaques at 40°C, several were still defective in subgenomic strand synthesis or minus-strand shutoff. However, all revertants cleaved the polyprotein into the individual proteins normally, indicating that defects in the protease function were causal for the *ts* phenotype, and perhaps led to the other phenotypes [109]. Homology to papaya plant cysteine proteinase (papain) and mutagenesis studies identified Cys-481, His-558 and Trp-559 on the carboxyl end of nsP2, which are two-dimensionally near the *ts*7, *ts*17 and *ts*18 lesions, as the active site residues (Figure 1.15) [117, 118].
Several nsP2 mutations were clearly temperature sensitive and showed none of the above phenotypes, indicating that nsP2 must have additional functions which can also lend a ts phenotype upon perturbation [109]. The amino end of nsP2 was discovered to contain helicase domains, perform nucleotide triphosphatase (NTPase) and RNA triphosphatase (RTPase) activities and be involved in cellular transcriptional shutoff [119–123]. These functions however have not yet been mapped to ts phenotypes. nsP1, nsP2 and nsP4 have

---

**Figure 1.15. Map of nsP2 ts Mutants and Active Site Residues.** All ts mutants isolated by Burge and Pfefferkorn with mutations mapped within nsP2 is shown at its amino acid position. Systemic amino acid replacement analysis of all conserved cysteine and histidine residues in nsP2 identified Cys-481 and His-558 as the active site residues, as all replacements at these positions abolished processing and was lethal for the virus. Trp-559 is also conserved among alphaviruses and did not tolerate replacement. Solid filled symbols above the protein indicate reduced or abolished proteolytic activity upon replacement, and open symbols below indicate tolerated replacements with wild-type proteolytic activity. Checkered symbols represent ts mutants with reduced proteolytic activity at the non-permissive temperature. Adapted from [118].
homologies to proteins from several plant and animal viruses, providing functional clues, but nsP3 does not [124]. The RNA- phenotypes of ts4 and the nsP3 mutation in ts7 in isolation showed that nsP3 is clearly involved in viral genome replication, but additional functions took a while to discover [125]. nsP3 is a phosphoprotein and contains ADP-ribose binding and hydrolysis activities [126]. The recent 2018 review, "The Enigmatic Alphavirus Non-Structural Protein 3 (nsP3) Revealing Its Secrets at Last" by Dr. Gerald McInerney's laboratory at Karolinska Institute in Sweden, details recent investigations of this protein [127]. For more contemporary information on the role of the four nonstructural proteins in alphavirus infection and RNA synthesis please see the 2015 review, "Alphavirus RNA synthesis and non-structural protein functions" by Dr. Richard Hardy's laboratory at Indiana University in Bloomington, and the 2017 review, "Alphavirus polymerase and RNA replication" by Dr. Tero Ahola's laboratory at the University of Helsinki in Finland [128, 129]. For a very comprehensive and detailed review of alphaviruses I recommend the 1994 Review "The alphaviruses: gene expression, replication, and evolution" by Jim and Ellen Strauss - my appendix, "Sindbis Genome Index of Strauss & Strauss (1994): The Alphaviruses" can point the reader to pages of discussion and references within this review for particular genome features of interest [17].
The link between extensive work on Sindbis virus in Elmer Pfefferkorn's laboratory in the 1960's and Jim Strauss's laboratory from the 1970's onwards seemed to be a brief stint by the virus in Jim Darnell's laboratory in the late 1960's to early 1970's. I often see Jim Darnell around the Rockefeller University campus, and asked him why he started work on this particular virus. He explained; "Wow no one has asked me that question in 25 years! We started on Sindbis because I had an errant idea that the viral membrane lipid composition may be different than that of the plasma membrane and be coded for by viral genes. I had moved to Albert Einstein [College of Medicine] from MIT because they'd tripled my salary, and Boyce Burge also moved from Boston and started in my lab as a postdoc. Boyce brought Sindbis with him and because it had an envelope, we thought it would be a good system to look at the lipids. Salvador Luria at MIT was looking for young virologists and Boyce went and joined MIT, and Jim Strauss took over the project. We were very wrong about that notion, it's been conclusively shown that the composition of lipids in the viral envelope are the same as the cellular membrane, with some viral proteins in it. I sent Sindbis over to Stephen Harrison in Boston, and he showed it budding from the cell membrane [130]. [Upon asking me how Dr. Harrison has been]. It's good to know that me sending him Sindbis didn't derail his career. Another poor fellow came over from Chicago and I had bought an HPLC machine for this project, and upon finding that the lipids in the envelope were just the same ones as in the cellular
membrane he went and joined another lab. Jim [Strauss] characterized some of the proteins and sugars [75, 131, 132]. Deborah Bernhardt, who had discovered pre-tRNAs [133], moved with me to Columbia [University] and brought Sindbis with her and found that messenger RNA that's one-third the size of the genome [134]. That was the last experiment I remember that we did with Sindbis. Jim Strauss took the project with him to establish his professorship over in California [at Caltech]. Charlie [Rice], as you know, was in Jim [Strauss's] lab when DNA sequencing arrived and made a name for himself by being the first to sequence Sindbis and a flavivirus [98, 99, 135]. This makes me Charlie's scientific grandfather". Jim Darnell laughed out loud at the last statement, and I reminded him that this consequently makes him one of my scientific great-grandfathers.

1.8. Discovery of Viral Translational and Transcriptional Control.

The discovery of an opal stop codon (UGA) between nsP3 and nsP4 upon sequencing of Sindbis virus revealed a mechanism of translational control of the nonstructural proteins [99]. In most cases, the nonstructural polyprotein translation stops at this codon forming protein P123. However, translational readthrough produces the full polyprotein P1234 at a small frequency, between 5% and 20% in mammalian cells, in which nsP4 is immediately cleaved in early infection [136]. Consequently nsP4, the viral polymerase, is produced around tenfold less than the other three nonstructural proteins. And
there are two different forms of nsP3, one in which translation ends at the opal stop codon, and the readthrough product from which nsP4 is cleaved, and is several amino acids longer. Whether these extra amino acids provide any functional significance is still unknown [17]. The efficiency of readthrough at this opal stop codon is also temperature sensitive, and has about a 20% efficiency at 30°C but less than a 5% efficiency at 40°C [137, 138].

The discovery of an opal stop codon upstream of the polymerase was not a great surprise, as several plant viruses and retroviruses were already shown to utilize amber stop codons for translational control [139]. The precise readthrough mechanism and whether it utilizes misreading or suppressor tRNAs, and the amino acid inserted in place of the opal stop codon during readthrough, is still unknown. Bacterial polymerases were shown to misread opal stop codons around 3% of the time and insert tryptophan, which produces minor products critical to Q-Beta phage infection [140]. Additionally, vertebrate cells were found to contain small amounts of opal stop suppressor tRNAs which likely compete with release factor proteins to allow readthrough at small frequencies. Opal suppressor tRNAs which insert tryptophan have been isolated from rabbit reticulocytes and opal suppressor tRNAs which insert serine or phosphoserine were found in bovine and chicken livers [139].

As more alphavirus sequences were reported, most alphaviruses were found
to contain a stop codon, which was always an opal codon, at this position, and the ones without stop codons were found to contain an arginine codon (Figure 1.16 A) [17]. The availability of the Sindbis infectious clone enabled mutagenic insertion and testing of tryptophan, serine, arginine and the alternative stop codons at this position, by Guangpu Li in the Rice laboratory (Figure 1.16 B) [136].

A)

B)
Figure 1.16. Translational Control in Alphaviruses Using Opal Stop Codon.
A) Sequence alignment of various alphaviruses indicated a predominant presence of the opal stop codon (boxed) between nsP3 and nsP4. The solid line indicates the P3/4 cleavage site in readthrough translation products. Sindbis (SIN) isolates AR339 and AR86 is shown, with changes in AR86 shaded; AR86 contains a UGU cytosine codon instead. The initially sequenced isolates of O'Nyong Nyong (ONN) and Semliki Forest (SF) contain an arginine codon (CGA) instead. WEE, EEE, VEE; Western, Eastern and Venezuelan Equine Encephalitis respectively, MID; Middleburg, RR; Ross River viruses. B) Various sense and alternative stop codons were inserted into Sindbis virus Toto1000 infectious clone and grown in CEF cells at MOI 20. The opal and ochre stop codons were the most and least fit, respectively, in CEFs, and this fitness advantage is reduced at higher MOI's. Adapted from [17, 136].

Introducing a sense codon does not lead to detection of equimolar amounts of nsP4 as the other nsP's, indicating involvement of a degradation pathway for excess nsP4 [136]. Interestingly, the original sequences for O'Nyong Nyong virus and Semliki Forest virus do not contain an opal stop codon. However, as sequencing technologies advanced, sequences of early passage isolates of O'Nyong Nyong was found to contain both the opal stop codon and the arginine codon, indicating that they exist in equilibrium in nature [141].

Infection of live Anopheles gambiae mosquitoes with bloodmeals containing a mix of 10% opal codon and 90% arginine codon-containing O'Nyong Nyong virus led to a predominance of opal stop codons at this position after 8 to 10 days, indicating that live mosquitoes exert enough selective pressure to give the stop codon a fitness advantage. However, O'Nyong Nyong virus with arginine codons at this position had the highest fitness in cell culture, indicating that perhaps a high number of cell culture passages prior to
sequencing could have contributed to the detection of arginine in the initially sequenced isolates [141]. Perhaps there are host-specific selective advantages such that containing such an equilibrium in nature is of advantage to certain alphaviruses.

Interestingly the cytosine nucleotide next to the opal or arginine codon is conserved in all alphaviruses (Figure 1.16 A). The Stollar laboratory had isolated several host-range Sindbis RNA+ mutants which were ts in vertebrate cells but not in mosquito cells, and several RNA- mutants which were ts in both cell types [142, 143]. Julie Lemm in the Rice laboratory found that these RNA- mutations individually inserted into an infectious clone caused the virus to also exhibit a host-range phenotype, and was ts in mosquito cells but not in CEFs [144]. One such mutation was a synonymous C→U mutation next to the opal stop codon which caused translational leakiness, further indicating that the importance of translational termination at this stop codon was perhaps dependent on host. Both sense codons in place of the opal codon and the silent C→U mutation of the adjacent cytosine cause an overproduction of uncleaved P34 in cell culture, which may underlie the host-dependent effects of viral infection [136, 144].

Each virion only contains one genomic RNA but contains several hundred copies of each of the structural proteins, and thus producing the structural
polyprotein at a vast molar excess to the nonstructural polyprotein would provide better efficiency of cellular resource usage. This is clearly the case in alphaviruses as shown by the relative ease of detection of intracellular structural proteins compared to nonstructural proteins using radiolabel incorporation assays (Figures 1.11 B and 1.13 B). The other major form of translational control by alphaviruses is the exclusive use of the subgenomic mRNA for structural protein translation. The amount of structural proteins largely correlates with the amount of subgenomic RNA in an infected cell. In general, in later stages of infection there is about threefold more subgenomic RNAs than genomic RNAs, and many genomic RNAs are packaged into nucleocapsids leaving only a fraction of these available for translation. At these stages polysomes containing viral RNA were found to contain a 10:1 ratio of subgenomic:genomic RNAs [139]. Thus translational control of the ratio of structural:nonstructural proteins is largely achieved by the ratio of subgenomic:genomic RNAs available for translation. Consequently, transcriptional control of the ratio of these RNAs is critical for alphaviral infection efficiency.

All viral RNA species - genomic plus strand, subgenomic plus strand and minus strand, and the dsRNA forms of each - were thought to be under transcriptional control due to the disruptive effect most RNA-\textit{ts} mutants had on these ratios. Sequence alignments of alphaviruses revealed an invariant
glycine at the penultimate position of each nonstructural protein cleavage sites, and provided a mutagenesis target which helped elucidate the elegant mechanism by which the virus modulates these RNA ratios (Figure 1.17).

Yukio Shirako, in the Strauss laboratory, tested the effects of mutating these positions on virus growth, in an opal stop codon → serine infectious clone backbone to minimize confounding effects of multiple nonstructural protein translation products (Figure 1.17 & Table 1.8) [138, 145]. As detailed in the table below, after a period of a variety of naming conventions, these mutants were generally named with the cleavage site number containing the mutation of the penultimate glycine followed by the letter indicating the amino acid in place of the glycine (Table 1.8).

Figure 1.17. Identity of Nonstructural Protein Cleavage Sites. Amino acid alignment of various alphaviruses at the N-terminal and C-terminal regions
around the nonstructural protein cleavage sites reveal an invariant glycine (shaded) at the penultimate position to each cleavage site (asterisks). Boxed X represents the opal stop codon. The mutations were introduced and assayed in Sindbis infectious clones with a serine in place of the opal stop codon, such that only P1234 is produced instead of a combination of P123 and P1234, simplifying analyses. The naming convention indicated, i.e., cleavage site 1 between nsP1 and nsP2 = P1/2, is used throughout this chapter and the literature. Alphavirus abbreviations are listed in figure 1.16. Adapted from [145].

Table 1.8. Growth and Temperature Sensitivity of Nonstructural Protein Cleavage Mutants. The mutations were inserted into the Sindbis virus Toto1101.S infectious clone, which differs from Toto1101 by replacement of the opal stop codon with serine (*). The penultimate glycine at the P1/2 and P2/3 position was created by replacing the G at nucleotide 1676 and/or 4096 to C, U or A, which resulted in substitution of Ala, Val or Glu respectively. Growth was determined via plaque assay by infection of CEFs at either temperature. Italics represent an approximation. Adapted from [145].

<table>
<thead>
<tr>
<th>Sindbis Virus Site 1</th>
<th>Site 2</th>
<th>Virus Growth</th>
<th>Plaque Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>* Opal→Ser P1/2 Site</td>
<td>P2/3 Site</td>
<td>30°C</td>
<td>40°C</td>
</tr>
<tr>
<td>WT</td>
<td>Gly</td>
<td>Gly</td>
<td>2.3E+09</td>
</tr>
<tr>
<td>1A</td>
<td>Gly→Ala</td>
<td>Gly</td>
<td>1.9E+09</td>
</tr>
<tr>
<td>1V</td>
<td>Gly→Val</td>
<td>Gly</td>
<td>1.0E+04</td>
</tr>
<tr>
<td>1E</td>
<td>Gly→Glu</td>
<td>Gly</td>
<td>1.0E+04</td>
</tr>
<tr>
<td>2A</td>
<td>Gly</td>
<td>Gly→Ala</td>
<td>2.0E+09</td>
</tr>
<tr>
<td>2V</td>
<td>Gly</td>
<td>Gly→Val</td>
<td>8.0E+08</td>
</tr>
<tr>
<td>2E</td>
<td>Gly</td>
<td>Gly→Glu</td>
<td>8.9E+08</td>
</tr>
<tr>
<td>1V2V</td>
<td>Gly→Val</td>
<td>Gly→Val</td>
<td>1.0E+04</td>
</tr>
<tr>
<td>1E2E</td>
<td>Gly→Glu</td>
<td>Gly→Glu</td>
<td>1.0E+04</td>
</tr>
</tbody>
</table>

Yukio found that mutating the penultimate glycine position affected the growth of the virus and made them temperature sensitive in CEFs; "After transfection of chicken cells with WT transcripts, large plaques were formed at both 30°C or 40°C. The 2V mutant formed small plaques at 30°C and minute plaques at
40°C. The 1V mutant and the 12V mutant formed minute plaques at 30°C and no plaques at 40°C. There was a correlation between virus growth and plaque size at both temperatures (Figure 1.18 A & B) [138]. Rodion Gorkachov in the laboratory of Elena Frolova and Ilya Frolov at the University of Texas Medical Branch in Galveston later found that these mutants are even more defective in mosquito cells relative to vertebrate cells, indicating that mutation of the penultimate glycine in each cleavage site gives rise to both ts and host-range phenotypes (Fig 1.18 C) [146].

Figure 1.18. Growth and Translation Characteristics of Nonstructural Protein Cleavage Mutants. Plaque assay of WT Sindbis virus and cleavage mutants 2V and 1V2V on CEFs after RNA transfection (A) and titers upon infection (B) at 30°C and 40°C; plaque size correlates with growth characteristics. C) Growth characteristics upon RNA transfection of GFP-tagged variants of these viruses in vertebrate and invertebrate cells, indicating a host-
range phenotype. SINV/1V2V/GFP/A (mosquito cells panel) refers to an adaptive 1V2V/GFP variant with an nsP4 E451→A mutation which was adaptive in vertebrate cells allowing similar growth to 2V/GFP and similar plaque sizes to the WT GFP virus, but was still heavily detrimental in mosquito cells. D) Viral proteins synthesized in an *in vitro* translation system containing rabbit reticulocyte cell lysate mix and radiolabeled $^{35}$S-methionine, using the cleavage mutant constructs indicated in table 1.8. Adapted from [138, 145, 146].

The translation products from these cleavage mutant constructs were then tested in a rabbit reticulocyte cell lysate *in vitro* translation system ([Figure 1.18 D](#)) [138, 145]. Translation analysis confirmed the expectation that certain substitutions of the penultimate glycine affected nonstructural protein cleavage at the mutated site, but also contained surprises - certain mutations at the 1/2 site affected cleavage at both the 1/2 and 2/3 sites. Consequently, the 1V and 1V2V mutants behaved fairly similarly; "Substitution of the penultimate Gly in the 1/2 or 2/3 cleavage sites with either Val or Glu [using 1V, 1E, 2V or 2E] abolished the processing at the mutated site, that abolition of processing at the 1/2 site [using 1V or 1E] also prevented efficient processing of the 2/3 site but abolition of cleavage at the 2/3 site [using 2V or 2E] did not interfere with processing at the 1/2 site, and that in either case the 3/4 site was cleaved efficiently" ([Table 1.9](#)) [138]. The 1V mutant would occasionally produce normal sized plaques, which when sequenced had mutations to alanine or reversions to glycine, indicating that alanine is also tolerated well at this penultimate position. Mutation of the 3/4 cleavage site using the 3V mutant only affected cleavage at that site, and cleavage at the 1/2 and 2/3 sites were
normal (Table 1.9) [138]. Additionally, mutation of the protease active site cysteine abolished cleavage at all sites, as expected, and behaved similarly to mutating all three cleavage sites (Table 1.9, P2C481G and 1V2V3V rows).

Table 1.9. Translation Products of Nonstructural Protein Cleavage Mutants. Detection of Sindbis virus nonstructural polyprotein cleavage products for various cleavage mutants and a protease active site mutant (P2C481G) in an in vitro translation system (see figure 1.18 D). (+) with parentheses indicates that only trace amounts were detected. (a) Abolishing cleavage at the 1/2 site, which prevents processing of mature nsP1, also decreased cleavage at the 2/3 site, such that mature nsP2 was undetectable and only trace amounts of mature nsP3 was visualized. Adapted from [138].

<table>
<thead>
<tr>
<th>Construct</th>
<th>nsP1</th>
<th>nsP2</th>
<th>nsP3</th>
<th>nsP4</th>
<th>P12</th>
<th>P23</th>
<th>P34</th>
<th>P123</th>
<th>P234</th>
<th>P1234</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>1V</td>
<td>–</td>
<td>–</td>
<td>(+)</td>
<td>(+)</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>2V</td>
<td>+</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>3V</td>
<td>+</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>1V2V</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>1V3V</td>
<td>–</td>
<td>–</td>
<td>(+)</td>
<td>–</td>
<td>(+)</td>
<td>–</td>
<td>(+)</td>
<td>+</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>2V3V</td>
<td>+</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>1V2V3V</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>P2C481G</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
</tr>
</tbody>
</table>

Each cleavage site also seemed to be temperature sensitive, and cleavage at the 1/2 and 2/3 sites were more efficient at the higher temperature, 40°C, than at 30°C, whereas cleavage at the 3/4 site was more efficient at the lower temperature, 30°C, than at 40°C. Perhaps these underlie the growth of certain cleavage mutants having a somewhat ts phenotype; "these differences in processing efficiencies may be due to changes in proteinase activities, to altered substrate structures, or to changes in the kinetics of translation at the two temperatures" (Table 1.8 and Figure 1.18 A & B) [138].
These studies indicated that these events may have a preferred order, and a picture of nonstructural polyprotein cleavage began to emerge. Based on *in vitro* translation experiments, the following sequence of events during cellular infection was postulated by Yukio Shirako; "Early in infection, when the numbers of genomic RNAs and translation products are limited, the 3/4 site in P1234 is cleaved rapidly in cis [producing P123 and nsP4], whereas the 1/2 and 2/3 sites in P123 are cleaved only slowly because a *trans* cleavage at the 1/2 site is required to start the cleavage cascade. P123 and mature nsP4 are thus the major translation products at this stage. Later in infection, the concentration of the viral proteinase [nsP2, within P123 at this stage] increases to the point at which the 2/3 site is cleaved rapidly, resulting in the production of P12 and nsP3 from P123 and P12 and P23 from P1234; P12 is further cleaved into nsP1 and nsP2, whereas P34 accumulates [in late infection] as a result of the lack of active enzymes that cleave the 3/4 site" [138].

The production of viral RNA's during infection was visualized for viable mutants in the above studies, and there were detectable effects of nonstructural protein cleavage on the amount and relative ratio of Sindbis genomic RNA and subgenomic RNA [138]. The Sawicki laboratory quantified these ratios for GFP-tagged versions of 2V and 1V2V during late infection.
using radiolabels and found that the subgenomic:genomic molar ratio decreases when 2/3 cleavage is abolished, and decreases further when both 1/2 and 2/3 cleavages are abolished (Figure 1.19 A); "The molar ratio of [subgenomic]-to-genomic RNA synthesis was less than 3 for SINV/2V/GFP and was 8 for SINV/GFP. Surprisingly, the SINV/1V2V/GFP/A variant synthesized amounts of genome RNA that were similar to those of SINV/GFP, but its [subgenomic-to-genomic] molar ratio of ~1 indicated that the P123-containing replication complexes were specifically defective in the internal promoter recognition needed for 26S [subgenomic] RNA transcription" [146, 147]. The amount of both RNA species is fairly equivalent during early infection and diverges later in infection in the wild-type virus. This divergence does not occur during infection by the 2V mutant, the most viable cleavage mutant, and similar amounts of genomic and subgenomic RNAs continue to be synthesized in late infection (Figure 1.19 B) [147].

Figure 1.19. Effect of nsP Cleavage Mutations on Ratio of Genomic and Subgenomic RNA During Infection. BHK cells were infected with the
indicated virus in the presence of $^3$H-uridine and gel bands were quantified by liquid scintillation counting. A) Ratio of Sindbis viral subgenomic:genomic RNAs produced by GFP-tagged versions of WT and 2V and 1V2V/A cleavage mutants, described in figure 1.18C. B) Time course of genomic (□) and subgenomic (△) RNAs synthesis during infection by GFP-tagged 2V cleavage mutant. Error bars represent range of values from three experiments. Adapted from [146, 147].

However, these analyses were limited to only the few cleavage mutants which replicated well enough in cell culture such that these various RNA's could be visualized as gel bands and quantified (Figure 1.18). To study the production of various RNAs and decouple virus viability from cleavage defects, Julie Lemm developed a trans-expression system using vaccinia virus that can express the nonstructural polyproteins and cleavage products within vertebrate cells without requiring Sindbis infection [148–150]. In this system, normal cleavage allowed this divergence and produced around fivefold more subgenomic RNA than genomic RNA, but all tested defects in 1/2 or 2/3 cleavage or protease activity affected this divergence and kept the ratio fairly equivalent (Table 1.10). Thus it seemed likely that during Sindbis virus infection of vertebrate cells, uncleaved P123 + nsP4 replication complexes produced equal ratios of both genomic and subgenomic RNA during early infection, whereas later in infection fully processed nsP1 + nsP2 + nsP3 + nsP4 replication complexes preferentially recognized the subgenomic promoter and produced far more subgenomic RNA than genomic RNA. Consequently, the nonstructural protein cleavage cascade underlies the
aforementioned phenomenon of the majority of polysomes containing subgenomic RNA during late infection, which results in translation of mostly structural proteins at these times. More recent studies in the Stollar laboratory has suggested that this divergence also occurs in mosquito cells, and the ratio can be flipped to produce more genomic than subgenomic RNA in mosquito cells via mutations in the subgenomic promoter region [151].
Table 1.10. Effect of Transexpression of nsP Cleavage Mutants on Ratio of Genomic and Subgenomic RNA. Any disruption of cleavage of P123 led to equivalent amounts of genomic and subgenomic RNAs, whereas WT P123 produced around fivefold more subgenomic RNA as genomic RNA. BHK cells were infected with vaccinia virus and transfected with the various indicated plasmids. * indicates the mutated protein in the immediate rows below, to which the other protein is added. Rows 1-4 are analyses of various polyproteins with and without replacement of the opal stop codon between nsP3 and nsP4. Rows 5-9 indicate mutations of P123 cleavage sites cotransfected with P34.S. Rows 10-14 indicate mutations of protease active site mutants cotransfected with ubiquitinated nsP4. Cleavage of the ubiquitin allows retention of the N-terminal tyrosine which is required for nsP4 function. The minus, genomic and subgenomic RNA values are normalized to the WT at the top of each set of rows, which is set to 1. The subgenomic to genomic RNA ratio is calculated within each sample and is not normalized. Viral RNAs were subjected to a RNase protection assay and quantified using a Betagen Betascope. Adapted from [149, 150].
<table>
<thead>
<tr>
<th>#</th>
<th>Sample</th>
<th>Normalized RNA Accumulation</th>
<th>Ratio</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Minus</td>
<td>Genomic</td>
<td>Subgenomic</td>
</tr>
<tr>
<td>1</td>
<td>WT</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>P1234.S</td>
<td>2.2</td>
<td>1.8</td>
<td>1.6</td>
</tr>
<tr>
<td>3</td>
<td>P123 + P34</td>
<td>1.7</td>
<td>1.3</td>
<td>1.5</td>
</tr>
<tr>
<td>4</td>
<td>P123 + P34.S</td>
<td>2.7</td>
<td>2.4</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td>P123* + P34.S</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>WT P123 (#4)</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1V</td>
<td>0.9</td>
<td>0.8</td>
<td>0.3</td>
</tr>
<tr>
<td>7</td>
<td>2V</td>
<td>1.5</td>
<td>3.3</td>
<td>1.2</td>
</tr>
<tr>
<td>8</td>
<td>1V2V</td>
<td>1.4</td>
<td>1.1</td>
<td>0.2</td>
</tr>
<tr>
<td>9</td>
<td>1E2E</td>
<td>0.4</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>nsP4 + P123*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>WT P123</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>P2-C481G</td>
<td>1.2</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>12</td>
<td>P2-C481S</td>
<td>4.3</td>
<td>1.4</td>
<td>0.3</td>
</tr>
<tr>
<td>13</td>
<td>P2-H558A</td>
<td>2.4</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>14</td>
<td>P2-H558Y</td>
<td>2.9</td>
<td>0.6</td>
<td>0.2</td>
</tr>
</tbody>
</table>
The protease mutants that Julie Lemm tested all produced more minus strands than the WT protease, indicating that perhaps replication complexes with uncleaved P123 preferentially transcribes minus strands from the plus strand template, and nonstructural protein cleavage shifts this preference to plus strand synthesis from the minus strand template (Table 1.10).

Experiments using viable cleavage mutant infections supported these results. Infection of vertebrate cells by WT Sindbis virus usually has a potent shutoff of new minus-strand synthesis after early infection; "Minus-strand synthesis stops by 6 h p.i. at 37°C in SIN virus-infected cells. infected BHK-21 cells showed the expected exponential increase in plus-strand RNA synthesis early in infection, converting to a linear and constant, maximal rate by 3 to 4 h p.i. [and] also showed an exponential burst of minus-strand synthesis.. and this burst was followed by a similar rapid decline and then overall cessation of minus-strand synthesis" (Figure 1.20 A & B) [152]. Due to the shutoff of new minus strand synthesis, the amount of minus strands detected remains constant in later infection in the WT virus; "The [total] amount of minus-strand RNA continued to increase during the first 3 h of incubation and then appeared to remain constant for up to 24 h at either 30° or 40°C. minus-strand RNA synthesis during the early stages of infection is indistinguishable in WT virus-infected and 2V [and 1V2V] mutant-infected cells" (Figure 1.20 D) [138].
Figure 1.20. Effect of nsP Cleavage Mutants on Minus Strand Shutoff. Abrogation of nonstructural protein cleavage prevents Sindbis minus strand shutoff; A & B show new synthesis of (-)RNA during 1 hr windows, C represents new (-)RNA synthesis in 2-3 hr windows, and D represents total accumulation of (-)RNA. For comparison, A also shows new total RNA synthesis in 1 hr windows. A) During growth of WT Sindbis virus in CEF cells at 30°C, 1 hr pulses of radioactive uridine incorporated into (O, left y-axis) total acid insoluble RNA or (●, right y-axis) specifically into minus strands. The y-axis scales for total incorporation and for minus strands differ by a factor of 10. B) Incorporation of 1 hr pulses of radioactive uridine into minus strands during infection of BHK cells at 37°C by WT Sindbis (■) or WT Sindbis replicons (□), which do not contain structural protein-coding regions of the genome. C) Incorporation of 2 or 3 hr pulses of radioactive uridine into minus strands during early and later infection of BHK cells at 37°C by Sindbis WT (open bars) or 2V cleavage mutant (filled bars). These were detected as RNase-resistant replicative forms for genomic (G) or subgenomic (SG) RNA; see figure 1.12 for description. D) Total minus strand accumulation by infection of CEFs at 30°C or 40°C by Sindbis WT, 2V and 1V2V mutants. (*) Bands represent the PCR product of a minus-strand specific RT-PCR assay run on an agarose electrophoresis gel. Adapted from [116, 138, 139, 147, 152].
As described above, ts17 and ts24 cannot turn off minus strand synthesis at the non-permissive temperature, and continue to synthesize new minus strands [116, 153]. Viable cleavage mutants replicated this phenotype during infection at permissive temperatures; "Minus-strand synthesis [by 2V mutant] does not stop as it does with WT SIN that produces [mature] nsP2 [and] nsP3.. Between 1 and 12 h p.i. [it was] estimated that 220% to 260% more minus strands were made during the infectious cycle than were made in WT" (Figure 1.20 C) [147]. Consequently, the total amount of minus strands continues to increase in cleavage mutants; "At 30°C [in CEFs].. with [the 1V2V mutant] the amount of minus-strand RNA continued to increase during a 12-h incubation period.. [quantification results indicate] that a significant fraction of the input [1V2V] mutant RNA could be transcribed into minus-strand RNA" (Figure 1.20 D) [138]. Thus, the disruption of nonstructural polyprotein cleavage, in addition to preventing the upregulation of the ratio of subgenomic:genomic RNA in late infection, also prevented the downregulation of minus strand synthesis after early infection.

Viral RNA can exist as both ssRNA and dsRNA, and studies have indicated that at least some of the minus strands are present complexed with plus strands as dsRNA in infected vertebrate and mosquito cells, which can be detected by immunofluorescence using dsRNA antibodies (Figure 1.21 A) [154, 155]. Such dsRNA foci is first detectable near the plasma membrane of
cells, and later is detectable throughout the cytoplasm [155, 156]. More
dsRNA foci is detectable in infections by viable cleavage mutants than by WT
during early infection, and the number of foci sees a moderate decrease in all
cases as the infection progresses (Figure 1.21 B) [155]. Additionally, cleavage
mutants induce more interferon in cells containing intact immune systems, and
do not have cytopathic effects in such cells (Figure 1.21 C) [146]. The
increased interferon induction by cleavage mutants could be due to higher
amounts of dsRNA, but could also be due to other factors [152, 155, 157–
159].
In summary, Sindbis virus controls translation by (i) using an opal stop codon upstream of nsP4 such that translational readthrough produces far less nsP4 than the other nonstructural proteins, and (ii) using genomic RNA for the translation of nonstructural proteins and subgenomic RNA for translation of structural proteins, whereby transcriptional control of amounts of these RNAs available for translation largely modulates the relative amounts of proteins produced. The virus has three RNA species during infection: genomic plus strand, complementary minus strand, and subgenomic mRNA strand. The ratios of these RNA species is temporally controlled during infection, and undergoes a couple ‘switches’: (i) plus / minus strand ratio: minus strand
production is shut off after early infection, switching to exclusively plus strand production, and (ii) genomic / subgenomic strand ratio: Both strands are produced in equal numbers in early infection, and switches to primarily producing subgenomic RNA later in infection.

These transcriptional switches occur due to successive processing of the P123 polyprotein, which is facilitated by increasing concentration of the nsP2 protease domain (figure 1.22 A). This mechanism has been studied mostly using (i) cleavage site mutants which alter polyprotein cleavage activity at specific sites (figure 1.22 B), (ii) protease active site mutants which reduces or enhances cleavage activity at all sites, and (iii) ts mutants wherein cleavage activity can be controlled with temperature differences. Unraveling the various steps of this mechanism using viable cleavage mutants, such as 2V and 1V2V, and by using cleavage and protease mutants in trans-expression and in vitro translation systems finally enabled the assignment of RNA- nonstructural protein ts mutant defects at the nonpermissive temperature to specific steps along this pathway (figure 1.22 C).
Figure 1.22. Summary of Sindbis Nonstructural Protein Cleavage Mechanism. A) Summary of characteristics of Sindbis nonstructural protein cleavage site mutants. Shaded boxes and white boxes represent detected and undetectable polyproteins and cleavage products, respectively, and X's represent abrogated cleavage sites. Present (+), intermediate (±) and absent (-) RNA synthesis at 30°C and 40°C in CEFs was detected by total radiolabeled uridine incorporation for plus strands and by qPCR assay for minus strands. Naming convention of cleavage site mutants is described in table 1.8. B) Model of how nonstructural protein cleavage forms early minus strand and later plus strand specific replication complexes. The nsP1, nsP2, nsP3 and nsP4 domains are shown with variable shading within polyproteins and as individual proteins. The proposed mechanism is described within the text. C) Placement of mapped Sindbis and Semliki Forest RNA- ts mutants which at the nonpermissive temperature are defective along specific steps in the proposed nonstructural protein cleavage mechanism. RC\textsubscript{initial} refers to minus strand complexes which require continuous protein synthesis, and are no longer made when cyclohexamidine, which stops cellular protein synthesis, is added to infected cells. RC\textsubscript{stable} refers to plus strand replication complexes which are not affected by cyclohexamidine addition. Involvement of host-factors specific to vertebrate and mosquito cells have been proposed, but their identity is still elusive. Adapted from [17, 160].
<table>
<thead>
<tr>
<th>Virus</th>
<th>Wild Type</th>
<th>Nonstructural Cleavage-Site Mutants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>12V</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonstructural Proteins Synthesized</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precursors</td>
<td>nsP1 nsP2 nsP3 nsP4</td>
<td></td>
</tr>
<tr>
<td>Cleaved products</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Growth Characteristics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNA Synthesis</td>
<td></td>
<td>30°</td>
</tr>
<tr>
<td>Minus strand</td>
<td></td>
<td>++</td>
</tr>
<tr>
<td>Plus strand</td>
<td></td>
<td>++</td>
</tr>
<tr>
<td>Plaque Size</td>
<td></td>
<td>Large</td>
</tr>
<tr>
<td>PFU/ml</td>
<td></td>
<td>$10^8$</td>
</tr>
</tbody>
</table>
B) Translation of 49S RNA

Cleavage of 3/4 bond

nsP2 proteinase

P123 + nsP4 (+ host factors?)

Minus-strand RNA synthesis

CAP

nsP1, nsP2, nsP3, + nsP4 (+ host factors?)

Cleavage of 1/2 and 2/3 bonds

nsP2 proteinase

Plus strand RNA synthesis

CAP
Therefore, this mechanism represents viral protein concentration-dependent transcriptional switches which may be of interest in synthetic biology. Perhaps the increased concentration of nsP2 protease domain relative to nsP4, due to the opal stop codon and/or selective degradation, enables proper timing of the cleavage events and consequent transcriptional switches. Perhaps the cleavage cascade also plays a role in limiting templates for innate immune system recognition resulting in muted interferon signaling. The interplay between these elegant alphaviral translational and transcriptional control systems enables the performance of different functions during early and late infection to maximize the use of cellular materials for infection.

1.9. Discovery of Viral RNA Sequence Elements.

Based on what had been discovered about the viral life cycle from the above studies, it was obvious that the nucleotide context of the viral RNA itself had to at least perform two functions; (i) contain stretches of nucleotides which can be specifically recognized by the viral replication complex to serve as promoters for the various RNA species described in the prior section, and (ii) contain nucleotide patterns which can be recognized by the capsid protein to at least initiate selective packaging of genomic RNA into nucleocapsids. The viral RNA likely performs many more functions as well, outside of its protein coding functions. However, very few synonymous mutants of Sindbis virus
were isolated from the use of screens for phenotypes previously described, such as temperature sensitivity. The aforementioned synonymous C→U mutation next to the opal stop codon, which made Sindbis ts in mosquito cells but not in vertebrate cells, is one such example [144]. Consequently, most of these RNA functional elements were discovered after the advent of DNA sequencing.

When the full genomic Sindbis RNA was sequenced and putative protein coding regions of the RNA were determined, regions outside of these that do not code for proteins seemed to be an obvious home for RNA elements. However, pinpointing sites within these untranslated regions (UTRs) that were most functionally important, and can serve as targets for mechanistic mutagenesis studies, only became possible when additional alphavirus genomes were sequenced and multiple sequence alignment was an option. Some of these became available as full genome sequences of alphaviruses were elucidated and published by researchers around the world, and some of these were acquired by targeted sequencing of UTR's by Dr. James Ou, currently at the University of Southern California in Los Angeles, while he was in the Strauss laboratory [23, 24, 26]. As concisely explained by Jim and Ellen Strauss; "It seems intuitively obvious that segments of the alphavirus genome that are bound in a specific fashion to virus proteins during RNA replication or encapsidation will tend to be conserved during alphavirus evolution. Domains
involved in specific interactions should change more slowly than other regions of the genome because changes in the nucleotide sequence or structure would require compensating changes in the structure of the protein that binds to that domain. Furthermore, host proteins, which evolve much more slowly than viral components, may form a component of the replicase. In such cases, the interacting viral RNA sequence would be under constraints not to change. Thus, conserved nucleotide sequences or structures in alphavirus RNAs may serve as regulatory elements for initiation of RNA transcription or for initiation of encapsidation of the genomic RNA" [139]. Four such regions of alphavirus RNA sequence, termed conserved sequence elements (CSE's), were thus discovered soon after just a few additional alphavirus genome sequences were revealed (Figure 1.23).
Figure 1.23. Alphavirus Conserved RNA Elements Identified Using Multiple Sequence Alignments. Shown is a map of Sindbis virus RNA with the four conserved nucleotide elements indicated with open boxes and the corresponding alphavirus multiple sequence alignments of A) 3' UTR, B) 5' UTR, C) region adjacent to subgenomic RNA start site which overlaps with C-terminal of nsP4 and the integenic junction region, and D) the 51-nt CSE region in nsP1. HJ: Highlands J virus. Additional alphavirus abbreviations are listed in the caption to figure 1.16. Adapted from [139].
A. 3' TERMINAL SEQUENCES IN GENOME RNAs

B. SEQUENCES COMPLEMENTARY TO THE 5' TERMINI

C. SEQUENCES COMPLEMENTARY TO THE JUNCTION

D. 51 NUCLEOTIDE CONSERVED SEQUENCE
As was expected, multiple sequence alignment of alphavirus genomes revealed a 5' CSE within the 5' UTR (Figure 1.23 B), a 3' CSE within the 3' UTR (Figure 1.23 A), and a CSE which includes the untranslated intergenic junction region between the coding regions for the nonstructural protein and the structural protein, but also overlaps with the 3' end of nsP4 (Figure 1.23 C). As more of a surprise, there was also a strongly structured region within nsP1 and a little downstream of the 5' UTR, which was descriptively named as the 51-nt CSE and predicted to contain two stem loop structures (Figure 1.23 D) [19]. Mutagenesis studies indicated that the 5' and 3' CSE's serve as promoters for full length RNA synthesis, and the 3' UTR contains sequence elements that produce host range phenotypes upon perturbation [18, 23, 24]. For example, an A→C mutation 7 nucleotides upstream of the polyA tract in the Sindbis 3' UTR gave it a ts phenotype in vertebrate CEFs and was very detrimental for growth in mosquito cells [18]. For a more current and comprehensive treatment of the 5' and 3' UTR's, I recommend the 2015 review, "The 5′ and 3′ ends of alphavirus RNAs – Non-coding is not non-functional" by Drs. Jeffrey Wilusz, William Klimstra, Scott Weaver and Michael Diamond and their respective laboratories [161]. As is intuitive, the CSE that overlaps with the 3' end of nsP4 and the intergenic region was identified to contain a portion of a promoter for 26S subgenomic RNA synthesis, as mutations within this region reduce the amount of subgenomic RNA transcribed during viral infection [26, 162]. The region of 24 conserved
nucleotides from position -19 to +5 relative to the 26S RNA start site (Figure 1.23 C) was found to form a minimally active promoter which is threefold to sixfold less active for this transcription than in the WT, whereas an extended 64 nucleotide region from -40 to +14 has full activity [163, 164].

Studies of the 51-nt CSE were more complicated. The CSE is located fully within the nsP1 coding region, from nucleotides 155 to 205 on the Toto1101 genome. However, nucleotide-level conservation strongly indicated that it was an important RNA element; "It should be noted here that the conservation of nucleotide sequences found cannot be explained simply by the necessity to conserve a stretch of amino acid sequence... the alphaviruses have diverged so extensively that even in regions in which amino acid sequence is conserved, the nucleotide sequence encoding this amino acid sequence is not conserved" [139]. Since it is exclusively within a coding region, mutagenesis studies were limited to probing synonymous mutations to prevent disruption of nsP1 protein function. Dr. Hubert Niesters, currently at the University of Groningen in the Netherlands, constructed and analyzed many such mutants as a postdoc in Jim Strauss's laboratory [19]. The intuition that the 51-nt CSE functions as a copromoter along with the nearby 5' UTR in full length viral RNA synthesis made sense, since many synonymous mutations in this region reduced viral titers (Table 1.11).
### Table 1.11. Phenotypic Properties of Sindbis Virus 51 Nucleotide CSE Mutants

Various phenotypic properties of 51-nt CSE mutants of Sindbis virus, wherein there does not seem to be much correlation between the various properties. These mutants are named for the mutation made in the Toto51 Sindbis virus infectious clone. For example, 158A indicates that the mutant contains an A at nucleotide 158, whereas it is a G in the WT sequence. Silent mutations are indicated by 'yes' in the synonymous column, and the specific nsP1 amino acid mutation is indicated for ones that are not silent. * indicates a double mutant. Stem loop numbers indicate whether the mutation is in the first or second hairpin within the 51-nt CSE. Titers and plaque sizes were assayed in CEF cells grown at 30°C. RNA secondary structure free energies were calculated using methods developed by Tinoco et. al. [165]. Accumulated titer values and consequently titer reduction values represent estimates calculated from one-step growth curve plots. Hubert Niesters and Jim Strauss graciously put an effort to find these original values and tables from three decades ago, but they were unrecoverable. Data adapted from [19].

<table>
<thead>
<tr>
<th>Mutant</th>
<th>Sensitivity</th>
<th>Fertility</th>
<th>nsP1 Mutation</th>
<th>Stem Loop</th>
<th>Titer Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>158A</td>
<td>yes</td>
<td>yes</td>
<td>nsP1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>158G</td>
<td>no</td>
<td>no</td>
<td>nsP1</td>
<td>2</td>
<td>0.6</td>
</tr>
<tr>
<td>158C</td>
<td>yes</td>
<td>yes</td>
<td>nsP1</td>
<td>1</td>
<td>0.7</td>
</tr>
<tr>
<td>158T</td>
<td>yes</td>
<td>yes</td>
<td>nsP1</td>
<td>2</td>
<td>0.8</td>
</tr>
<tr>
<td>158A</td>
<td>yes</td>
<td>yes</td>
<td>nsP1</td>
<td>1</td>
<td>0.9</td>
</tr>
<tr>
<td>Mutant</td>
<td>Synonymous</td>
<td>Stem Loop</td>
<td>Free Energy</td>
<td>Titer</td>
<td>Plaque Phenotype</td>
</tr>
<tr>
<td>----------</td>
<td>------------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
<td>------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>kcal/mol</td>
<td>PFU/ml</td>
<td>Relative to WT</td>
</tr>
<tr>
<td>WT Toto51</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.8E+08</td>
<td>WT</td>
</tr>
<tr>
<td>158A.LP</td>
<td>Yes</td>
<td>1</td>
<td>-3.8</td>
<td>5.3E+07</td>
<td>Large</td>
</tr>
<tr>
<td>158A.SP</td>
<td>Yes</td>
<td>1</td>
<td>-3.8</td>
<td>3.8E+08</td>
<td>Small</td>
</tr>
<tr>
<td>158U</td>
<td>Q33H</td>
<td>1</td>
<td>-8.8</td>
<td>2.9E+06</td>
<td>WT</td>
</tr>
<tr>
<td>161A</td>
<td>Yes</td>
<td>1</td>
<td>-5.6</td>
<td>6.1E+04</td>
<td>WT</td>
</tr>
<tr>
<td>161G</td>
<td>Yes</td>
<td>1</td>
<td>-5.6</td>
<td>4.9E+07</td>
<td>WT</td>
</tr>
<tr>
<td>161U</td>
<td>Yes</td>
<td>1</td>
<td>-7.6</td>
<td>4.7E+05</td>
<td>WT</td>
</tr>
<tr>
<td>170C</td>
<td>Yes</td>
<td>1</td>
<td>-10.8</td>
<td>4.5E+05</td>
<td>WT</td>
</tr>
<tr>
<td>173U.LP</td>
<td>Yes</td>
<td>1</td>
<td>-4.8</td>
<td>1.5E+08</td>
<td>Large</td>
</tr>
<tr>
<td>173U.SP</td>
<td>Yes</td>
<td>1</td>
<td>-4.8</td>
<td>1.7E+08</td>
<td>Small</td>
</tr>
<tr>
<td>176A</td>
<td>H39Q</td>
<td>1</td>
<td>-8.6</td>
<td>0.0E+00</td>
<td>Lethal</td>
</tr>
<tr>
<td>176C</td>
<td>Yes</td>
<td>1</td>
<td>-8.6</td>
<td>1.3E+06</td>
<td>WT</td>
</tr>
<tr>
<td>176G</td>
<td>H39Q</td>
<td>1</td>
<td>-8.6</td>
<td>0.0E+00</td>
<td>Lethal</td>
</tr>
<tr>
<td>185G</td>
<td>Yes</td>
<td>2</td>
<td>-7.6</td>
<td>2.7E+05</td>
<td>WT</td>
</tr>
<tr>
<td>188U</td>
<td>R43S</td>
<td>2</td>
<td>-17</td>
<td>0.0E+00</td>
<td>Lethal</td>
</tr>
<tr>
<td>197A</td>
<td>Yes</td>
<td>2</td>
<td>-10.4</td>
<td>7.5E+05</td>
<td>WT</td>
</tr>
<tr>
<td>197C</td>
<td>Yes</td>
<td>2</td>
<td>-10.4</td>
<td>Low</td>
<td>WT</td>
</tr>
<tr>
<td>197U</td>
<td>Yes</td>
<td>2</td>
<td>-10.4</td>
<td>Low</td>
<td>WT</td>
</tr>
<tr>
<td>200C</td>
<td>Yes</td>
<td>2</td>
<td>-13.4</td>
<td>1.7E+06</td>
<td>WT</td>
</tr>
<tr>
<td>201U</td>
<td>Yes</td>
<td>2</td>
<td>-12.4</td>
<td>3.2E+05</td>
<td>WT</td>
</tr>
<tr>
<td>203A</td>
<td>Yes</td>
<td>2</td>
<td>-7.6</td>
<td>Low</td>
<td>WT</td>
</tr>
<tr>
<td>203C</td>
<td>Yes</td>
<td>2</td>
<td>-7.6</td>
<td>Low</td>
<td>WT</td>
</tr>
<tr>
<td>203U</td>
<td>Yes</td>
<td>2</td>
<td>-7.6</td>
<td>Low</td>
<td>WT</td>
</tr>
<tr>
<td>158A173U*</td>
<td>Yes</td>
<td>1</td>
<td>-3.6</td>
<td>8.1E+07</td>
<td>WT</td>
</tr>
</tbody>
</table>
Further mechanistic analysis was tricky, as not much else about these phenotypes made sense - there was absolutely no correlation between plaque size and viral growth rates, as often occurs with detrimental mutants, and no correlation between either and calculated free energies of stem loop secondary structures (Table 1.11). Additionally, things got even weirder with the two mutations that should have disrupted the secondary structure the most, 158A and to 173U, according to free energy calculations. Both grew fine relative to the WT when introduced into chicken cells, but always produced a two plaque phenotype in the same wells - one much smaller than the WT and one much larger. Plaque purifying and passaging either the small or large plaques only produces its respective plaque phenotypes, indicating that they were both unique genomes (Figure 1.24). Additionally, as Hubert describes, "These two mutants that grow well are also strange in that both.. the large and small plaque variants grow [more or less] similarly, and thus neither has a significant selective advantage over the other. Furthermore, the specific infectivity of the RNA transcript is the same (within the margin of experimental error) as that of wild-type transcripts, so that there is no reason to suspect that revertants of some sort are being selected from the transfection mix" [19]. As whole genome sequencing at the time was time-intensive and expensive, the source of any of these phenotypes is still unknown; "Sequence analysis of the genomic RNA of all four mutants between nucleotides 151 and 240 showed
that the original mutation was still present and that no other change was present in this region. The source of the plaque size variation is thus obscure. [Thus] both the large and small plaque variants have the predicted sequence within the 51-nucleotide region but clearly must differ somewhere else in the genome." [19]. Options to sequence the entire genome are cheap and plentiful today, and discovering the source of these weird plaque phenotypes shown in Figure 1.24 would be of interest and may provide powerful tools to further study viral RNA element mechanisms.

Figure 1.24. Weird Plaque Phenotypes of Sindbis Virus 51 Nucleotide CSE Mutants. Large plaque (LP) and small plaque (SP) variants were plaque purified from select 51-nt CSE mutants, which were predicted to highly disrupt RNA secondary structure, and assayed for growth in both vertebrate CEF and invertebrate C6/36 cells grown at 30°C. The authors describe the range of phenotypes which do not make much intuitive sense; "All four mutants grew well in both cell types, but for both 158A and 173U, the small plaque strains grew better than the large plaque strains in both cell types (thus the very large difference in plaque size is not correlated with the titer of virus produced). Intriguingly, 158A.SP grew even better than the wild type in both cell lines. In chicken cells, virus was produced at faster rates than for the wild type early in infection, but the rates of virus release for the mutant and wild type were the
same after 8 h. In mosquito cells, the rate of virus production by the 158A.SP mutant was about threefold greater than that of the wild type throughout the infection cycle. Mutant 173U.SP also outgrew the wild type in mosquito cells but not in chicken cells" [19]. The viral sequence for all four mutants shown is equivalent to the wild type virus at the 51-nt CSE. The remainder of these genomes have never been sequenced and therefore the source of these phenotypes is still unknown. Adapted from [19].

Hubert summarizes these studies of the 51-nt CSE using mutagenesis; "The results make it clear that the sequence in this region is quite important in some way for virus replication, as predicted from its conservation.. All of the mutants with silent mutations were viable [and] the relationship of plaque size to virus growth rate is also unpredictable. In the case of the two mutants 158A and 173U, the small plaque variant grows at a faster rate than the large plaque variant, which produces plaques that are six times as large. But even more strangely, all of the other mutants produce plaques that are wild type in morphology, despite the fact that most produce virus at a rate only 1%, and in extreme cases 0.01%, that of the wild type rate and many exhibit a pronounced lag in the growth curve. Thus, primary plaque morphology cannot be used to predict the growth properties of the virus other than that it is viable" [19]. More recently in 2018, Dr. Katrina Kutchko in the laboratory of Dr. Alain Laederach, along with the laboratories of Drs. Nat Moorman and Mark Heise at the University of North Carolina in Chapel Hill, confirmed the predicted secondary structure of this region using RNA SHAPE analysis (Figure 1.25 A) [166]. They found a significant defect in viral growth rates when a total of 20
synonymous mutations were introduced into the 51-nt CSE and neighboring regions of the Girdwood strain of Sindbis virus, when grown both in C6/36 mosquito cells and in vertebrate Vero cells (Figure 1.25 B) [166]. Many of the functions of these stem loops and the sources of the mutant phenotypes still remain a mystery, and consequently are ripe for further analysis with more modern tools.

Figure 1.25. RNA SHAPE Secondary Structure of Sindbis Virus 51 Nucleotide CSE. A) Experimental high throughput secondary structure analysis of Sindbis genomic RNA using RNA SHAPE confirmed previous computational secondary structure predictions of the two hairpins in the 51-nt CSE of Sindbis virus. Also shown is a large hairpin upstream which includes the viral nonstructural protein start codon. Stem loops are named SL3 and SL4 in reference to the two stem loops in the 5' UTR, and correspond to stem loops 1 and 2 in table 1.11. B) Growth curve of Sindbis virus Girdwood strain wild type (black) and RNA element mutants in Vero cells infected at an MOI of 0.01. Mut hairpin + CSE (orange): 20 total synonymous mutations were introduced in the nucleotides circled in blue in (A), in the nsP start-codon containing hairpin and the 51-nt CSE stem loops. Mut pack. sig. (blue): 70 total synonymous mutations were introduced into the viral packaging signal region described below. Adapted from [166].
In addition to the aforementioned functions, the viral RNA must also contain sequences that can be specifically recognized by the capsid protein to initiate the packaging of genomic RNA into nucleocapsids. For a contemporary treatment of this topic discussing (i) capsid protein packaging studies, (ii) RNA sequence packaging studies, and (iii) the multinucleation theory please see the excellent 2018 review "Alphavirus Nucleocapsid Packaging and Assembly", by Dr. Richard Kuhn's laboratory at Purdue University in Indiana [167]. Within an infected cell, the vast majority of RNA is cellular RNA, and primarily ribosomal RNA. Of the minority viral RNA fraction, during alphavirus infection there is many fold more subgenomic RNA than genomic RNA, along with a small fraction of dsRNA which contains minus-strand RNA. However, in virions, the vast majority of RNA is full-length genomic RNA [17, 168]. Therefore, the capsid protein specifically captures the minor species of genomic RNA due to some combination of RNA size, location, and specific sequences. Given the exclusive packaging of genomic RNA, these specific sequences were theorized to be in the 5' two-thirds of the genome that do not overlap with the subgenomic RNA. Testing the binding capacity of RNA fragments of various lengths from the 5' end of the Sindbis virus genome to purified capsid protein in vitro, by the Schlesinger laboratory, provided the first indication of the location of these sequences, termed the packaging signal [169]. There was a drop in binding when the 3' side of the truncated viral RNA ended at n.t. 1147 (BanII restriction site), and a further drop in binding ability
when the RNA was truncated at n.t. 765 (Avall restriction site), both within the
nsP1 region (Figure 1.26 A). RNA in vitro transcribed from an insertion of a
Sindbis viral fragment comprising nucleotides 683 to 1255 (Hhal restriction
site) into a bacterial plasmid vector effectively bound to purified capsid protein
(Figure 1.26 B).

Cell culture infections using defective interfering (DI) RNAs further refined the
packaging signal region in these initial studies. DI particles are partial
fragments of full-length viral RNAs that are effectively encapsidated and
released during infections, particularly high MOI infections, which can
effectively self replicate during coinfections with viruses that contain
complementary functions (helper viruses). Deletion of the Hhal fragment from
DI viruses allowed the fragment to replicate upon transfection with a helper
virus but prevented packaging of these DI RNAs; "The binding studies and the
in vivo packaging data, taken together, provide strong support for the
conclusion that there is a specific capsid recognition domain in Sindbis virus
RNA that plays a role in nucleocapsid assembly" [169]. Sequencing of a
Sindbis DI virus that is effectively packaged, DI25, showed that it contained
two copies of the genome region between nucleotides 726 and 1226, and this
region was initially considered to be the Sindbis virus packaging signal [169].
Furthermore, studies using chimeric alphaviruses indicated that this packaging
signal region in Sindbis nsP1 interacts with the capsid protein of several other
alphaviruses, and vice versa (Figure 1.26 C). Therefore, there seemed to be a conserved packaging mechanism in several alphaviruses, but not all alphaviruses, whereby capsid protein features can bind to the RNA packaging signal from a different virus to initiate encapsidation, perhaps through recognition of similar RNA secondary structures [170].

**Figure 1.26. Binding of Sindbis RNA Packaging Signal Region to Homologous and Heterologous Capsid Proteins.** A) *in vitro* binding assay of Sindbis Toto1102 *in vitro* transcribed RNA fragments from the indicated 5' region of the genome to purified Sindbis capsid protein. Binding is normalized relative to that of truncated DI25 RNA *in vitro* transcribed from a DI25 plasmid cut at the *SacI* site. B) Binding of RNA *in vitro* transcribed from bacterial pBR322 plasmid containing the chloramphenicol acetyltransferase (CAT) gene downstream of an SP6 promoter to purified Sindbis capsid protein. The *Hhal* fragment containing nucleotides 683 to 1225 (H572) from the plus strand (+) and the reverse complement from the minus strand (-) was inserted into CAT and assayed for binding. Binding is normalized relative to that of DI25 RNA used as control. C) Comparative replication in BHK-21 cells of Sindbis Toto1101
strain and a chimeric alphavirus with the nonstructural proteins of Sindbis (white boxes) and the structural proteins of Venezuelan equine encephalitis virus (black boxes). The structural proteins of VEEV, including the capsid protein, can efficiently recognize packaging signal in the 5' Sindbis portion of the genome and assemble and release infectious particles containing full-length chimeric genome RNA. Adapted from [169, 170].

Multiple sequence alignment and subsequent mutagenesis studies by the Frolov laboratory identified conserved GGG trinucleotides that lie apical on stem loops as critical for alphavirus nucleocapsid recognition and packaging [170]. In these set of elegant experiments, the packaging signal region of Venezuelan equine encephalitis virus (VEEV) strain TC-83, in nucleotides 856 to 1150, were copied and inserted downstream of nsP4 (Figure 1.27 A).

Introducing numerous synonymous mutations into the packaging signal at the original position still allowed virus growth, indicating that the packaging signal can function independent of its original position in nsP1, and providing the advantage that the cloned packaging signal can be mutated without regard for protein coding context (Figure 1.27 A, filled circle). Introduction of these numerous mutations to the cloned packaging signal was detrimental to the virus, and mutating only these GGG domains was equally detrimental (Figure 1.27 A, open circle and triangle, respectively). Reintroduction of a synthetic packaging signal that does not resemble the original one except for having stem loops with apical GGG trinucleotides then restored viral titers to nearly that of the wild type packaging signal (Figure 1.27 B) [170].
Figure 1.27. Alphavirus RNA Packaging Signal Mutagenesis Reveals Motifs Critical for Packaging Function. VEEV growth curves in BHK cells of viruses with dual packaging signals (PS). The PS region of VEEV TC-83 strain (nt 856-1150) was cloned and inserted downstream of nsP4, as indicated by the square boxes in each schematic. Then numerous synonymous mutations were introduced into the original packaging signal region (black squares), and the efficiency of packaging using the copy downstream of nsP4 was assayed by viral growth titer. White squares adjacent to nsP4 indicate the WT packaging signal at this position. A) The red square indicates a packaging signal with only three GGG stretches mutated (3xGm). Addition of these mutations to the PS reduced viral titers to the same levels as the addition numerous mutations to the PS. B) The red square indicated a synthetic PS that contains stem loops with apical GGG's and no other homology to the WT PS (artPS). Having only these domains from the WT PS in a similar secondary structure context nearly restores viral growth rate to WT. Adapted from [170].

This system was also used to study heterologous packaging signals - introduction of the nsP1-derived regions of Sindbis and eastern equine
encephalitis viruses into the construct shown in figure 1.27 grew as efficiently as that of VEEV, indicating that the VEEV capsid can efficiently recognize the packaging signal of these viruses [170]. Introduction of Semliki Forest virus (nsP1 region) and chikungunya virus (nsP1 and nsP2 regions), however, stunted growth indicating that the VEEV capsid does not efficiently recognize the packaging signal of these viruses. Interestingly, the vice versa case does not apply; keeping the VEEV packaging signal and replacing the VEEV structural proteins with those of chikungunya grew efficiently, indicating that chikungunya’s capsid can recognize the VEEV packaging signal (nsP1 region). Further investigation of this phenomena led to the finding that chikungunya’s capsid does not package its own nsP1 region efficiently but can package its own nsP2 fragment, indicating that chikungunya’s packaging signal lies within the nsP2 region instead of nsP1.

These studies, summarized in Table 1.12, seem to highlight a curious case of evolutionary divergence; "we hypothesize that a nsP1-specific [packaging signal] was present in an ancestral alphavirus. While the GGG motif-containing [stem loops] remained in the nsP1 region of most alphaviruses during their evolution, they disappeared in the genomes of [Semliki Forest virus]-like viruses and were replaced by a [packaging signal] located in the nsP2 gene" [170]. Consequently, capsid proteins of viruses in the Semliki Forest virus clade can still recognize the “ancestral”-esque packaging signal,
but not vice versa. The packaging signal regions from viruses in the Semliki Forest virus clade do not appear to contain GGG motifs, and it would be of interest to identify specific sequence motifs within the nsP2 region that aid packaging in these viruses, if they exist. Additionally, Aura virus has been shown to package its subgenomic RNA within its virions, and it would be interesting to see if this is due to another case of evolutionary divergence where the packaging signal has shifted to the 3’ end of the genome, or if the capsid protein of Aura virus has lost specificity for the genomic RNA, as described below [171, 172].

Table 1.12. Growth Properties of Chimeric Alphaviruses with Heterologous Packaging Signals and Capsid Proteins. Indicated genomic nucleotide regions from various alphaviruses were inserted into the VEEV construct described in figure 1.27, at the position downstream of nsP4, outside of its protein coding context. Additionally, a few constructs were generated where the structural proteins of the VEEV construct were replaced with those of chikungunya (CHIKV). Virus titer is assayed using chimeric virus grown on BHK cells at 10 hours p.i. Growth summary is arbitrarily chosen as efficient indicating titers about 10^7, and poor indicating titers below 10^6. * indicates that the region contains numerous synonymous mutations to disrupt packaging signal efficiency. Additional alphavirus abbreviations are listed in figure 1.16. Adapted from [170].

<table>
<thead>
<tr>
<th>Nucleotides</th>
<th>Packaging Signal</th>
<th>Capsid Protein</th>
<th>Virus Titer (pfu/ml)</th>
<th>Growth Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>774-1159</td>
<td>VEEV</td>
<td>VEEV</td>
<td>1.00E+08</td>
<td>Efficient</td>
</tr>
<tr>
<td>774-1159*</td>
<td>Mutated VEEV</td>
<td>VEEV</td>
<td>2.00E+05</td>
<td>Poor</td>
</tr>
<tr>
<td>731-1213</td>
<td>SINV</td>
<td>VEEV</td>
<td>5.00E+07</td>
<td>Efficient</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>---------</td>
<td>---</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>726-1185</td>
<td>EEEV</td>
<td>VEEV</td>
<td>2.00E+07</td>
<td>Efficient</td>
</tr>
<tr>
<td>797-1212</td>
<td>SFV</td>
<td>VEEV</td>
<td>2.00E+05</td>
<td>Poor</td>
</tr>
<tr>
<td>826-1182</td>
<td>CHIKV nsP1</td>
<td>VEEV</td>
<td>5.00E+05</td>
<td>Poor</td>
</tr>
<tr>
<td>2501-3078</td>
<td>CHIKV nsP2</td>
<td>VEEV</td>
<td>5.00E+05</td>
<td>Poor</td>
</tr>
<tr>
<td>774-1159</td>
<td>VEEV</td>
<td>CHIKV</td>
<td>1.00E+08</td>
<td>Efficient</td>
</tr>
<tr>
<td>774-1159*</td>
<td>Mutated VEEV</td>
<td>CHIKV</td>
<td>1.00E+05</td>
<td>Poor</td>
</tr>
<tr>
<td>826-1182</td>
<td>CHIKV nsP1</td>
<td>CHIKV</td>
<td>2.00E+05</td>
<td>Poor</td>
</tr>
<tr>
<td>2501-3078</td>
<td>CHIKV nsP2</td>
<td>CHIKV</td>
<td>5.00E+06</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

Whereas VEEV virus growth was reduced by a few logs in each of these cases with mutated GGG domains, neither mutations to the GGG domain nor the introduction of hundreds of mutations to the packaging signal were lethal, indicating the presence of a significant amount of functional redundancy across the viral genome. During the RNA SHAPE study described above, Katrina Kutchko introduced a total of 70 synonymous mutations to the packaging signal of Sindbis virus Girdwood strain, including all possible silent mutations to these apical GGG’s, and found that viral growth was similarly only decreased by a few logs, lending additional evidence to the existence of significant amounts of redundancy in the viral packaging mechanism (Figure 1.25 B, blue line) [166]. The aforementioned study also revealed the secondary structure of the Sindbis virus packaging signal, confirming the computationally predicted apical positions of these GGG trinucleotides (Figure 1.28) [166].
Figure 1.2. RNA SHAPE Secondary Structure of the Sindbis Virus Packaging Signal. Experimental high throughput secondary structure analysis of the Sindbis PS using RNA SHAPE. These confirmed previous computational secondary structure predictions indicating apical GGG trinucleotides on multiple stem loops, highlighted in green circles. Viral growth curve of a Sindbis PS variant with 70 total synonymous mutations introduced to the nucleotides circled in blue is shown in figure 1.25 B (blue line). Adapted from [166].

Parallel studies, primarily conducted in Richard Kuhn’s laboratory, analyzed the role of the capsid protein in packaging. While the C-terminal region of the capsid protein contains the autoprotease, the N-terminal region is thought to be involved in packaging and dimerization [167]. The inability to crystallize this N-terminal region indicates that it is likely disordered. Studies have subdivided
the N-terminal region by function (Table 1.13). Region I is highly positively charged and interacts with the negatively charged viral RNA, and also contains a helix that facilitates dimerization - the capsid protein was still functional upon replacement of the helix with a yeast protein dimerization domain. Mobility shift assays revealed that region II is primarily involved in interaction with the RNA packaging signal [168].

**Table 1.13. Sindbis Virus Capsid Protein Domains.** A description of the various domains of the Sindbis virus capsid protein and their functions - the N-terminal domain has evaded structural analysis due to the inability to crystallize the region, likely owing to it being quite flexible. A portion of region II (amino acids 99-114 in Sindbis virus) is highly conserved among all alphaviruses. CP = capsid protein, PS = packaging signal, cdE2 = C-terminus of E2 protein. Adapted from [167].

<table>
<thead>
<tr>
<th>Domain</th>
<th>Region/Sub-Domain</th>
<th>Amino Acids</th>
<th>Proposed Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-terminal</td>
<td>Region I</td>
<td>1–81</td>
<td>Specificity of RNA packaging</td>
</tr>
<tr>
<td></td>
<td>Helix I</td>
<td>38–55</td>
<td>CP dimerization</td>
</tr>
<tr>
<td></td>
<td>Region II—conserved sequence</td>
<td>81–114</td>
<td>Interaction with RNA PS</td>
</tr>
<tr>
<td>C-terminal</td>
<td></td>
<td>114–264</td>
<td>Protease domain, interaction with cdE2</td>
</tr>
</tbody>
</table>

Deletion and mutagenic analyses revealed that region II was involved in the packaging selectivity of full-length genomic RNA [168]. Whereas the wild type virion contains predominantly genomic RNA, various deletions and mutations also allowed the packaging of subgenomic RNA (Figure 1.29 A). In particular, deletion of a partially highly conserved region results in the packaging of 11 times more subgenomic RNA than genomic RNA (Figure 1.29 A, Δ97-106).
Individual mutations within and outside this region have a far more modest
effect on the genomic:subgenomic RNA ratio, indicating a high level of
redundancy [168]. Aside from using deletions, the proportion of subgenomic
RNA can also be significantly increased by incorporating a large number of
mutations to partially overcome this redundancy; "One of the most compelling
results was that mutagenesis of all the positively charged amino acids [in the
N-terminus] still resulted in virus particle formation, albeit ones that packaged
predominantly subgenomic RNA. These particles were termed
pseudoinfectious viruses owing to their dramatically decreased infectivity..
both in vivo and in vitro studies have thus far concluded that [nucleocapsid]
assembly is not a reaction simply mediated by a [capsid protein and packaging
signal] interaction. The subdomains in the N-terminal domain appear to govern
many different reactions from selecting the RNA and binding adjacent [capsid
protein] molecules as well as sending the protein to the correct intracellular
location. Both specific and non-specific interactions are necessary in this
regard. The C-terminal protease domain is not passive in the assembly
process since it governs the geometry and symmetry of the particle but can
also drive assembly when the N-terminal domain is mutated significantly"
[167]. Therefore, studies of the RNA packaging signal and capsid protein
indicate a high level of redundancy for the packaging mechanism, on both the
RNA and protein levels.
Figure 1.29. Sindbis Subgenomic:Genomic RNA Ratio in Virions and Multinucleation Theory of Packaging. A) Molar ratio of subgenomic RNA to genomic RNA for Sindbis virus Toto64 (WT) and a variety of capsid protein N-terminus deletion and substitution mutants. Grey dashed line shows the intracellular ratio of subgenomic to genomic RNA in infected BHK cells at 10 hours p.i. The copy number of subgenomic and genomic RNA's were assayed using qRT-PCR on BHK cell infection supernatants, on three biological replicates. Modern next-generation sequencing on highly purified supernatant virions can additionally reveal percentages of nonspecific cellular RNA incorporation. Adapted from [168]. B) Schematic of the two-step multinucleation theory for nucleocapsid assembly. Studies using single molecule fluorescence spectroscopy on in vitro assembly reactions using MS2 bacteriophage and satellite tobacco necrosis virus revealed that the first stage in nucleocapsid assembly involves compaction of the viral RNA's hydrodynamic radius. In the second stage, further capsids "follow the [Hamiltonian energetically favorable] path laid out by the RNA to form fully assembled particles. This hypothesis shifts away from traditional capsid-centric based assembly ideas towards an RNA-centric one. The hypothesis can be compared to stapling together proteins from the inside. Each [packaging signal] acts as a staple on a path, which influences the path itself as well as future [capsid-capsid] interactions" [167]. Further studies in hepatitis B, hepatitis C and picornaviruses have added evidence that this theory may be broadly applicable to viruses. However, such studies have not yet taken place for alphaviruses. Multiple packaging signals of equal strengths have been identified for some of the viruses in which these studies have been conducted, which differs in alphaviruses. CP = capsid protein, PS = packaging signal (highlighted in red). Adapted from [167].
More recently, the use single molecule fluorescence spectroscopy on *in vitro* assembly reactions of the MS2 bacteriophage, satellite tobacco necrosis virus, hepatitis B virus, hepatitis C virus and several picornaviruses have led to the proposal of a two-stage nucleocapsid assembly mechanism which may be a more broadly applicable paradigm [167]. In this model, initial binding by a few capsids to the viral RNA to specific sites is followed by nonspecific binding of additional capsid molecules to the viral RNA, which continually compacts the RNA as the nucleocapsid forms (**Figure 1.29 B**). For alphaviruses, this hypothesis would propose specific capsid binding to the packaging signal to initiate nucleation, followed by multiple sites of nonspecific electrostatic neutralization between primarily the N-terminal region of the capsid protein and the viral RNA to drive the rest of the packaging process. Indeed, studies using purified Sindbis virus capsid proteins have shown that they can nonspecifically encapsidate a number of negatively charged macromolecules to form nucleocapsid-like structures *in vitro* [173]. However, this remains a proposal as the requisite experiments have not yet been performed using alphaviruses. Thus, details of the alphavirus packaging process remain elusive; "For many, the question of how a [nucleocapsid] forms is a fundamental one. To some this is seen as a potential antiviral target independent from the host cell. Our view is that virology has always been a tool for understanding the basic building blocks of biology and we see that the
molecular machinery of how viruses are created as another means for appreciating the world within a cell" [167].

1.10. Further Discovery Using Phenotypic Screens and Concluding Remarks.

In addition to the aforementioned low methionine medium screen which allowed mapping of low-methionine resistance to mutations to nsP1, additional phenotypic screens elucidated portions of the Sindbis genome that were involved in establishment of the respective screened phenotypes. Noncytopathic screens in the Schlesinger laboratory, whereby BHK cells were infected with Sindbis virus containing high amounts of DI particles and viable cells were selected, identified a proline in nsP2 residue 726 as critical for inducing cell death [174]. Studies using a substitution of a valine at this position (nsP2-726V) indicated that the wild type residue was involved in degradation of RPB1, the catalytic subunit of cellular RNA Polymerase II, which in turn induces cellular transcriptional shutoff [175]. These studies also indicated that the wild type residue is involved in translational shutoff via PKR-dependent and independent mechanisms. Infections with the nsP2-726V mutant also upregulated interferon production, indicating that nsP2 is also involved in mitigating the cellular innate immune response. Additionally, a variety of substitutions at the nsP2-726 position drastically decreased viral replication rates, indicating that this position is also critical for viral RNA
In order to further tease apart these functions of nsP2, the Frolov laboratory performed a similar screen, in which nsP2-726 substitutions arose multiple times, but also selected for high viral replication by fusing GFP to nsP2 and isolating bright cells [175]. Among multiple hits, they discovered that substitutions of a proline in nsP2 residue 683 to either serine, asparagine or glutamic acid abolished the cellular transcriptional shutoff phenotype, shown by the lack of RPB1 degradation (Figure 1.30 A), and also induced high levels of interferon. However, they all retained the translational inhibition and cytopathic phenotypes akin to the wild type virus (Figure 1.30 B), had similar nsP2 localization as the wild type, and also replicated to the same levels as the wild type. Therefore, the transcriptional shutoff phenotype in isolation was able to be mapped to nsP2-683 [175]. Using the nsP2-683S mutant and screening for noncytopathic cells identified a deletion of residues 24-29 in nsP3, which in combination with nsP2-683S makes the virus less cytopathic in immune-competent cells and noncytopathic in immune-deficient cells. In contrast with the nsP2-683S single mutant, the double mutant failed to induce cellular translational shutoff (Figure 1.30 B). However, the double mutant also induced interferon similarly to the single mutant, and also replicated to the same levels and produced similar amounts of nsP2 with the same localization as the wild type and the single mutant viruses [175]. A double mutant with a
substitution of an asparagine in nsP3 residue 24 to an alanine instead of the deletion in nsP3 mirrored the same phenotypes. Therefore, a specific cellular translational inhibition function was mapped to nsP3-24. This nsP3 mutation alone does not produce a detectable phenotype, and needs to be in combination with the nsP2-683 substitution for detection. This residue is part of the nsP3 macrodomain which functions as a mono-ADP-ribosylhydrolase, and nsP3-24A was shown to abolish hydrolase activity, which postulates a link between the hydrolase activity of alphaviruses and cellular translational shutoff [175].

Figure 1.30. Effect of Viral Mutations on Cellular Transcriptional and Translational Shutoff. Mutations in nsP2 and nsP3 can render Sindbis virus incapable of inducing cellular transcriptional and translational inhibition. Mutations in nsP2-726 are incapable of either and are noncytopathic, and produce less viral RNA than the wild type. Several mutations in nsP2-683 prevent transcriptional shutoff, and double mutants with an additional deletion
of nsP3 residues 24-29 or substitution of nsP3-24 with alanine prevent translational shutoff, whilst replicating to the same levels as the wild type. NIH 3T3 cells were infected with the indicated virus at an MOI of 20 for both panels. A) RNA was radiolabeled using $[^{3}H]$uridine between 3 and 7 hrs p.i. and run on an agarose gel. The positions of prominent viral RNA and cellular RNA bands are indicated. B) Proteins were radiolabeled using $[^{35}S]$methionine at 6 hrs p.i. for 30 minutes and run on a SDS-PAGE gel. Prominent viral and cellular protein bands are indicated. Adapted from [175].

In the past couple decades there has been much focus on cellular host factors that aid and control alphavirus infection, perhaps due to the advent of robust cellular genetics technologies which enable their study and consequent interest from funding agencies; "genome-wide genetic screens, including loss of function screens, microarrays, RNA-sequencing, and candidate gene studies, have further elucidated the role host genetics play in the response to virus infection, with the immune response being found in particular to majorly influence the outcome" [176]. For example, such screens conducted in our laboratory have identified stress granule proteins G3BP1 and G3BP2 and innate immune factors ZAP and TRIM25 as important host factors in controlling alphavirus infections [177–179]. The reader can find contemporary information on this rapidly expanding list in the 2018 reviews "Genetic Control of Alphavirus Pathogenesis" and "Innate Immune Control of Alphavirus Infection", and the 2019 review "New World Alphavirus Protein Interactomes From a Therapeutic Perspective" [176, 180, 181]. Due to the relatively recent nature of many of these cellular genetics tools, information on host factors in older reviews is generally far more limited.
Most papers in the field start with a paragraph or two listing what is known about alphaviruses. This chapter was written with the goal of informing the reader how they became known and putting those discoveries in context. A tremendous amount of information as it pertains to features in the viral genome that facilitate infection have been revealed through these described efforts. However, it should also be apparent that many of the tools available today were not present when many of these discoveries were made. The goal of our study is to expand our knowledge of the viral genome by using modern, powerful next-generation sequencing methods on populations of the virus and seeing how they evolve. The following chapters in this thesis dissertation detail our efforts therein. Finally, if the reader has made it this far, I would like to take a moment to offer my congratulations and thanks for taking an interest in the history of alphavirology. Even if one were forced (strongly suggested) to, I hope it was as much a pleasure to read as it was me to write.
1.1. Chapter I References.


41. DICK GWA (1952) Zika virus. II. Pathogenicity and physical properties. Transactions of the Royal Society of Tropical Medicine and Hygiene, 46(5):521–534.


44. Hurlbut HS (1940) A study of the larval and pupal chaetotaxy and the eggs of the four common inland species of anophelines in southeastern United States with notes on their biology.


47. MANSHARAMANI HJ, DANDAWATE CN, KRISHNAMURTHY BG, NANAVATI AN, JHALA HI (1965) EXPERIMENTAL VACCINE AGAINST KYASANUR FOREST DISEASE (KFD) VIRUS FROM MOUSE BRAIN SOURCE. Indian journal of pathology & bacteriology, 12:159–177.


144. Lemm JA, Durbin RK, Stollar V, Rice CM (1990) Mutations which alter the level or structure of nsP4 can affect the efficiency of Sindbis virus replication in a host-dependent manner. *Journal of Virology*, 64(6):3001–3011.


Chapter II. Methods: How to Purify a Virus.

Practical Considerations in the Isolation of Select Arbovirus Virions and Viral RNA.

2.1. Introduction.

I have spent quite some time during my PhD training enriching viruses to aid in downstream analyses wherein sample purity is paramount. Herein I list the various considerations I have empirically learned are important to such isolation, many such details that are outside the scope of what is typically espoused in a methods section. The primary goal of such exposition in this chapter is twofold; 1) so a reader may be able to troubleshoot their own purification experiments when they run into problems while following the paper's methods section, and 2) more importantly, that they may be able to effectively adapt the methods and successfully enrich other viruses and target RNA of interest without necessitating personal consultation.

Oftentimes at the very core of successful study of a target biomolecule is the successful enrichment or isolation of that molecule. To that end, I relate one such anecdote from a student lunch I had with Dr. Bruce Buetler of University of Texas Southwestern, winner of the 2011 Nobel prize in physiology or medicine. When asked what led to his discovery, he stated, "I was just the guy in lab who was known for being good at purifying stuff from mice, so they
asked me to purify this particular thing people were interested in and having trouble with, and I did, and it turned out to be important”.

Viral enrichment within this chapter falls within two main categories:
1. Isolation of intracellular viral RNA from the surrounding cellular RNA.
2. Concentration of packaged virions from cell culture supernatant into small volumes, and subsequent isolation of pure viral RNA from these virions.

To quickly determine whether the following will be useful to the reader, I start with an example of the expected results from each type of enrichment procedure (Figure 2.1), as perhaps more efficient methods have become available since the writing of this chapter, or increases in sequencing throughput has rendered target enrichment unnecessary:
Figure 2.1. Viral RNA Enrichment Efficiency of Oligonucleotide Target Capture and Polyethylene Glycol Precipitation. A) Enrichment of Zika viral RNA from infected Huh 7.5 cells using two rounds of our oligo capture method. The method allows the enrichment of target viral RNA to 98+ % in the post-capture RNA pool, relative to the starting RNA extracted from infected cells in which the vast majority is human cellular RNA. Percent of each type of RNA in the RNA pool (y-axis) is detected by qPCR in the pre-capture sample and by next-generation sequencing read alignments in the post-capture sample, and the number indicates the percent of viral RNA in each sample. For analysis of Zika RNA, $987 of a typical $1000 MiSeq sequencing run will go towards the intended target if the post-capture sample is sequenced. If the infected cell RNA is sequenced without enrichment, as is commonly done in the literature, only $17 of a $1000 run will go toward the intended target. B) Retention of Sindbis virus Toto1101 strain virions in a PEG precipitation reaction, which allows the volume of cell culture media, which contains the virions released by infected cells, to be reduced thousand-fold. Such concentration is essential for effective further processing. The y-axis represents the total amount of viral RNA extracted from virions produced by a confluent Huh 7.5 cell monolayer in a 500 cm2 tissue-culture glass plate (input) and after concentration (precipitate). Error bars represent SEM.
Zika virus Puerto Rico strain

Percent RNA (of Total RNA Sample)

- Pre-Capture: 1.7%
- Post-Capture: 98.7%

Cellular RNA  Viral RNA
B) Sindbis virus packaged virion RNA

P.E.G. Precipitation
1,000-fold Volume Reduction

Nanograms Viral RNA

Input
Precipitate

90%
It seems to me that there is no reason to believe that such efficiencies cannot be achieved for other viruses or RNA targets. Several examples of similar efficiencies achieved with other arboviruses are provided within the text.

Importantly, both methods are very cost-efficient and consequently highly scalable. Typical dollar values of key reagents is listed below and expanded within the text (excluding stock reagents such as salts for buffers, which I assume will not fluctuate much in price). Also provided are examples of kits available for purchase and commonly used in contemporary literature, which as of 2019 are orders of magnitude more expensive per reaction than these methods.
Table 2.1. Cost Comparison of Oligonucleotide Capture RNA Enrichment Method. The per-reaction (\$/rxn) cost of key reagents in our enrichment method is listed and compared to those of the Agilent SureSelectXT kit, which is currently the most common RNA enrichment method used in the literature for similar purposes. The nominal cost of buffers and salts is excluded in our calculations, which may increase the per reaction cost by a few cents. Prices listed are from 2016, after which prices of the SureSelectXT kit is no longer publicly advertised and a quote must be requested. Further details are within the text of this chapter.

<table>
<thead>
<tr>
<th>Method</th>
<th>Baits / Rxn</th>
<th>Beads / Rxn</th>
<th>Total Cost / Rxn</th>
<th>Rxn's Per $1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Enrichment Method</td>
<td>$0.60</td>
<td>$19.80</td>
<td>$20.40</td>
<td>49</td>
</tr>
<tr>
<td>Agilent SureSelect XT</td>
<td>$365.63</td>
<td>$72.13</td>
<td>$437.76</td>
<td>2.3</td>
</tr>
</tbody>
</table>

Table 2.2. Cost Comparison of Polyethylene Glycol Based Virus Concentration. The cost of an equivalent volume of the key solution in our PEG concentration method is listed and compared to that of ABCAM’s PEG Virus Precipitation Kit (Product # ab102538). The nominal cost of buffers and salts is excluded in our calculations, which may increase the per reaction cost by a few cents. Further details are within the text of this chapter.

<table>
<thead>
<tr>
<th>Method</th>
<th>Amount of 5x PEG Solution</th>
<th>Price</th>
<th>Media Concentrated / $1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCAM PEG virus kit</td>
<td>125 ml</td>
<td>$395.00</td>
<td>1.3 liters</td>
</tr>
<tr>
<td>Our PEG Protocol</td>
<td>125 ml</td>
<td>$1.53</td>
<td>326.8 liters</td>
</tr>
</tbody>
</table>

Whilst this chapter will be more detailed than my memory, the reader is welcome to send questions, comments, success announcements etc. to pmambrose@gmail.com.
2.2. Quantitative PCR Assays for Sindbis virus, Zika virus, and Human Cellular RNA.

For the development of an enrichment procedure, it was critical to have an accurate assay that measures the amount of target RNA we hope to retain, and the amount of non-target RNA that we hope to remove. Additionally such assays are also useful in detecting successful infections before embarking on downstream applications, in addition to a variety of applications in virology. After several attempts, we found that qPCR using the Taqman system gave us a far better dynamic range than protocols based on intercalating dye (such as SYBR Green), whereby dynamic range is defined by the segment of the standard curve that is linear and above where the water control produces a signal (Figure 2.2). For viruses, the standard curve is generated using aliquots of the in vitro transcribed RNA, the same RNA that is used for electroporation into cells to produce the virus. The in vitro transcribed RNA is quantified using a nanodrop spectrophotometer and diluted ten-fold from the maximum dynamic range value.
Figure 2.2. Quantitative PCR Standard Curves of Zika, Sindbis and Cellular RNA. The cycle number (y-axis) corresponding to dilutions of RNA (x-axis) used for standard curve generation, of Sindbis Toto1101 in vitro transcribed RNA (A), Zika Puerto Rico Strain in vitro transcribed RNA (B) and Huh 7.5 Total Cellular RNA (C). Cycle numbers are calculated as the second derivative peak values using Roche Lightcycler 480 Software. The dynamic range, wherein the standard curve is linear, is indicated with a red line. The amounts of RNA in query solutions are calculated against a standard curve run concurrently on the same plate in every assay. SEM values were calculated using technical replicates but error bars are too small to be displayed.
18s rRNA qPCR Standard Curve

Cycle Number

Nanograms RNA / Well (Total Huh 7.5 Cell RNA)
For Sindbis virus Toto1101 strain, Dr. Hachung Chung designed the following primers which give a fantastic dynamic range, between 1 picogram per well and 100 nanograms per well in a 96 well plate. The standard curve is made by using ten-fold dilutions of \textit{in vitro} transcribed Sindbis Toto1101 RNA starting with 100 ng per well (50 ng/ul).

Forward Primer: GGTAGCTCATTGGGACAACA (Sense)
Reverse Primer: GCTGGAACACCGGAATCTA (AntiSense)
Taqman Probe: /56-FAM/TGGCGTGATCGTACCCATACTTGC/36-TAMSp/ (AntiSense)

For Zika virus, Dr. Joseph Luna acquired the following primers that can detect multiple strains of Zika effectively, which gives us a dynamic range between 1 picogram per well and 10 nanograms per well in a 96 well plate [1]. The standard curve is made using ten-fold dilutions of \textit{in vitro} transcribed Zika virus starting with 10 ng per well (5 ng/ul).

Forward Primer: AACTCCACAYTGGAACAACAARG (Sense)
Reverse Primer: CWCCATCCATCTCAGCCTCYA (AntiSense)
Taqman Probe: /56-FAM/GG+AGCHGTTCAYA+C+GG/3IABkFQ/ (AntiSense)

Note: “+” precedes LNA base .R = A or G, Y = C or T, W = A or T.

To detect cellular RNA, we decided on using primers against ribosomal RNA (rRNA) as a proxy, which is the most abundant RNA species in total cellular RNA. We built a standard curve based on dilutions cellular RNA extracted
from uninfected Huh 7.5 cells, the cells we grow Zika virus and Sindbis virus in. We achieved best results using Eurogentec's FAM-TAMRA 18s rRNA Control Kit (Taqman based, Eurogentec Product # RT-CKFT-18S) which gives us a dynamic range between 50 nanograms per well and 5 picograms per well of Huh 7.5 total cell RNA (Figure 2.2 C). We tested all rRNA qPCR kits that were commercially available at the time, and most of the others were rather terrible. The primer sequences are proprietary. Given that rRNA is the most abundant RNA species we reasoned that it would provide a good proxy for total cellular RNA, in the absence of ribodepletion or viral infection drastically affecting the 18s rRNA to total cellular RNA ratio. Good agreement between our cellular RNA quantitation and next-generation sequencing values indicate that such drastic rRNA ratio disregulation is not the case.

Several dilutions of the query RNA solution is made and assayed to ensure that some dilutions fall within the dynamic range. Our assays all used the Roche Lightcycler 480 RNA Master Hydrolysis Probes kit (Roche Product # 04991885001) and the Roche Lightcycler 480 machine. Each well of a 96 well plate included 2 ul of query RNA sample, 6.3 ul of d.d.H2O, 1.3 ul of Roche Activator, 7.4 ul of Roche MasterMix, and 1 ul each of 10 uM resuspensions of the forward primer, reverse primer and Taqman probe. The forward and reverse primers come pre-mixed in the Eurogentec kit, and thus 2 ul of the primers and 1 ul of the probe was added in those instances. The reverse
transcription and amplification cycle details, designed by Dr. Hachung Chung, for all assays is provided in Figure 2.3.

<table>
<thead>
<tr>
<th>Programs</th>
<th>Reverse Transcription</th>
<th>Denaturation</th>
<th>amplification</th>
<th>cooling</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cycles</strong></td>
<td>1</td>
<td>1</td>
<td>45</td>
<td>1</td>
</tr>
<tr>
<td><strong>Analysis Mode</strong></td>
<td>None</td>
<td>None</td>
<td>Quantification</td>
<td>None</td>
</tr>
<tr>
<td><strong>Target (°C)</strong></td>
<td>Acquisition Mode (hh:mm:ss)</td>
<td>Ramp Rate (°C/s)</td>
<td>Acquisitions (per °C)</td>
<td>Sec Target (°C)</td>
</tr>
<tr>
<td>63</td>
<td>None</td>
<td>00:03:00</td>
<td>4.40</td>
<td>0</td>
</tr>
<tr>
<td>95</td>
<td>None</td>
<td>00:00:30</td>
<td>4.40</td>
<td>0</td>
</tr>
<tr>
<td>95</td>
<td>None</td>
<td>00:00:15</td>
<td>4.40</td>
<td>0</td>
</tr>
<tr>
<td>60</td>
<td>Single</td>
<td>00:00:30</td>
<td>2.20</td>
<td>0</td>
</tr>
<tr>
<td>72</td>
<td>None</td>
<td>00:00:01</td>
<td>4.40</td>
<td>0</td>
</tr>
<tr>
<td>40</td>
<td>None</td>
<td>00:00:10</td>
<td>2.20</td>
<td>0</td>
</tr>
</tbody>
</table>

**Figure 2.3. Quantitative PCR Program Details for One Step Assay.** The temperature cycles we used for the Sindbis, Zika and 18s rRNA assays in the Roche 480 Lightcycler machine. In summary, the RNA is reverse transcribed to DNA in 63°C for 3 minutes and denatured in 95°C for 30 seconds. Then 45 minutes of PCR amplification occurs with 95°C for 15 seconds, 60°C for 30 seconds during which data acquisition takes place, and 72°C for 1 second. After the cycles the plate is cooled at 40°C for 10 seconds to make it safe for removal from the machine.
2.3. Introduction to Oligonucleotide Based Hybridization Capture.

This section will be presented to the reader as follows - I begin with a brief description of the method and a short protocol, outlining the precise steps to isolate Sindbis and Zika viral RNA, along with a discussion of materials and steps used. The aforementioned materials and steps that are not explained further happened to work and did not require further adjustment, and may leave room for experimenting with alternatives or optimization. Following these steps precisely will yield Sindbis virus genomic RNA (Figure 3.32) or Zika Puerto Rico strain RNA (Figure 2.1 A) to 95+% purity in the final RNA pool that is to be sequenced using next-generation sequencing, or alternatively characterized.

This will be followed by an exposition of various findings and trials during the development of the method that will alert the reader to what we discovered to be critical parameters for enrichment. This should allow the development an enrichment protocol targeting any RNA of the reader's interest without stumbling through the multitude of issues we encountered during our development phase. The goal of this section is to provide the reader with the capabilities to develop and optimize a cost-effective and scalable enrichment protocol for any target RNA molecule that will enhance subsequent analyses.
2.4. Extraction of RNA from Infected Cells: A Short Protocol.

These are the steps we followed to extract infected cellular RNA ready for enrichment. Our protocol is designed for a 500 cm2 dish with approximately 40 million infected Huh-7.5 cells at time of harvest (Corning 500 cm2 Square TC-treated Culture Dish, Product #431110). This can be scaled accordingly. This protocol was originally designed by Dr. Yingpu Yu, with a few changes by me.

Acronyms:

- TRIS - tris (hydroxymethyl) aminomethane (used as 10 mM, 100x dilution in distilled deionized H2O of Tris 1M pH 8.0, RNase-free, Ambion Product # AM9855G). We usually do not adjust the pH after dilution.
- DPBS - Dulbecco’s Phosphate Buffered Saline (Gibco Product # 14190-144).
- DMEM - Dulbecco's Modified Eagle Medium (Gibco Product # 11905-065) + 5% Fetal Bovine Serum (Hyclone, GE Healthcare Product # SH30071.03) & Nonessential Amino Acids (Gibco Product # 11140-076).

Note: *The DMEM mix is just the media used to grow and maintain the Huh 7.5 cells and had handy, and can likely be substituted with whichever cellular growth medium is used for the cells the RNA is being extracted from. The specific product numbers above represent what our cell culture facility had in stock and can likely be substituted with similar available reagents.*
**Key:**

+ I Add Reagent
- I Remove Supernatant
* I Centrifuge or Shake Step
^ I Repeat from Indicated Step
> I Transfer to Indicated Tube
= I Incubation Step

---

**A | Trizol Cell RNA Extraction**

To Tray (p500), after media is removed:

A.01 I + 50 ml DPBS -> wash

A.02 I - Remove Supernatant

A.03 I + 20 ml Trypsin (0.05% with EDTA, Gibco Product # 25300-062)

A.04 I = 10 min @ 37°C 5% CO2 incubation

A.05 I + 20 ml DMEM

A.06 I > Transfer -> 50 ml Conical

A.07 I * 5 min 4°C spin @ 1,300 rpm (400 g on our Allegra 6R GH-3.8 tabletop)

A.08 I - Remove Supernatant

A.09 I + 30 ml DPBS -> mix to wash

A.10 I * 5 min 4°C spin @ 1,300 rpm (400 g on our Allegra 6R GH-3.8
A.11 \( | \) - Remove Supernatant

A.12 \( | \) + 200 ul DPBS \( | \) -> pipette mix thoroughly

A.13 \( | \) > Transfer \( | \) > PhaseLock Gel Tube (Yellow, 2 ml)

*Note: 5Prime phaselock tubes, which is essentially RNase-free gasoline jelly, make phase separation while using Trizol, phenol or chloroform easier, but it is not necessary. The company was recently acquired by Qiagen and these have been renamed to MaXtract (Qiagen MaXtract High Density 200 x 2 ml, Product # 129056).*

B \( | \) Trizol RNA Phase Extraction

To \( \sim \)200 ul Cells in DPBS:

B.01 \( | \) + 800 ul Trizol Reagent

*Note: Trizol is a brand name for phenol with guanidium thiocynate added, we had plenty in hand so chose to continue using it instead of mixing our own (Invitrogen Product # 15596026).*

B.02 \( | \) * 15 second Shake (with hand)

B.03 \( | \) = 5 to 10 min @ room temperature incubation

[Optional Pause] Overnight @ \(-80^\circ \text{C}\) incubation

B.04 \( | \) + 250 ul Chloroform:IAA

B.05 \( | \) * 15 second Shake (with hand)

B.06 \( | \) = \( \sim \) 10 min @ room temperature incubation
B.07 | * 10 min RT spin @ 12,000 rpm

B.08 | > Transfer Top Phase (Aqueous Phase, ~ 750 ul) -> 2 ml Eppendorf Tube

C | Qiagen RNAeasy Columns RNA Isolation (Critical!).

To Aqueous Phase:

C.01 | + 700 ul Ethanol 70% (total volume now should be ~1.45 ml)

C.02 | > Immediately transfer 700 ul -> RNAeasy Spin Column

C.03 | * 15 second RT spin @ 12,000 rpm -> Empty Spin Tube

C.04 | ^ Repeat from Step C.02 until all volume has passed through column (~2x)

**Note:** Remainder Steps are from Qiagen RNeasy Kit Protocol (Qiagen Product # 74104):

C.05 | + 500 ul Buffer RPE (EtOH Added Wash)

C.06 | * 15 second RT spin @ 12,000 rpm -> Empty Spin Tube

C.07 | + 500 ul Buffer RPE (EtOH Added Wash)

C.08 | * 2 min RT spin @ 12,000 rpm -> Empty Spin Tube

C.09 | * 1 min RT spin @ Max Speed (To Dry Membrane)

C.10 | > Transfer Spin Column -> 1.5 ml Eppendorf Tube

C.11 | + 50 ul 10 mM TRIS pH 8 -> Center of Spin Column

C.12 | = 1 min @ room temperature incubation

C.13 | * 1 min RT spin @ 12,000 rpm
C.14 | > Transfer ~ 50 ul RNA Eluate -> Center of Spin Column

C.15 | = 1 min @ room temperature incubation

C.16 | * 2 min RT spin @ 12,000 rpm -> Discard Spin Column

D | Characterization

D.01 | Quantify total RNA amount with spectrophotometer (eg. Nanodrop).

Given the above number of input cells, we typically get multiple micrograms per microliter of total infected cellular RNA.

D.02 | Dilute to ~1.5 ug/ul with 10 mM TRIS pH 8. Input per reaction is ~50 ug total RNA in 35 ul, and input for 2 round capture is ~150 ug, explained further below.

D.03 | Optional: qPCR viral RNA (or other RNA target) to get percent of target RNA in total cell RNA. This will give a good idea of whether the infection was successful and within the ranges we get (described below), which would set the stage for a high final ratio upon enrichment.
2.5. Enrichment of Target RNA from Total RNA: A Short Protocol.

These are the steps we followed to enrich viral RNA from total infected cellular RNA, purified using the RNA extraction protocol detailed above. Specific probe sequences for hybridizing Sindbis virus and Zika virus are listed in the next section. The following protocol has been effective for every viral RNA target lab members have attempted since, and should be similarly effective against non-viral RNA targets as well.

**Reagents:**

Invitrogen Dynabeads Streptavidin Myone C1 Magnetic Beads (Thermo Fisher Product # 65001)

DynaMag-2 Magnetic Stand (Thermo Fisher Product # 12321D)

Tris pH 8, 1M (Ambion Product # AM9855G)

EDTA pH 8, 0.5M, Fluka Bioultra Grade (Sigma Adlrich Product # 03690)

Sodium Chloride, 5M, molecular biology grade (Promega Product # V4221)

Sodium Acetate pH 5.5, 3M (Ambion Product # AM9740)

Glycogen, RNA grade (Thermo Fisher Product # R0551)

CoStar Spin-X 0.22 um Cellulase Acetate 2 ml Centrifuge Tube Filters

RNase/DNase Free (Corning # CLS8160)

Ethanol, Anhydrous 200 Proof (Decon Labs Product # 2716)

2-Propanol (Isopropanol), Bioreagent for Molecular Biology (Sigma Aldrich Product # I9516)
Distilled deionized water (ddH2O) was purified using a Millipore Milli-Q Synthesis system and Q-gard 1 cartridges (Millipore Sigma #QGARD00R1).

Note: the above is the list of specific brands we used. With the exception of those described as important below, many of them can likely be replaced with what the user has in hand. Most of the above are stated to be tested RNase-free (molecular biology grade), which is preferable. The rather expensive magnetic stand is what we already had in lab, but one can likely assemble a similar stand using neodymium magnets or buy a 3D-printed stand on eBay for a small fraction of the cost (search for "Magnetic Beads Separator Rack Stand").

Solutions (And Preparation Using Reagents Listed Above).

1. Bind / Wash Buffer (2x): 10 mM Tris pH 8, 1 mM EDTA, 2 M NaCl (Table 2.3).

Table 2.3. Preparation of 2x Bind & Wash Buffer. To prepare 50 ml's of 2x Bind/Wash Buffer, mix the following components to a homogenous solution:

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Final Concentration</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tris pH 7.5 (1M Stock)</td>
<td>10 mM</td>
<td>+ 0.5 ml</td>
</tr>
<tr>
<td>EDTA pH 8 (0.5M Stock)</td>
<td>1 mM</td>
<td>+ 0.1 ml</td>
</tr>
<tr>
<td>Sodium Chloride (5M Stock)</td>
<td>2 M</td>
<td>+ 20 ml</td>
</tr>
<tr>
<td>Nuclease Free H2O</td>
<td>d.d.</td>
<td>Fill to 50 ml</td>
</tr>
</tbody>
</table>

2. Bind / Wash Buffer (1x): 5 mM Tris pH 8, 0.5 mM EDTA, 1 M NaCl.
Preparation: Simply dilute 2x buffer above 1:1 with d.d.H2O.

3. Annealing Buffer (10x): 100 mM Tris pH 8, 10 mM EDTA, 500mM NaCl

(Table 2.4).

Table 2.4. Preparation of 10x Annealing Buffer: To prepare 1 ml of 10x annealing buffer, mix the following components to a homogenous solution:

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Final Concentration</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tris pH 8 (1M Stock)</td>
<td>100 mM</td>
<td>+ 100 ul</td>
</tr>
<tr>
<td>EDTA pH 8 (0.5M Stock)</td>
<td>10 mM</td>
<td>+ 20 ul</td>
</tr>
<tr>
<td>Sodium Chloride (5M Stock)</td>
<td>500 mM</td>
<td>+ 100 ul</td>
</tr>
<tr>
<td>Nuclease Free H2O</td>
<td>d.d.</td>
<td>+ 780 ul</td>
</tr>
</tbody>
</table>

The annealing reaction takes place in 10 mM Tris pH 7.5, 50 mM NaCl, and 1 mM EDTA.

4. Elution Buffer: 10 mM Tris pH 8.

Preparation: Dilute Tris pH 8.0, 1M, 100x with d.d.H2O.

Note: We do not calibrate the pH of the solutions after mixing the solutions.

Setup:

1. Set Heating Block to 70°C (must accept 1.5 ml tubes)
2. Set Thermocycler or another Heating Block to 68°C
3. Set vortex at medium speed setting (setting 5 in our Fisher Vortex Genie 2 #12-812)
4. Thaw Biotinylated Oligos (a 200 uMolar resuspension should stored at @ -20°C)

4a. For Sindbis Toto1101 (Oligo #4):

5’-Biotin-TEG-GGCTGTACATTCTTCATGTCGGGATAGTCGATTTGCAAT (HPLC Purified)

4b. For Zika Puerto Rico Strain:

5’-Biotin-TEG-GCCCTATGAGAGATCCACACCACAAGTCTTCCCTTTT

Note: Upon familiarity with the protocol, each of these can be set shortly prior to the step in which the instrument is required.

Key:

+ | Add Reagent

- | Remove Supernatant

* | Centrifuge or Shake Step

^ | Repeat from Indicated Step

> | Transfer to Indicated Tube

= | Incubation Step

E | Wash and Prepare Beads

E.01 | * Mix MyOne C1 Magnetic Beads (hand shake & gentle vortex)

E.02 | > 225 ul Magnetic Beads -> 1.5 ml eppendorf tube

E.03 | = 1 min @ magnetic stand incubation
E.04 l - Remove Supernatant

E.05 l + 225 ul 1x Bind/Wash Solution

E.06 l * Vortex until Emulsion

E.07 l = 1 min @ magnetic stand incubation

E.08 l - Remove Supernatant

E.09 l ^ Repeat wash once from step E.05 (with +225 ul 1x Bind/Wash Solution)

E.10 l + 65 ul 2x Bind/Wash Solution

F | Oligo Treatment and Hybridization

Note: Choose tube size depending on whether you are using a thermocycler or heating block set to 68°C. I have typically used 200 ul tubes and a thermocycler for this step, but I don’t think it would matter.

To 35 ul Viral RNA (~1.5 ug/ul), add:

F.01 l + 4 ul 10x Annealing Buffer

F.02 l + 1 ul of 200 uMolar Biotinylated Hybridization Oligonucleotide

Note: The total annealing reaction is 40 ul per sample, but multiple reactions of the same sample can be combined into the same annealing reaction and subsequently split into 40 ul aliquots after annealing. The annealing reaction is scalable in our hands, however we have not attempted to scale any of the steps involving magnetic beads, mostly because our magnetic stand only accepts 1.5 ml tubes. See below for sequence and ordering information.
F.03 l = 10 min @ 68°C incubation (can use thermocycler or heat block)

F.04 l > Place out at RT immediately (for annealing)

F.05 l = 10 min @ room temperature incubation

Note: For step F.04 I have always simply removed the tube from the thermocycler set to 68°C and placed it out on my bench for gradual calibration to room temperature. Dr. Inna Ricardo-Lax has used the thermocycler to shift down temperature and incubate at 25°C and that method also works well.

G | Magnetic Bead Depletion

G.01 l + 40 ul Annealing Reaction (from B) –to-> Magnetic Bead Resuspension (from A)

Immediately & thoroughly:

G.02 l * 10x pipette mix

G.03 l * 10 second vortex

G.04 l = 10 min @ room temperature incubation

G.05 l * 2 second vortex (just to resuspend bottom)

Immediately:

G.06 l = 1 min @ magnetic stand incubation

G.07 l - Remove Supernatant (Aqueous Phase should be around ~ 105 ul)

Proceed Immediately to Wash & Elution:
H | Wash 2 Times
H.01 | + 400 ul 1x Bind/Wash Solution
H.02 | * 30 second Vortex @ medium setting (Critical!)
H.03 | = 1 min @ magnetic stand incubation
H.04 | - Remove Supernatant
H.05 | ^ Repeat once from step H.01 (with +400 ul 1x Bind/Wash Solution)

I | Elution
I.01 | + 400 ul 10 mM TRIS (pH 8.0)
I.02 | = 10 min @ 70°C incubation

Immediately:
I.03 | = 30 seconds @ magnetic stand incubation
I.04 | > Transfer supernatant to CoStar column (0.22 um cellulose acetate)
I.05 | * 2 min RT Spin @ 12,000 rpm
I.06 | > transfer to 1.5 ml eppendorf tube
I.07 | > Place immediately on ice

J | DNase Treatment

Note: Set heating block to 37 C, and allow sufficient time to cool if using the same block as above.

To 400 ul Filtered Elution RNA, add:
J.01 | + 45 ul 10x DNase I Buffer
J.02 | 16 ul Ambion DNase Enzyme

J.03 | 30 min @ 37°C incubation

K | Ethanol Precipitation

To DNase Treated RNA Elution (~460 ul), add:

K.01 | 340 ul Nuclease Free H2O

K.02 | 2x Split: Transfer 400 ul -> new 1.5 ml Eppendorf Tube (Maybe Critical!)

To each tube of 400 ul Diluted RNA Elution, add:

K.03 | 50 ul NaOAC 3M

K.04 | 2 ul Thermo Fisher Glycogen, RNA Grade

Note: Ashley describes in her paper for CirSeq, "Critical! In our experience, glycogen from other vendors can result in lower yield of nucleic acids after ethanol precipitation." [2]. As such, I always used this brand and never tried anything else, and can't speak to how critical this specific brand is to effective oligo capture.

K.05 | 1000 ul Ethanol/Isopropanol 1:1

K.06 | 30 min @ room temperature incubation, mix well

K.07 | 1 hour @ dry ice incubation

K.08 | overnight @ -80°C incubation

K.09 | * 45 min 4°C spin @ max speed

K.10 | - Remove Supernatant with 1 ml pipette tip
K.11 | 700 ul ice cold 70% Ethanol
K.12 | 20 min 4°C spin @ max speed

Note: This particular method of ethanol precipitation, using a mix of ethanol
and isopropanol and both room temperature and low temperature incubations
was taught to me by Dr. Troels Scheel when performed my first ethanol
precipitation, and I have been using it since. I would guess that, if the reader
has a preferred ethanol precipitation method that they have successfully used
in the past, it should work fine as well. I would caution however, that certain
changes I would have hypothesized to be minor had tremendously negative
effects, as explained in the following text. Thus once something worked, we
generally stuck with it, especially if there wasn't a large benefit to switching. I
learned via such experiences not make changes to the protocol, however
minor, without testing the relative efficacy of the change to what worked
previously.

K.12 | Remove Supernatant with 1 ml pipette tip
K.13 | Remove Remaining Supernatant with 20 ul pipette tip until minimal
       liquid left
K.14 | Few seconds 4°C spin @ ~ max speed
K.15 | Remove Supernatant with 20 ul pipette tip
K.16 | Air dry pellet for ~ 3 mins
K.17 | Resuspend in desired volume of 10 mM TRIS (pH 8.0)

Note: Nuclease free water can be used in this step if it would be preferable to
your downstream application.

Note: In the case of a typical 2 step capture whereupon the first round has 3 reactions (explained below), each pellet is resuspended in 5.8 ul of 10 mM TRIS and combined to equal 35 ul, upon which the reaction is taken through all the above steps for a second round of capture. Alternatively, each pellet can be subsequently resuspended in the same 35 ul (after resuspension of one pellet the supernatant with dissolved RNA is transferred to the next pellet in the next tube and so on).

2.6. General Principles of Oligonucleotide Based Hybridization Capture.

The theory behind oligo capture is fairly straightforward. We order synthetic DNA oligonucleotides complementary to the RNA of interest (initially Sindbis viral RNA in our case) that have biotin covalently attached to one end. First we incubate the total RNA pool and DNA oligos together and allow them to anneal. Second, we add these DNA:RNA hybrids to magnetic streptavidin beads that strongly bind to the biotin, and will immobilize these hybrids. Afterwards the binding supernatant, which contains unbound RNA's, is removed. Preferably the RNA target is immobilized on the beads whereas the other RNA in the pool is removed in the supernatant. Third, we elute the RNA off of the beads into a solution that is concentrated and purified such that the RNA is amenable to downstream applications. During our testing phase, we
also performed a second elution in harsher conditions to test various elution protocols (Figure 2.4).

**Figure 2.4. Oligonucleotide Based Hybridization Capture Schematic.** The biotinylated DNA oligo (blue) and RNA target (red) is hybridized in the first step, followed by binding of the hybrid to magnetic streptavidin beads (black squiggle). Upon removal of the supernatant containing unbound RNA, the DNA:RNA hybrids are eluted in a mild condition and then in a harsh condition. Dr. Inna Ricardo-Lax designed a far more professional diagram (Figure 3.31), however as much of scientific development occurs on napkins I have included here my original napkin and Sharpie marker drawing of the intended protocol [3]. Whereas the protocol underwent a multitude of changes during the development phase, the general procedure except for the second elution step still applies to the final protocol.

Our plan was to first optimize the protocol using pure in vitro transcribed Sindbis Toto1101 strain RNA. Having 100% of the input RNA pool be the target of interest reduces variables during initial testing. CirSeq optimally requires 5 ug of RNA and minimally 1 ug of RNA for library preparation, and

174
thus we always tested 5 ug of input in vitro transcribed RNA unless otherwise
stated. Next we aimed to test the non-specific binding properties of cellular
RNA. Finally we would optimize the protocol using our intended sample of
infected intracellular RNA, where our target viral RNA is surrounded by an
excess of cellular RNA. The goal was to acquire 1-5 ug of Sindbis RNA that is
relatively pure for CirSeq library preparation. The first part of this section
details our efforts toward this goal.

Around the later stages of protocol development an epidemic of Zika virus
surprised the world. An increase in microcephaly in newborns in regions
affected by the epidemic and a possible link with Zika virus increased interest
in the RNA virus field [4]. Our lab was not immune from such interest, and Dr.
Mohsan Saeed and Dr. Yingpu Yu approached me to collaboratively adapt the
enrichment procedure to Zika virus infected cells as well. Their specific
downstream analyses would be aided by enriched intracellular Zika viral RNA.

Zika virus produces lower titers than Sindbis, and relatively pure Zika RNA
extracted from purified Zika virions does not yield enough RNA for their
downstream analyses. In cultured cells, Zika does not replicate as well as
Sindbis, and consequently is a smaller fraction of the total infected cell RNA
pool than Sindbis. Thus, Zika virus required further optimization on top of what
was done with Sindbis, and we learned some additional lessons during the
process, and the latter part of this section details our efforts therein.
We intended to analyze intracellular Sindbis genomic viral RNA in comparison with extracellular virion RNA, which is purely genomic (Chapter 3). The desire to develop such a protocol was borne when we ran out of other options. Both Sindbis plus strand species, genomic RNA and subgenomic mRNA, contain poly-A tails and cannot be differentiated using a poly-A based selection of intracellular RNA. Subgenomic RNA is present in a vast excess, and the additional pull down of cellular mRNA will result in only a small fraction of next-gen sequencing reads being of our intended target. We would be unable to differentiate between reads from genomic RNA and subgenomic mRNA in the 5' region, confounding our downstream analyses. Since they are both different sizes, we reasoned that we can run a denaturing agarose gel and gel extract the genomic RNA band using electroelution. Electroelution using Millipore (dialysis) D-tubes (Millipore Product # 71506) was highly efficient against pure in vitro transcribed genomic RNA and in a 1:1 mix of it with cellular RNA (Figure 2.5 A). However, we were not able to achieve success using real-world infected cellular RNA samples wherein the Sindbis genomic RNA is only 5-10% of the total RNA pool. We then thought that perhaps we can first ribodeplete the sample and then gel extract the genomic RNA band once the viral RNA is in a more preferable ratio within the total RNA pool. We next-generation sequenced a control sample of 5 ug of pure Sindbis viral genomic RNA from purified virions that had been subjected to ribodepletion using
EpiCentre’s RiboZero kit. Unfortunately we lost 73% of viral RNA during the procedure, and the kit seems to drastically reduce coverage in certain regions of the viral genomic RNA (Figure 2.5 B). Perhaps the viral RNA is being degraded, and probes are binding to select fragments and depleting them. Since the probes and reagents are proprietary and we do not know the specific sequences, we chose to abandon these efforts. The specific target enrichment kits that were used in the field were quite expensive and had to be ordered on blind faith that their computationally determined oligo pool would be effective with no prior data on our specific target. We decided to first try specific target enrichment on our own before such capital expenditure.
Figure 2.5. Electroelution and Ribodepletion Results. (A) Separation of Sindbis genomic RNA in a denaturing agarose gel and subsequent electroelution of the RNA using Milliopore D-tubes was effective against a 1:1 ratio of in vitro transcribed genomic RNA and Huh 7.5 cellular RNA. The goal was to ultimately isolate Sindbis genomic RNA (11,670 bp) from Sindbis subgenomic mRNA (4,103 bp), 28S rRNA (5,070 bp), 18S rRNA (1,869 bp), and tRNAs (73-93 bp), with most of the other RNA species assumed to be minor. We recovered 70% of input Sindbis genomic RNA, and the recovered pool was 99.5% composed of Sindbis RNA, with 0.5% cellular RNA. However, the procedure was not effective against intracellular RNA from infected Huh 7.5 cells. Error bars represent SEM of qPCR technical replicates. (B) A coverage plot of pure Sindbis genomic RNA from virions subjected to ribodepletion reveals uneven coverage across the genome, with several severe low coverage regions in the nsP2 and nsP3 regions. Aligned number of reads per nucleotide is on y-axis, and nucleotide position of Sindbis genome is on x-axis.
A) 1:1 Sindbis RNA + Cell RNA Electroelution

B)
The protocol is not difficult nor time intensive. Initially when Dr. Charles Rice (Charlie) asked me to develop such a protocol in-house, I had assumed that it would be near impossible, since it is rare in the literature and offered by corporations that have teams of experienced scientists for thousands of dollars. Additionally, I was an inexperienced graduate student who first worked with RNA less than a year prior, and yet here we are. Everyone in the lab that has been taught the protocol has performed it successfully, though that is not saying much as I am the least experienced with RNA viruses amongst them. Since then, other targets have been successfully enriched using the same procedure, for example Dr. Inna Ricardo Lax has purified Sindbis Toto1101 strain subgenomic RNA and Stephen Bluethgen has purified Coxsackievirus CVB3 Woodruff strain to 95+% purity. Equipped with an accurate target measurement qPCR assay, anyone should be able to do so in a few days’ time.

As such, I am convinced that this method is applicable to enrich any target RNA of the reader’s interest from any total RNA pool. Such enrichment will be useful for a variety of downstream applications in which RNA level information is characterized and such information would be lost in a reverse transcription step prior to PCR amplification, such as detecting low frequency RNA variants or RNA methylation events. In this section I include many details on all the parameters we tested, largely in chronological order. My goal in doing so is to
impair the reader with a sense of which aspects of the protocol were
extensively tested and should be changed with utmost caution, and which
aspects can likely be optimized further or changed without immediate failure.
Several parameters continue to not make much sense as to why they are
important. I have noticed that a lack of details lead users to assume that every
step is equally important or unimportant as described and troubleshooting or
further optimization is initially aimless, and often completely unimportant steps
continue to be performed as if they involve 'voodoo magic'. I had such an
experience when I was implementing CirSeq directly from the short methods
section of a paper, and additional details and input from Dr. Ashley Acevedo
(who developed the protocol), Dr. Chris Mason, and others on various
parameters was critical to successful library preparation.
2.7. Hybridization Oligonucleotide Probe Design for Sindbis Genomic Viral RNA.

For design of the sequence of the DNA probe that anneals to Sindbis viral RNA, we used the Oligo 7 software from Molecular Biology Insights Inc [5, 6]. I chose a 40 base pair length as a starting point for the oligos, as that is the maximum length beyond which the per-nucleotide pricing by IDT moves to the next tier (up to 40 bp is the lowest pricing tier in 2015). The software was used to computationally generate hybridization probe sequences 40 bp in length in the plus-strand full length Sindbis Toto1101 genome, free of duplexes and hairpins (Figure 2.6). Minimal RNA secondary structures should theoretically assist the RNA region in binding to the DNA probe instead of to itself. The remaining settings were largely set to the software's default.
Figure 2.6. Oligo 7 Software Predicted Secondary Structure Free Hybridization Probes. The (A) probe type, (B) length, and (C) structure settings used to query the Sindbis Toto1101 sequence and predict optimal hybridization oligonucleotide sequences. (D) The output window with predicted sequences shows 10 sequences in the genomic only RNA region, which should only pull down Sindbis genomic RNA, and 6 oligos in the genomic and subgenomic mRNA region that would pull down both RNA species.
A) **Probe Type:** Adjust in Search for Primers & Probes -> Search Option

B) **Oligo Length:** Adjust in Search Parameters -> Parameters -> Constraints
C) Oligo Structure: Adjust in Search for Primers & Probes -> Subsearches

Secondary Structure Restrictions

D) Oligo 7 Output: Predicted Secondary Structure Free Oligo Sequences

- genomic
- RNA only
- subgenomic
- mRNA + genomic

5' Genome Position  Oligo Length  + Strand Sequence (reverse complement is hybridization oligo)
The software predicted 10 sequences in the genomic-only region of the plus strand. We ignored the 6 sequences predicted in the genome and subgenomic region as we did not want to pull down subgenomic mRNA for our analysis (Figure 2.6 D). The sequences were aligned to the whole human and mosquito genomes (Aedes aegypti and Aedes albopictus) using NCBI BLAST to ensure that cellular RNA from human and mosquito cell infections will not also be targeted. The maximum identity for all 10 genomic-only sequences was 56% in both cell types, which did not seem like cause to worry (Table 2.5).
Table 2.5. Hybridization Oligos Aligned to Human and Mosquito Genome and Transcriptome. The 10 oligos from figure 2.6 D in the genomic RNA only region of Sindbis were aligned using NCBI BLAST to predict off-target binding. The identity percentages of alignment to human genome and transcriptome database is shown in 'Human BLAST' column. The maximum identity percentage between such alignment to Aedes aegypti and Aedes albopictus is shown in the 'Mosquito BLAST' column. The Oligo 7 software outputs plus strand sequence, and thus the reverse complement shown should be synthesized.
<table>
<thead>
<tr>
<th>Oligo #</th>
<th>SINV Position</th>
<th>SINV Region</th>
<th>Length</th>
<th>Score</th>
<th>Reverse Complement</th>
<th>Human BLAST</th>
<th>Mosquito BLAST</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1748</td>
<td>nsP2</td>
<td>40</td>
<td>898</td>
<td>GGCATTCTTCAG</td>
<td>45%</td>
<td>35%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CACAGAGTTTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GCGAGACAACG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ATATACT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2009</td>
<td>nsP2</td>
<td>40</td>
<td>903</td>
<td>TGTAACCTTGTA</td>
<td>50%</td>
<td>37%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CTGCTCCTCTTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TGTATTCTTGGC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GGGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2323</td>
<td>nsP2</td>
<td>38</td>
<td>867</td>
<td>CATACCCCTCAG</td>
<td>44%</td>
<td>36%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TCTTAGCACGTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GGCCTCAAATTTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2786</td>
<td>nsP2</td>
<td>40</td>
<td>870</td>
<td>GGCTGTcATAC</td>
<td>42%</td>
<td>35%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TTTAGTGCcCGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GATAGTCGATTT</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GCAAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5019</td>
<td>nsP3</td>
<td>37</td>
<td>849</td>
<td>ATGCCGGGAGTG</td>
<td>48%</td>
<td>37%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TGCCGATTAAAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AGGACTACTTTTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5061</td>
<td>nsP3</td>
<td>40</td>
<td>928</td>
<td>GGTAGGcGTcTT</td>
<td>45%</td>
<td>32%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CTGGCAcCTTCTA</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TGACTTTACcGGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CGGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>5932</td>
<td>nsP4</td>
<td>40</td>
<td>902</td>
<td>TTACGGAGACTG</td>
<td>45%</td>
<td>35%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GTACcCTACTTTT</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GTTGcGCcTTGGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TGGGcA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>6649</td>
<td>nsP4</td>
<td>40</td>
<td>899</td>
<td>TTTcGcGTcTCT</td>
<td>50%</td>
<td>37%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TCTGTGTGT TTTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GTGCcTGcGTGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AACcTT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>6829</td>
<td>nsP4</td>
<td>40</td>
<td>919</td>
<td>AGTACCcGGGC</td>
<td>55%</td>
<td>40%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GCCTTGcGTGAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GTGTcTGcGTcTAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GATTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>7565</td>
<td>Intergenic</td>
<td>41</td>
<td>882</td>
<td>CTGaCTATTTTAG</td>
<td>41%</td>
<td>56%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GACCcCGcTAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AGATGcTTTATT</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TcCcCCT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
It is important to remember that this software search using these parameters lists positive-strand sequences, and thus the reverse complement must be ordered as the DNA oligo for hybridization. The ribosomal RNA removal protocol by Ignolia et al states that "oligonucleotides should be modified by the addition of the 5′-biotin-TEG and purified by HPLC to eliminate unbiotinylated products that could compete with effective, biotinylated molecules during subtraction." [7]. TEG refers to tetraethylene glycol, which is used as a 15 atom spacer between the biotin and the DNA to minimize steric hindrance of biotin-avidin binding (Figure 2.7).

![Chemical Structure of a Biotin Tetraethylene Glycol Oligonucleotide](image)

**Figure 2.7. Chemical Structure of a Biotin Tetraethylene Glycol Oligonucleotide.** The DNA oligonucleotide sequences were ordered from Integrated DNA Technologies (IDT) with biotin linked to the DNA using the specific 15 atom TEG spacer shown [8]. The structure of such spacers from different DNA synthesis providers seem to differ slightly, and we have not tested alternatives.
We chose oligo #4 because it was the closest predicted oligo to the amplified site in our qPCR assay. This should minimize RNA degradation as a variable during the testing phase. Additionally for comparison we ordered a biotinylated version of our 20 bp qPCR oligo, since that complementary sequence binds selectively at least in a PCR setting. Both were ordered from IDT with a 5' Biotin TEG linker as follows and with the HPLC purification option.

Oligo #4:
/5BiotinTEG/GGCTGTCATTACTTCATGTCCGGATAGTCGATTTGCAAT (+HPLC Purification)

Oligo HC: /5BiotinTEG/GCTGGAACACCGGAAATCTA (+HPLC Purification)

2.8. Water and Low Heat Effectively Elutes Target RNA.

First we wanted to find an effective elution protocol, such that consequent experiments can give us insight on the binding steps and not be confounded by variability in elution. We attempted to stick to a mantra of adjusting only one variable at a time, and we initially used the Epicentre RiboZero kit which we had prior experience with (Figure 2.5 A), while using our Sindbis target oligos instead of the supplied rRNA-binding oligos. The reagents in the kit are proprietary and the use of a biotin-avidin bond was not explicitly mentioned, however given that most rRNA removal methods in the literature use it we predicted that the 'RiboZero rRNA Removal Solution' was likely a mix of
biotinylated oligos and the 'RiboZero Microspheres' were likely magnetic beads containing some variant of avidin [7].

A short description of the RiboZero protocol is described here as the stand-alone kit is no longer available for purchase as of 2019, and the online protocol website is defunct. Briefly, a maximum of 5 ug total RNA in 26 ul volume is added to 10 ul RiboZero rRNA removal solution and 4 ul RiboZero reaction buffer and incubated for 10 mins at 68°C and then 5 mins in room temperature. 225 ul of RiboZero microspheres is washed with 225 ul d.d.H2O and resuspended in 65 ul RiboZero resuspension buffer and 1 ul RiboGuard RNase Inhibitor. Both solutions are mixed together and incubated for 5 mins at room temperature, vortexed and incubated for 5 mins at 50°C, and placed in a magnetic stand. The supernatant with the rRNA depleted sample is then removed and ethanol precipitated for downstream applications.

For the annealing reaction we combined 5 ul of 200 uM of our biotinylated oligo #4 and oligo HC with 2.5 ug of in vitro transcribed Sindbis Toto1101 RNA in 31 ul ddH2O and 4 ul of the 10x RiboZero kit reaction buffer. As per kit instructions, the reaction was incubated for 10 mins at 68°C proceeded by 5 mins at room temperature, and the rest of the kit protocol was followed. The RiboZero kit performs a negative selection protocol, whereby the material stuck to the beads (rRNA) is discarded and the supernatant with depleted
rRNA is preserved. We were after the material stuck on the beads, and thus continued after the final step of the kit protocol by testing various methods of eluting and acquiring the nucleic acids from the RiboZero kit beads.

The Invitrogen website for Dynabeads states that "The biotin-streptavidin bond is broken by harsh conditions.. It has also been reported that the biotin-streptavidin interaction can be broken by a short incubation by non-ionic water at a temperature above 70°C [9]." Elsewhere Invitrogen's website provides conflicting information in a table of biotinylated DNA elution efficiency (Table 2.6):

**Table 2.6. Typical Elution Efficiencies of Biotin-Streptavidin Interaction in Different Elution Conditions.** Elution efficiency of water is in bold. The table is from a ThermoFisher Scientific website titled "Elution of the Streptavidin - Biotin Bond" that accompanies the Invitrogen Dynabeads page. The website warns that these efficiencies are for free biotin, and may differ for biotinylated ligands. Adapted from [10].

<table>
<thead>
<tr>
<th>Temperature &amp; Time</th>
<th>Elution Solution</th>
<th>Elution Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>90°C for 10 minutes</td>
<td>10 mM EDTA pH 8.2 and 95% formamide</td>
<td>96.8%</td>
</tr>
<tr>
<td>90°C for 5 minutes</td>
<td>10 mM EDTA pH 8.2 and 95% formamide</td>
<td>96.4%</td>
</tr>
<tr>
<td>90°C for 2 minutes</td>
<td>10 mM EDTA pH 8.2 and 95% formamide</td>
<td>96.8%</td>
</tr>
<tr>
<td>65°C for 5 minutes</td>
<td>10 mM EDTA pH 8.2 and 95% formamide</td>
<td>96.4%</td>
</tr>
<tr>
<td>Temperature</td>
<td>Solution</td>
<td>Nucleic Acid Recovery</td>
</tr>
<tr>
<td>-------------</td>
<td>----------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>65°C for 2 minutes</td>
<td>10 mM EDTA pH 8.2 and 95% formamide</td>
<td>97.9%</td>
</tr>
<tr>
<td>37°C for 10 minutes</td>
<td>10 mM EDTA pH 8.2 and 95% formamide</td>
<td>41.9%</td>
</tr>
<tr>
<td>90°C for 10 minutes</td>
<td>H2O</td>
<td>7.3%</td>
</tr>
<tr>
<td>90°C for 10 minutes</td>
<td>10 mM EDTA pH 8.2</td>
<td>52.0%</td>
</tr>
<tr>
<td>90°C for 10 minutes</td>
<td>95% formamide</td>
<td>35.9%</td>
</tr>
<tr>
<td>90°C for 10 minutes</td>
<td>30 mM NaOAc pH 9 and 95% formamide</td>
<td>95.5%</td>
</tr>
<tr>
<td>90°C for 10 minutes</td>
<td>80 mM NaOAc pH 9 and 95% formamide</td>
<td>97.3%</td>
</tr>
<tr>
<td>90°C for 10 minutes</td>
<td>140 mM NaOAc pH 9 and 95% formamide</td>
<td>95.4%</td>
</tr>
</tbody>
</table>

We tested the above statements by performing a first elution in d.d.H2O at around 70°C, and subsequently performing a second elution in harsher conditions to see the relative amounts retrieved by both conditions. The elutions were performed in a volume of 400 ul as that is the maximum volume that can be effectively ethanol precipitated within one eppendorf tube. In the following experiment, we first eluted at 75°C for 10 mins followed by 95°C for 15 mins, and ethanol precipitated the elutions and the supernatant into an equal small volume of 10 ul and characterized the total nucleic acid content via nanodrop and the viral RNA using qPCR (Figure 2.8). The qPCR was first run using undiluted inputs of the ethanol precipitations into the qPCR mastermix solution (Figure 2.8 B), which fell well within the standard curve, and then using 10x and 100x dilutions of the solution (Figure 2.8 C).
Figure 2.8. Majority of Target RNA is Eluted in Water and 75°C.
Comparison of binding and elution efficiency between predicted 40 bp Oligo #4 (purple) and 20 bp Oligo HC (orange, qPCR primer sequence). All samples were ethanol precipitated into an equal volume of 10 ul before quantitation. (A) Total nucleotide content measured via nanodrop spectrophotometer. (B & C) Sindbis viral RNA content measured via qPCR of (B) undiluted solutions and (C) diluted solutions. 1 ug of viral RNA was ethanol precipitated directly as a control (green bar in B), showing that the reagents themselves did not cause interference. Error bars represent SEM of technical replicates.
C)

qPCR (Diluted)

- **Supernatant**
- **Elution 1**
- **Elution 2**
- **Supernatant**
- **Elution 1**
- **Elution 2**

<table>
<thead>
<tr>
<th>Concentration (ng/ul)</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
</tr>
<tr>
<td>300</td>
</tr>
<tr>
<td>200</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

**Legend:**
- **Oligo HC**
- **Oligo #4**
Comparison of the undiluted and diluted qPCRs (Figure 2.8 B vs C) alerted us that something was being brought along during the ethanol precipitation that interfered with the qPCR reaction, even after the input is diluted 10x with the Roche Lightcycler Taqman qPCR mastermix solution for the qPCR run. This is included to alert the reader to keep their eyes peeled for such issues when optimizing an oligo capture protocol. It is a good idea to include multiple dilutions and watch out for such inconsistencies to avoid our initial confusion. This was our first indication that perhaps the ethanol precipitation itself was also being interfered with by these mystery component(s), which turned out to be true.

Since oligo #4 bound more in vitro transcribed Sindbis RNA than oligo HC, we discarded oligo HC and all further experiments were conducted with oligo #4. Perhaps the increased length of oligo #4, at around 40 bp, compared to oligo HC at around 20 bp accounts for this difference. Or perhaps there is less secondary structure in the binding region of oligo #4, and whereas oligo HC without linked biotin performs well for qPCR that utilized higher temperatures, the biotinylated form does not bind as effectively during the capture protocol.

Next, we tested an even harsher condition and tried 90°C for 10 mins in 100 ul of 95% formamide and 10 mM EDTA as our second elution, as recommended in Table 2.6. This solution was then diluted with d.d.H2O to 400 ul prior to
ethanol precipitation [11]. We also sought to use magnetic streptavidin beads that can be purchased at scale, and further reduce our reliance on propriety kit reagents. We ran the same experiment as above, this time replacing the kit's 225 ul of 'RiboZero Microspheres' with 50 ul of Invitrogen MyOne C1 Dynabeads brand magnetic streptavidin beads [7]. We also replaced the 'RiboZero Magnetic Bead Resuspension Solution' with 10 mM TRIS-HCl pH 7.5, 1 mM EDTA, and 2M NaCl, as recommended in Invitrogen's Dynabeads accompanying manual [12, 13].

Invitrogen advertises that ~500 pmol of ssDNA oligos are bound per mg of MyOne C1 Dynabeads. Thus for each ul of our 200 uM oligo resuspension, which has 200 pmol of ssDNA oligos, we would 0.4 mg of beads. Beads are provided in a slurry as 10 mg/ml, and thus we'd need 0.04 ml (40 ul) of beads. We later learned that, at least for ssDNA:viral RNA hybrids in our conditions, this is wildly inaccurate. These experiments used 1000 pmol of our oligo, but in a misguided attempt to save beads to preserve the bottle that I had for as many experiments as possible, I figured I'd try 50 ul of beads. We also suspected that in spite of the magnetic incubation steps, there is always a small amount of beads in the elutions. We incorporated a 0.22uM cellulose acetate purification step before ethanol precipitation (step I.04), and the filter is always stained brown from this step (the same color as these beads), confirming our suspicions.
Perhaps because I violated the mantra of changing and testing one variable at a time in comparison to what previously worked, the result was a tremendous failure, where we retrieved ~3% of viral RNA in the elution and could not account for ~80% of input RNA (Table 2.7). Given that oligo HC is one of our qPCR oligos and oligo #4’s binding site is nearby, it seems unlikely that RNA degradation accounts for this difference. However, running the results on a agarose gel confirmed that the elutions contained large amounts of DNA oligos at around 40 bp, and thus the probes were also being eluted in our conditions. These would need to be removed for the RNA in our elution to be effective in certain downstream application (Figure 2.9).

Table 2.7. Harsh Elution Comparison Reveals High Efficiency of Mild Conditions. Sindbis virus qPCR of the input and output solutions from 50 ul Dynabeads oligo capture experiment. A second elution using 95% formamide in 90°C shows that the majority of RNA is already eluted in the first elution condition of water and 75°C. The vast majority of input RNA is unaccounted for in the three output solutions, and is seemingly lost during ethanol precipitation.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Nanograms Viral RNA</th>
<th>Percent of Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>7,863</td>
<td>100%</td>
</tr>
<tr>
<td>Supernatant</td>
<td>1,534</td>
<td>19.5%</td>
</tr>
<tr>
<td>First Elution (Mild)</td>
<td>207</td>
<td>2.6%</td>
</tr>
<tr>
<td>Second Elution (Harsh)</td>
<td>0.8</td>
<td>0.01%</td>
</tr>
<tr>
<td>Unaccounted</td>
<td>6,121</td>
<td>77.9%</td>
</tr>
</tbody>
</table>
Figure 2.9. DNA Oligos Are Eluted by Mild and Harsh Conditions. Agarose gel of the oligo capture experiment in table 2.7, where oligo capture solutions (lanes 5-7) were ethanol precipitated into 10 ul and run. Lane 1: NEB 1 kb DNA ladder, where the lowest band is 500 bp, lanes 2 & 3: 400 ng and 40 ng respectively of Sindbis *in vitro* transcribed RNA, lane 4: ~400 ng aliquot of the Sindbis *in vitro* transcribed RNA input in this experiment, lane 5: binding supernatant from oligo capture experiment, lanes 6 & 7: mild elution and harsh elution, respectively, from oligo capture experiment. The majority of Sindbis RNA is in the supernatant (lane 5) and not being bound by the beads. The bright bands in lanes 6 and 7 indicate that the oligo is also being eluted in both conditions.
Whereas the total viral RNA bound in our early experiments were relatively low and most remained in the supernatant, we were surprised to see that most viral RNA was eluted in our first mild condition, and a minor amount remained to be eluted in the harsher condition of 90°C for 10 minutes, in either d.d.H2O or 95% formamide in 10 mM EDTA. The same was true in later experiments on Zika virus. Since its discovery in chicken and duck eggs, the biotin (vitamin H or B7) and avidin bond has been known to be one of the strongest non-covalent interactions in biology [14]. How is it that a femtomolar affinity bond can be disrupted by just water and low heat? (Note: low heat in molecular biology terms, a 70°C shower will still scald). In discussions with Dr. Jeremy Dittman, we postulated a few possibilities. Perhaps the DNA oligo + viral RNA hybrid is dissociating instead. The incongruence between our nanodrop and qPCR data (Figure 2.8 A vs C) shows that there is far more total nucleic acid in both elutions than there is viral RNA, indicating that a significant proportion of the DNA oligo is in the elutions. The agarose gel results in Figure 2.9 adds further evidence that the DNA oligo is also being eluted. It is possible however, that the eluted fraction consists of DNA oligos that nonspecifically bound with just enough affinity to prevent being pulled in the supernatant, but did not form a biotin-streptavidin high affinity bond, and this nonspecific binding got disrupted during elution.

Another possibility is that the TEG linker between the biotin and the DNA oligo
probe is being cleaved (Figure 2.7), releasing the DNA-RNA hybrid into the elution whereas the biotin molecule itself stays bound to the magnetic bead that is coupled with some form of avidin. Or perhaps the biotin-avidin bond is truly being dissociated by these conditions, as temperature has been shown have a negative correlation with binding affinity and perhaps 70°C in a deionized medium is sufficient [15]. We cannot distinguish between these possibilities as we never tested directly for biotin content, and it was outside the scope of our aim, though it would be interesting to do so.

2.9. Efficacy of Direct vs Indirect Binding and Identification of Interference During Ethanol Precipitation.

Next we reverted to using a 225 ul volume of MyOne C1 Dynabeads magnetic beads as in the RiboZero protocol and tested the non-inferiority of homemade annealing buffer for the biotinylated oligo to viral RNA hybridization reaction. Our successful trial used a final annealing buffer concentration of 10 mM TRIS, 1 mM EDTA and 50 mM NaCl, as detailed in Millipore Sigma's 'Protocol for Annealing Oligonucleotides' Webpage [16]. As we had our suspicions about ethanol precipitation interference, we assayed the solutions before and after ethanol precipitation via qPCR (Figure 2.10 A, B). We also ran 80% of the ethanol precipitated solutions on an agarose gel (Figure 2.10 C). Prior to ethanol precipitation, the solutions are too dilute to visualize bands on a gel.
Figure 2.10. Non-Inferiority of Prepared Annealing Buffer for Oligo Capture. Sindbis virus qPCR of pre (A) and post (B) ethanol precipitations solutions from an oligo capture experiment comparing *in vitro* transcribed viral RNA results from homemade annealing buffer (orange) and RiboZero kit annealing buffer (purple). Comparison of first and second elutions further confirms that most viral RNA is eluted in d.d.H2O and 70°C. Comparison of elution 1 before and after ethanol precipitation shows loss of RNA. C) Agarose gel of the ethanol precipitated solutions quantified in (B) show that the viral RNA band is not present in the first elution, whereas a DNA oligo band is. There is no difference between both annealing buffers.
A) Pre Ethanol Precipitation

![Bar chart showing total nanograms of viral RNA for different samples before ethanol precipitation.]

B) Post Ethanol Precipitation

![Bar chart showing total nanograms of viral RNA for different samples after ethanol precipitation.]

- Orange bars represent Homemade Annealing Buffer.
- Purple bars represent RiboZero Annealing Buffer.
Relative to the experiments in the prior section, in the above experiment we further reduced the temperature in the first elution to 70°C in d.d.H2O, and continued to retrieve a majority proportion of RNA in the first elution. Consequently we decided that the harsh elution condition was unnecessary for further experiments. Additionally, to keep the environment of the beads consistent, the use of d.d.H2O for washing the beads as in the RiboZero protocol was replaced with the binding solution recommended in the invitrogen Dynabeads manual diluted to 1x with d.d.H2O to 5 mM TRIS, 0.5 mM EDTA and 1M NaCl (this may not matter but it made me feel better). The recommendation of using Tris pH 7.5 was changed to Tris pH 8 in the bind and wash solutions because I had it available on my bench and the labmate who had a stock of Tris pH 7.5 was away from his bench at the time I made new buffers. These changes demonstrated non-inferiority in interim experiments and remain in the final protocol.

These data collectively suggest that the viral RNA in the elution is either not being precipitated into the pellet or is being degraded during ethanol precipitation. The majority of viral RNA that is unaccounted for in table 2.7 is likely being lost during this step. It is critical for downstream applications to have a working procedure that concentrates the RNA into a smaller volume and into either pure water or a buffer of choice, free of the contaminants leached from the beads. The ethanol precipitation control in Figure 2.8 B and
multiple successful ethanol precipitations of RNA in experiments outside of this protocol indicate that the reagents were not the problem, nonetheless they were freshly made several times and the problem persisted. I had not encountered such a problem with ethanol precipitation prior to these experiments.

Next we tested whether direct binding may increase efficiency. Indirect binding refers to first coupling the biotinylated DNA probes and the RNA target, and then coupling this hybrid with the streptavidin beads, as in the Epicentre RiboZero protocol and the above experiments. Direct binding refers to first coupling the biotinylated DNA probes with the streptavidin beads, and then coupling this mixture with the RNA target [17]. Direct binding is used in protocols accompanying New England Biolabs' and Qiagen's magnetic bead offerings, and may have the advantage that excess biotinylated DNA probes not bound to the beads will be washed away prior to RNA binding and should not be present in the elution. This may solve the eluted DNA oligo problem highlighted in Figure 2.9 and Figure 2.10 C. Essentially the same protocol as above was followed with one difference; 10 ul of 200 uM Oligo #4 was added to the beads in 2x bind/wash solution instead of to the RNA in annealing buffer, and incubated at room temperature for 15 minutes before the addition of the RNA. For comparison we conducted an indirect binding experiment as before also using 10 ul of 200 uM Oligo #4, as it was suggested to me to use
as much oligo as possible to drive the hybridization reaction forward. 5 ug of Sindbis *in vitro* transcribed RNA was used as input in this experiment (1.31 picomoles), which corresponds to a DNA hybridization oligo to Sindbis viral RNA molar ratio of 1500:1. As a control for this suggestion I included a very low amount of oligo for indirect binding, 0.1 ul of Oligo #4, representing an oligo:RNA molar ratio of 15x. These solutions before any further purification steps were highly diluted and assayed via qPCR (**Figure 2.11**).

**Figure 2.11. Comparison of Direct vs Indirect Biotinylated DNA : Streptavidin : RNA Binding Protocols.** Sindbis virus qPCR of highly diluted pre-purification solutions from an oligo capture experiment comparing viral RNA results between indirect binding at 1500x and 15x DNA:RNA molar ratios, and direct binding at 1500x DNA:RNA molar ratio. 5 ug of *in vitro*
transcribed Sindbis RNA was input into the capture experiment. The majority of viral RNA remained in the supernatant in both high molar ratio conditions. However, more viral RNA was retrieved in the elution than left in the supernatant in the low molar ratio control sample. Error bars represent SEM of technical replicates.

Direct binding had very poor efficiency using our protocol, but performed marginally better than indirect binding with the high amount of oligo. To our surprise, the low oligo control was the most efficient capture to date. We then sought to solve the ethanol precipitation problem by purifying the solutions in the above experiment with Qiagen RNeasy columns. We characterized the amount of viral RNA retained relative to the 5 ug of input viral RNA before and after column purification via qPCR (Table 2.8).

**Table 2.8. Loss of Viral RNA in Oligo Capture Solutions During Column-Based RNA Purification.** Sindbis virus qPCR of low dilution samples from the experiment in figure 2.11 before and after column purification and resuspension in d.d.H2O. For all columns, percentage of viral RNA is calculated relative to the input 5 ug of *in vitro* transcribed RNA. Unfortunately, it seems I had forgotten the lesson I learned in figure 2.8 and used low dilutions for the qPCR, which is likely why the pre-column values are different from those in figure 2.11 even though both were assaying the exact same samples. Calculations in this table are of the total amount from assaying 10x and 100x dilutions, and in figure 2.11 are from assaying 1,000x and 10,000x dilutions, which gave far more consistent and likely accurate values even though all dilutions fell within the standard curve.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Solution</th>
<th>Pre-Column</th>
<th>Post-Column</th>
<th>Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indirect 1500x Ratio</td>
<td>Supernatant</td>
<td>49.5 %</td>
<td>7.5 %</td>
<td>42 %</td>
</tr>
<tr>
<td></td>
<td>Elution</td>
<td>9.0 %</td>
<td>0.1 %</td>
<td>9 %</td>
</tr>
<tr>
<td></td>
<td>Supernatant</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>Indirect 15x Ratio</td>
<td>16.8 %</td>
<td>3.1 %</td>
<td>14 %</td>
<td></td>
</tr>
<tr>
<td>Elution</td>
<td>60.6 %</td>
<td>8.1 %</td>
<td>53 %</td>
<td></td>
</tr>
<tr>
<td>Direct 1500x Ratio</td>
<td>67.7 %</td>
<td>9.2 %</td>
<td>59 %</td>
<td></td>
</tr>
<tr>
<td>Elution</td>
<td>22.2 %</td>
<td>0.5 %</td>
<td>22 %</td>
<td></td>
</tr>
</tbody>
</table>

Most of the viral RNA in both the binding supernatants and elutions is lost during column purification. Given that the post-column purification samples are substantially lower than the suppressed pre-column purification values in the low dilution qPCR assay (Table 2.8), either (i) the mystery component affecting the qPCR as in Figure 2.8 is not being removed by the silica-based columns and the quantified difference (loss column) is accurate, or (ii) there is an even more substantial loss of viral RNA during column purification than indicated in the above table if the pre-purification values are suppressed but the post-purification values are accurate.

We were able to visualize viral RNA in the elution by running 66% of each column-purified solution on an agarose gel (Figure 2.12 A). This was the first time we retrieved more RNA in the purified elution than we lost in the supernatant (Figure 2.12 B). This was confirmed whereby the elution in lane 5 had a thicker band running at the same size as the input Sindbis viral RNA than the binding supernatant in lane 4 (Figure 2.12 A).
Figure 2.12. Low DNA:RNA Molar Ratio Condition Yields Higher RNA Retrieval Than Loss. A) Agarose gel of column purified RNA samples in table 2.8. Lane 1: 3 ug aliquot of input in vitro transcribed Sindbis RNA, lanes 2-3: Indirect protocol 1500x molar ratio binding supernatant and elution, lanes 4-5: Indirect protocol 15x molar ratio binding supernatant and elution, lanes 6-7: Direct protocol 1500x molar ratio binding supernatant and elution. Viral RNA band in lanes 2-7 being far dimmer than in lane 1 confirms that the majority of RNA is lost during column purification. According to pre-column qPCR values in figure 2.11, lanes 2, 5 and 7 should be equivalent to lane 1 if the loss was
only an artifact of qPCR. Having a low molar ratio results in an invisible DNA oligo band in the elution (lane 5 vs lanes 3 and 7). The indirect low molar ratio condition yields more RNA in the elution than is lost in the supernatant (lane 5 vs 4), and vice versa in the other conditions. B) Lab notebook entry from April 2016 celebrating higher viral RNA retrieval than loss.

Regardless, column-based purification in this context is not the answer to the ethanol precipitation problem. The reader will notice throughout this chapter that many aspects of the protocol were chosen due to convenience, and many aspects of the protocol may be changed with minimal detriment. However, the above data was included to alert the reader that the specific points at which ethanol precipitation or columns are used in the final protocol were decided carefully upon many failures and problems, and should be changed with utmost caution. I had previously assumed that they were largely interchangeable (as they likely are in many other downstream applications), which underpinned many of my numerous failed enrichment trials.

2.10. Solving Interference During Ethanol Precipitation.

Next we sought to both test DNA oligo : viral RNA molar ratios in a one log range around 1:15, which previously gave great results, and also find the cause of the problems that occur during ethanol precipitation. We tested molar ratios of 1.5x, 15x, and 150x against an input of 5 ug of in vitro transcribed Sindbis viral RNA, and characterized the solutions before further purification using qPCR (Figure 2.13). We included a control in which the input RNA was
taken through the protocol without the addition of the biotinylated oligos, primarily to see if the oligos interfered with ethanol precipitation. The protocol used was the same as previously until the elution step. Charlie stated that "RNA does weird things in distilled water", and thus we buffered the elution solution and changed it from 400 ul of d.d.H2O to 400 ul of 10 mM TRIS.

![Pre Ethanol Precipitation](Image)

**Figure 2.13. Influence of DNA:RNA Molar Ratio on RNA Retrieval in Elution.** Sindbis virus qPCR of highly diluted pre-purification solutions from an indirect oligo capture experiment comparing oligos in a one-log molar ratio range to what was previously effective in Figure 2.11, along with a control in which no oligos were added. 5 ug of *in vitro* transcribed Sindbis RNA was input into the capture experiment. A DNA:RNA molar ratio of 150x had the highest efficiency. Error bars represent SEM of technical replicates.
A molar ratio of 150x in 225 ul of MyOne C1 Dynabeads gave an excellent return of viral RNA! 85% of the input of 5 ul viral RNA was in the elution. Using these elutions, we tested various hypotheses as to why we were losing 80% of the RNA in the elution upon ethanol precipitation. We tested 4 hypotheses; (i) I am bad at ethanol precipitation, (ii) the binding buffer is causing problems, (iii) the bound oligos is causing problems, and (iv) the cellulose acetate column purification step is preventing the RNA/DNA hybrid from passing through.

Upon further discussions, we postulated that the first hypothesis is unlikely, due to the hundreds of ethanol precipitations during my CirSeq sample preps working well. If the second hypothesis is correct, then RNA in bead binding buffer will have a major loss of recovery upon ethanol precipitation. To test this hypothesis, we ethanol precipitated the binding supernatant of the no oligo control, which we presumed would contain the majority of the input ~5 ug of Sindbis RNA due to the absence of the biotinylated probe. The Sindbis RNA would be in 1x bead binding buffer (40 ul of 1x annealing buffer + 65 ul 2x bind wash buffer) and would contain any unknown component of the beads that is being leached during the biotin-streptavidin binding reaction. For all these tests that were ethanol precipitated in parallel, an overnight incubation step at -80 was added (step K.08) as I had a labmate's graduation party to attend. Previously I would mostly do a 30-60 min incubation in dry ice, or rarely an overnight incubation in -80°C instead if I preferred to continue the experiment the next day, but never both. We quantified the amount of RNA before and
after ethanol precipitation via qPCR of high dilutions such that there is only a trace amount of binding buffer remaining (Figure 2.14).

Figure 2.14. Influence of Binding Buffer on RNA Ethanol Precipitation.

Sindbis virus qPCR of the bead binding supernatant of the no oligo control (Figure 2.13) that contains ~5 ug of *in vitro* transcribed viral RNA, before and after ethanol precipitation. This sample was not filtered in the 0.22 um cellulose acetate column. The 1x binding buffer does reduce ethanol precipitation efficiency, but only by 25%. Thus trace amounts of binding buffer in the elutions does not explain the 80% loss of RNA during ethanol precipitation in previous experiments. Error bars represent SEM of technical replicates.

The bead binding buffer does reduce our quantified yield of viral RNA, either by inhibiting the ethanol precipitation itself or by the trace amounts inhibiting the qPCR, but only by 25%. Thus the trace amounts of bead buffer remaining during elution is unlikely to contribute to the 80% loss of RNA in our previous experiments.
experiments. Furthermore, qPCR quantification of viral RNA in the bead binding supernatant in our experiments is likely fairly accurate, especially in high dilutions where the remaining binding buffer is minute. Next we tested the third hypothesis, that the biotinylated oligos bound to the beads that are eluted are interfering with ethanol precipitation. Either the 40 bp of DNA or the bound biotin or TEG linker is increasing the solubility of the whole DNA:RNA hybrid molecule, making precipitation more difficult. If this was the case, the no oligo control precipitated in the above figure will have a far higher efficiency than both the 150x and 15x ratio samples, which would have equivalent efficiencies. Alternatively, the free amount of biotinylated oligos could be the interfering factor, in which case the 150x ratio sample would have a lower efficiency than the 15x ratio sample, and the no oligo control in the above figure would have the highest efficiency. Thus we ethanol precipitated the 150x and 15x elutions and quantified via qPCR (Figure 2.15).
Figure 2.15. Influence of Biotinylated DNA Oligonucleotides on RNA Ethanol Precipitation. Sindbis virus qPCR of the elution solutions of the 15x and 150x DNA:RNA molar ratio experiments in Figure 2.13, before and after ethanol precipitation. These samples was not filtered in the 0.22 um cellulose acetate column. The ethanol precipitation efficiency is not significantly affected by varying amounts of biotinylated probes input. Thus the varying amounts of DNA oligos used is unlikely to explain the 80% loss of RNA during ethanol precipitation in our previous experiments. Error bars represent SEM of technical replicates.

Neither was the case; the 150x and 15x molar ratio elutions had equal ethanol precipitation efficiencies, indicating that free biotinylated oligos is not the interfering factor. And both of these had a higher efficiency than the no oligo control in binding buffer, indicating that the fact that the viral RNA is hybridized to a biotinylated DNA oligo is not the interfering factor either (Table 2.9).
Table 2.9: Ethanol Precipitation Efficiencies of Unfiltered RNA Solutions. Compiled Sindbis viral RNA retention after ethanol precipitation of the experiments in figs 2.14 and 2.15. The percentage column reflects the amount of post-ethanol precipitation viral RNA divided by the amount of pre-ethanol precipitation viral RNA, detected by qPCR. These solutions were not filtered in the 0.22 um cellulose acetate column. In contrast to all previous experiments, we were somehow retaining the majority of RNA upon ethanol precipitation, and thus neither trace amounts of bead binding buffer in the elutions or varying amounts of DNA oligos in the elutions is the likely cause of our previous failures.

<table>
<thead>
<tr>
<th>DNA:RNA Molar Ratio</th>
<th>Buffer</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>150x</td>
<td>Elution Buffer</td>
<td>88 %</td>
</tr>
<tr>
<td>15x</td>
<td>Elution Buffer</td>
<td>89 %</td>
</tr>
<tr>
<td>0x (No Oligos)</td>
<td>1x Bead Binding Buffer</td>
<td>75 %</td>
</tr>
</tbody>
</table>

The high ethanol precipitation efficiencies above largely invalidate hypothesis one as well. Surely then, hypothesis four, that the CoStar 0.22 um cellulose acetate columns is somehow retaining the DNA:RNA hybrids, must be correct? To test this hypothesis, immediately after elution we had diluted the 400 ul solution to 800 ul with 10 mM TRIS (step K.02), and split the solution twofold. The supernatant of the no oligo control was similarly diluted twofold in 1x bead binding buffer and split. One half was ethanol precipitated without the CoStar column filtration step and quantified above (figs 2.14 and 2.15 and table 2.9). The high ethanol precipitation efficiencies therein indicate that the small amount of streptavidin beads that are usually filtered by the CoStar columns are not the interfering factor. We filtered the other half using the 0.22
um cellulose acetate columns, removing the remaining streptavidin beads, and then ethanol precipitated in parallel to the unfiltered solutions. The filtered solutions were quantified via qPCR and compared (Figure 2.16).

**0.22 μm Filter Column Interference Test**

<table>
<thead>
<tr>
<th></th>
<th>Unfiltered</th>
<th>Filtered</th>
</tr>
</thead>
<tbody>
<tr>
<td>150x Elution</td>
<td>5000</td>
<td>4500</td>
</tr>
<tr>
<td>15x Elution</td>
<td>4000</td>
<td>3500</td>
</tr>
<tr>
<td>0x Supernatant</td>
<td>3000</td>
<td>2500</td>
</tr>
</tbody>
</table>

Molar Ratio [DNA Oligo : Viral RNA]

**Figure 2.16. Influence of 0.22 um Cellulose Acetate Filter on RNA Ethanol Precipitation.** Sindbis virus qPCR of the ethanol precipitated solutions of the solutions described in table 2.9 with and without a CoStar column filtration step. The solutions were each diluted to double volume immediately upon acquisition and one half was subjected to a filtration step, and then both halves were taken through the rest of the ethanol precipitation protocol in parallel. Filtration through 0.22 um pores is effective in removing trace amounts of streptavidin beads and does not affect ethanol precipitation efficiency, and does not explain the 80% loss of RNA during ethanol precipitation in our previous experiments. Error bars represent SEM of technical replicates.
Cellulose acetate column filtration had no effect on ethanol precipitation efficiencies. None of our four postulated hypotheses as to why we were losing the majority of viral RNA during ethanol precipitation in previous experiments were correct, and ethanol precipitation suddenly started working well. Compared to previous experiments we made three changes during this experiment; (i) the elution solution was changed to 10 mM TRIS instead of d.d.H2O (upon Charlie's suggestion to buffer the solution), (ii) the elutions were diluted twofold prior to ethanol precipitation (to test CoStar column filtered vs unfiltered conditions), and (iii) the precipitation was incubated overnight in -80°C (I had a graduation party to attend). In delight that ethanol precipitation finally worked properly and in a desire to move on to the next phase, all three changes are retained in the final protocol (steps I.01, K.02 and K.08). Several lab mates have since skipped the overnight -80°C step and reported success with the protocol, but the relative efficiency has not been quantified. In all likelihood, the three incubation steps in room temperature, dry ice and -80°C (steps K.06, K.07 and K.08) contain some level of redundancy that is yet uncharacterized.
2.11. Importance of DNA:RNA Molar Ratio and Cost Comparison.

Compiling the results from the above experiments leads to an interesting observation; retrieval efficiency of 5 ug of input in vitro transcribed Sindbis RNA peaks at 150x molar ratio of biotinylated DNA oligos with over 85% recovery, and falls with both increasing and decreasing DNA oligo amounts (Figure 2.17). Moreover, there is a small amount of viral RNA recovered even without any DNA oligos added, indicating that the beads nonspecifically bind a proportion of nucleic acids that do not have biotin. Using very high amounts of DNA oligos performs no better than not adding any oligos at all!
Figure 2.17. DNA Oligo : Viral RNA Molar Ratio Determines Capture Success of 5 ug *in vitro* Transcribed Sindbis RNA. Compiled oligo capture efficiency values from previous experiments that varied DNA oligo amount while holding RNA input and amount of streptavidin beads consistent. As several experiments were conducted before we solved the ethanol precipitation problem, all values are Sindbis virus qPCR of pre-purification solutions in high dilution. There is a clear peak of efficiency in our tests at 150x molar ratio, below and above which the efficiency decreases. The beads also experience a small amount of non-specific binding as shown in the control in which no biotinylated probes were added (0x). Error bars represent SEM of technical replicates.
Oligo Capture Efficiency

Molar Ratio [DNA Oligo : Viral RNA]

- **Supernatant [Bad!]**
- **Elution [Good!]**

Total Nanograms Viral RNA

Fraction of input 5 ug Sindbis RNA
Since we optimized capture to a consistent input of 5 ug of in vitro transcribed Sindbis RNA, the 150x DNA:RNA molar ratio equates to 200 picomoles of DNA oligo. We hypothesize that at low amounts of oligo (2 and 20 pmol of oligo, corresponding to 1.5x and 15x molar ratios), the vast majority of viral RNA is not hybridizing with the biotinylated DNA probe. Consequently much of the viral RNA remains in the supernatant as it has no mechanism to bind to streptavidin outside of a small proportion binding nonspecifically. Additionally, we hypothesize that at high amounts of oligo (1 and 2 nmol of oligo, corresponding to 750x and 1500x molar ratios), the streptavidin in the beads is being saturated predominantly with biotinylated DNA oligos that are not hybridized with RNA, leaving most of the of the DNA:RNA hybrids in the supernatant. Even though all the RNA is likely hybridized to biotinylated DNA oligos, they are vastly outnumbered by unbound probes, and the hybrids likely have a lower binding affinity to the beads due to additional steric hindrance. 200 pmol of DNA oligo hybridized to 5 ug of viral RNA and bound to 2.25 mg of MyOne C1 streptavidin beads represents a favorable balance between these factors. We have not tested whether only the amount of DNA oligo, or the DNA oligo to RNA target ratio, is the more important parameter in our most efficient condition. This can be assessed by alternatively varying the RNA input amount while holding the DNA oligos consistent, and comparing the results to the above data. If the binding capacity of the beads to the RNA:DNA hybrid beyond 200 pmol of biotinylated DNA oligo is the limiting factor as in
our hypothesis, then that amount of oligo would be equally efficient at capturing higher or lower amounts of viral RNA, and would scale with varying the amount of beads.

Importantly, requiring the use of just 1 ul of 200 uM DNA oligo resuspension per reaction for an optimal 150x molar ratio given 5 ug of input target RNA allows the method to be extremely cost efficient, and consequently highly scalable. The majority of target enrichment described in genomics literature at this time is conducted using the SureSelectXT baits and kit provided by Agilent (Figure 2.18).

Figure 2.18. Agilent SureSelectXT Small Sample Full Kit (2016). This kit is the most commonly used method of enriching specific viral RNA targets in the literature currently, with examples in the text. As explained below, after 2016 the prices for full kits was no longer advertised on the website and a quote must be requested. The quote request process was more effort than I was willing to put in to update this figure, though I doubt it is much lower than listed here.
The above figure shows Agilent's prices as of 2016, as prices since are not listed publicly and a quote must be requested. The SureSelectXT baits allow for a multitude of targets. However, this is unnecessary for a single viral RNA target, or any target RNA molecule of the reader's interest. SureSelectXT is used for single virus targets, perhaps due to the lack of other options [18–21]. Ordering a 40 bp DNA oligo from IDT, biotinylated with a TEG linker and HPLC purified, as in our Sindbis viral RNA probe, allows an efficient capture reaction to be conducted with just 60 cents of baits (Table 2.10).

**Table 2.10. Cost Comparison of Baits for Our Protocol Versus Common Kit.** Per reaction (rxn) cost comparison of ordering baits for a commonly used kit and ordering a specific effective bait in our method, using 2016 prices. The number of reactions corresponds to a 150x molar ratio biotinylated DNA oligo to 5 ug of viral RNA in our method. A large percentage of our cost involves HPLC purification by the vendor to remove unbound biotin, and we have not performed a comparison with non-HPLC purified oligos on in-house HPLC purification.

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Probe</th>
<th>Quantity</th>
<th>Price</th>
<th># of Rxn's</th>
<th>Price / Rxn</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDT</td>
<td>40 bp Biotin-TEG DNA + HPLC</td>
<td>100 nmole</td>
<td>$125</td>
<td>40</td>
<td>$3.13</td>
</tr>
<tr>
<td>IDT</td>
<td>40 bp Biotin-TEG DNA + HPLC</td>
<td>250 nmole</td>
<td>$170</td>
<td>100</td>
<td>$1.70</td>
</tr>
<tr>
<td>IDT</td>
<td>40 bp Biotin-TEG DNA + HPLC</td>
<td>1 umole</td>
<td>$240</td>
<td>400</td>
<td>$0.60</td>
</tr>
<tr>
<td>Agilent</td>
<td>SureSelectXT custom library mix</td>
<td>Unknown</td>
<td>$5,850</td>
<td>16</td>
<td>$365.63</td>
</tr>
</tbody>
</table>
Perhaps it is unfair that our reaction prices at volume is compared to a 16 reaction set. In 2019 the only price listed on Agilent's website is for a similar high volume set. A volume of baits for 480 reactions is listed for the price of $149,039, or $310.50 per reaction, representing a 15% volume discount from their 2016 price for 16 reactions (Figure 2.19).

**SureSelect Custom Probes**

Enter the desired item number and quantity to obtain product availability and pricing.

<table>
<thead>
<tr>
<th>PART NUMBER</th>
<th>DESCRIPTION</th>
<th>PRICE</th>
<th>QTY</th>
</tr>
</thead>
<tbody>
<tr>
<td>5190-4823</td>
<td>SureSelectXT Custom 0.5-2.9Mb, 96, Auto, RO</td>
<td>Request Quote</td>
<td></td>
</tr>
<tr>
<td>5190-4822</td>
<td>SureSelectXT Custom 0.5-2.9Mb, 96, RO</td>
<td>Request Quote</td>
<td></td>
</tr>
<tr>
<td>5190-4821</td>
<td>SureSelectXT Custom 0.5-2.9Mb, 16, RO</td>
<td>Request Quote</td>
<td></td>
</tr>
<tr>
<td>5190-4820</td>
<td>SureSelectXT Custom 0.5-2.9Mb library (up to 55k oligos), 480 samples with Automation. Additional reagent kit needed, see details.</td>
<td>$149,039.00</td>
<td></td>
</tr>
<tr>
<td>5190-4857</td>
<td>SureSelectXT2 Custom 0.5-2.9Mb, 96</td>
<td>Request Quote</td>
<td></td>
</tr>
<tr>
<td>5190-4859</td>
<td>SureSelectXT2 Custom 0.5-2.9Mb, pre-capture index up to 480 samples</td>
<td>Request Quote</td>
<td></td>
</tr>
<tr>
<td>5190-4904</td>
<td>SureSelectXT Custom 12-24Mb library (up to 220k oligos) re-order, 480 samples</td>
<td>Request Quote</td>
<td></td>
</tr>
<tr>
<td>5190-4903</td>
<td>SureSelectXT Custom 12-24Mb, 96, Auto, RO</td>
<td>Request Quote</td>
<td></td>
</tr>
<tr>
<td>5190-4902</td>
<td>SureSelectXT Custom 12-24Mb, 96, RO</td>
<td>Request Quote</td>
<td></td>
</tr>
<tr>
<td>5190-4901</td>
<td>SureSelectXT Custom 12-24Mb, 16, RO</td>
<td>Request Quote</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2.19. Large Volume Purchase of Agilent SureSelectXT Baits (2019).** The only advertised price on Agilent's website at the time of writing this chapter is for 480 reactions, which is in the same range of the number of reactions we can perform using an order of 1 umole of our Sindbis probe. Perhaps in the future inflation has made it such that comparison to contemporary purchasing power is difficult for the reader. In 2019 the value is equivalent to three kilograms of gold, 40,000 bushels of corn, a new Mercedes-Benz S-Class AMG high performance model sedan, or 4 years of a Weill Cornell graduate student's stipend.
Similarly the MyOne C1 Dynabeads from Invitrogen can also be ordered with a volume discount. Due to the high number of such enrichment reactions that our lab conducts, we typically order 50 ml's of beads per order (5 x 10 ml bottles, Thermo Product # 65002), and receive a 52% volume discount, which amounts to $88 per ml of magnetic streptavidin beads, or $8.80 per mg. Similar volume discounts can likely be attained through discussions with the Invitrogen salesperson that represents the reader's campus. Using such volume discounts the total cost of our method is an order of magnitude cheaper per reaction, and consequently far more scalable within a specific budget, than the most commonly used method in the literature (Table 2.10). Upon the initial capital investment for a large volume of reagents, 49 reactions using our method can be conducted for $1,000, whereas the same amount provides just around 2 reactions using the SureSelectXT system (2016 prices).

Furthermore, the magnetic streptavidin beads are claimed to be reusable by the vendor for multiple reactions, further reducing cost. Given that we never tested if our elution conditions are degrading some component of the biotinylated DNA oligo instead of separating biotin from streptavidin, we never pursued this option. Other vendors offer beads for a lower price, however the reader must take caution and calculate the actual milligrams of beads provided instead of the volume stated as many vendors provide a lower density slurry.
Common options include beads of the agarose and magnetic variety, and biotin binders avidin, neutravidin (avidin with glycosylations removed), and streptavidin. The accompanying manual claims that streptavidin has the advantage of lower non-specific binding than avidin and neutravidin, though we have not tested this statement. Additionally, the biotin binding capacity and nonspecific binding ability may be different and optimization experiments likely must be performed.

Given such options, why were Invitrogen's MyOne C1 Dynabeads specifically chosen and optimized? Discussions and lab meeting presentations during the planning phase made it common knowledge within the lab that I was intent on attempting such an enrichment method. Postdoc Dr. Margaret Scull received a job offer to start her own lab around that time, and bequeathed me her bottle upon her departure. The procedure started showing signs of working while using that bottle I already had, and as such I continued using the brand. Discussions with Dr. Jeremy Dittman highlighted several features of these beads that I had not given much thought to prior. The 'one' in 'MyOne' and 'C1' likely refer to the 1 µm diameter of these beads, as compared to the 2.8 µm diameter of Invitrogen's M-280 Dynabeads for example. The production of spherical polystyrene beads of uniform diameter was discovered by Dr. John Ugelstad and commercialized with a company named Dynal AS in the 1970's, which explains the prefix in 'Dynabeads' [22]. The 'C1' beads seems to differ
from the 'T1' beads only in that the latter has been treated and pre-blocked with BSA, yielding it more appropriate for protein enrichment. Since we did not test alternative beads, the relative importance of such parameters is unknown.

2.12. The Biotinylated Probe is Selective but the Streptavidin Beads Non-Specifically Bind Input RNA.

The small proportion of Sindbis in vitro transcribed RNA in Figure 2.13 gave us cause for concern. When first moving to the 'real world' sample of Huh 7.5 RNA intracellular RNA infected with Sindbis, we initially tested the non-specific binding to the beads by running the experiment without an oligo added (Figure 2.20 A). In my notebook entry I chastise myself for forgetting to add the oligos, but it ended up being a good control. Without any oligo added, about 10% of the input cellular RNA and Sindbis RNA bound to the beads. Next we repeated the same experiment while including the same amount of DNA oligo as our 150x molar ratio against 5 ug Sindbis RNA in Figure 2.13 (200 pmol). As we had expected, we retrieved a higher proportion of Sindbis viral RNA when using the DNA probe (Figure 2.20 B).
Figure 2.20. A Proportion of Input RNA Non-Specifically Binds to Streptavidin Beads. Oligo capture experiments were performed in Sindbis infected Huh 7.5 Intracellular RNA without (A) and with (B) the biotinylated probe added. Sindbis genomic RNA (gRNA) and cellular RNA were assayed using qPCR (with 18s rRNA as a proxy for cellular RNA as described). In both cases, about 10% of the cellular RNA binds to the beads. Given that cellular RNA is a large majority of the input RNA, 10% constitutes a significant fraction in the eluted RNA pool. Error bars represent SEM of 4 technical replicates.
A) Sindbis Infected Intracellular RNA Capture Without Oligo

![Bar chart showing the comparison between Cellular RNA and Sindbis gRNA in the Input, Supernatant, and Elution stages of the capture process.](chart_A)

B) Sindbis Infected Intracellular RNA Capture With Oligo

![Bar chart showing the comparison between Cellular RNA and Sindbis gRNA in the Input, Supernatant, and Elution stages of the capture process.](chart_B)
Adjusting the axes of the above figure and specifically comparing Sindbis viral RNA shows that adding the biotinylated probe allows retention of over 50% of RNA in the elution (Figure 2.21 A). Thus our probe binds and allows retrieval of both synthetic in vitro viral RNA and viral RNA from cells, and the relative in vivo conditions of the latter do not induce secondary structures or such that inhibits binding. Being surrounded by a majority of cellular RNA, however, does decrease the efficiency relative to our previous results with pure in vitro transcribed Sindbis RNA. Additionally, the amount of cellular RNA bound to the beads and thus retrieved in our elution shows no difference with regard to the presence of oligo, at least in the 18s rRNA qPCR that we are using as a proxy for cellular RNA (Figure 2.21 B). This indicates that our DNA oligo is selective for Sindbis viral RNA and is not binding to the most predominant cellular RNA species.
Figure 2.21. Biotinylated DNA Probe Is Selective for Sindbis Genomic RNA. Adjusted axes from the above experiment (Figure 2.20) allows a clearer visualization of Sindbis genomic RNA (A) and Huh 7.5 cellular RNA (B) binding and retention. ~10% of Sindbis RNA nonspecifically binds to the beads, and this efficiency increases to ~50% with the addition of the biotinylated probe. In contrast, the percentage of input cellular RNA does not significantly change upon addition of the probe, indicating that the probe is selective for our target. Given that 18s rRNA is used as a proxy for cellular RNA in our qPCR assay, it is possible that the probe binds other off-target
RNA species, which next-generation sequencing can reveal. Error bars represent SD of 4 technical qPCR replicates.

Because cellular RNA is present in such a large majority in the infected intracellular RNA pool, the amount of cellular RNA non-specifically bound and retrieved is a higher than the amount of our target RNA specifically captured in the elution. Consequently the majority of the cost of a next-generation sequencing run on our eluted RNA pool would still go towards RNA we were not interested in. To solve this problem, we planned several washing procedures before elution that would hopefully prevent non-specific binding while not interfering with the specific biotin-streptavidin interactions. Previously we did not perform any wash steps (steps H.01 to H.05) because (i) the RiboZero protocol we used as an initial template did not contain any and (ii) we wanted to minimize the steps in which we could be losing RNA, which assisted in pinpointing ethanol precipitation as the source of RNA loss.

We decided on testing individually (i) a temperature method based on heating the sample after binding, (ii) a mechanical method based on vortexing the sample after binding, (iii) a chemical method based on blocking the beads prior to binding, and if all else fails, (iv) testing other types of beads that may hopefully exhibit less non-specific binding. This is a good point for the reader to hypothesize what the most effective method would be prior to reading the title of the next section. Heating is the most commonly used wash method in
all protocols that involve using the biotin-streptavidin bond for enrichment. The theory is to use a temperature that would disrupt non-specific binding and allow molecules thusly bound to be sucked up in the supernatant, while leaving molecules specifically bound via the biotin-streptavidin reaction to remain until elution. The RiboZero protocol incorporates some vortexing prior to a heating step, but vortexing is not present in most protocols, perhaps due to the technical difficulty of vortexing in anything higher than room temperature. Chemical blocking is used frequently in protein enrichment, such as with BSA in the 'T1' brand of dynabeads, and also in northern blots. Northern blot protocols provide a variety of blocking solutions to try, such as Denhardt's buffer, that incorporate exotic polymers and have been effective for selective RNA binding in a different context. Finally, if all else fails, we would test the manufacturer's claim that "streptavidin has the advantage of lower non-specific binding [than alternatives]" and compare it to chicken egg white avidin and neutravidin. Vendor statements accompanying a product have not always stood up to testing such claims. Perhaps instead of the type of avidin, the composition of magnetic beads is the cause of the problem, and using agarose beads would solve the issue.
2.13. Vortexing Selectively Removes Non-Specifically Bound RNA.

We first tested the efficacy of mechanical and temperature based methods on uninfected Huh 7.5 cellular RNA. For the mechanical method we performed 2 washes in 1x Bind/Wash buffer with 30 seconds of continuous heavy vortexing each time. We modeled the temperature based method on protocols that several lab members were performing regularly, and heated the beads at 55°C for 5 mins before the supernatant was pulled. This represented a slightly higher temperature than we had used in previous experiments, which was recommended by the RiboZero protocol. Many protocols recommend even higher temperatures for washes, but we felt those were too close to our elution temperature for our initial test, albeit in a different buffer and context [23]. For comparison, we included a control whereupon there was no heating or vortexing between the bead binding and elution steps (Figure 2.22). We would expect around 10% of the input cellular RNA to be non-specifically bound to the beads in the control condition, and hopefully less in the others.
Figure 2.22. Comparison of Vortexing-Based and Heating-Based Steps on Non-Specific Binding. An oligo capture experiment of uninfected Huh 7.5 cellular RNA before and after incorporating both protocols, in comparison to a control, assayed with qPCR. Heating to 55°C prior to supernatant removal made no difference in the amount of cellular RNA remaining in the elution compared to a room temperature control. In contrast, two washes with heavy vortexing removed the vast majority of non-specifically bound cellular RNA. Error bars represent SEM of 6 technical replicates.

To our complete surprise, the highly recommended heated washes did nothing compared to the room temperature control. And simple vortexing was highly effective and cleared the majority of non-specifically bound RNA! This was fortunate as vortexing represents the easiest and most convenient method, and it is rare that such convenience is also most effective. In an interesting observation of human behavior, presenting such findings to the lab did not
cause a single user of a protocol that incorporates heat-only wash steps to test the relative efficacy of heat versus vortex washing in their contexts, and make potential changes to the protocols they were following. To be fair, I am guilty of such dogmatic practices as well when it comes to making changes to things that have worked well enough previously, as the reader may have noticed at select points within this text. We had joked about such possibilities for this protocol; for example, what if the '5' setting on the vortex machine was gradually ascribed special importance by end users when it was simply what my machine happened to be set to when the above experiment was conducted!

Next we tested the effect of both methods on the biotin-streptavidin bond itself by performing the protocol on Sindbis virus \textit{in vitro} transcribed RNA. We reasoned that the simplest method to remove the eluted DNA oligos (\textbf{Figure 2.9}) would be via DNase digestion immediately after elution. Thus we performed an additional DNase digestion step on the mechanical wash elution sample and confirmed that viral RNA was not affected in our conditions (\textbf{Figure 2.23}).
Figure 2.23. Vortexing and DNase Digestion Does Not Affect Captured Sindbis RNA. An oligo capture experiment of Sindbis Toto1101 in vitro transcribed RNA incorporating both a vortex wash and DNase digestion protocol, in comparison to a control with neither step, assayed with qPCR. Vortexing and DNase digestion, which should remove a proportion of nonspecifically bound RNA (Figure 2.22) and the biotinylated DNA probes respectively, has no detectable off-target effects on viral RNA captured. Error bars represent SEM of 6 technical replicates.

Given that the easiest wash method was the most effective, this was retained in the final protocol (step H.02) and we abandoned our efforts with adjusting temperature, blocking buffers and different beads. Additionally the DNase digestion step was added such that only polymers of RNA is taken into the ethanol precipitation (step J.01 to J.03). Sufficient DNase digestion should release unbound Biotin-TEG molecules into the solution. With the hope that...
the unbound Biotin-TEG molecule would either not be precipitated during ethanol precipitation or if precipitated in the pellet, its presence would not inhibit downstream applications, we did not apply a size selection step for its removal.

Next we tested our new protocol additions (vortex wash and DNase digestion) against a 'real world' sample, Sindbis infected Huh 7.5 Intracellular RNA. As a control we included a sample captured in parallel without any wash or heating steps as above. We assayed the amount of Sindbis and cellular RNA in the input intracellular RNA and the output elutions, after DNase digestion but before ethanol precipitation, via qPCR (Figure 2.24).
Figure 2.24. Oligo Capture of Infected Intracellular RNA With Vortexing Steps. An oligo capture experiment was performed on Sindbis infected Huh 7.5 intracellular RNA samples with and without incorporating the vortexing wash steps, and viral and cellular RNA was assayed via qPCR. A) Analysis of total nanograms of RNA (y-axis) shows that an additional two-thirds of cellular RNA non-specifically bound to the beads is cleared by vortexing, while the percent of Sindbis genomic RNA captured (42% of input) is not affected. Numbers on bars represent the percentage of the respective RNA input. B) The input pool of RNA contains 9% Sindbis genomic RNA when compared to cellular RNA, and this increases to 30% if oligo captured without the wash steps, and to 51% if washed with vortexing. Y-axis and numbers on graph indicate the percent of Sindbis RNA of the total RNA in the pool (quantified as Sindbis RNA + Cellular RNA). The amount of Total RNA is indicated below the x-axis. Since total RNA is calculated by adding cellular RNA (using 18s rRNA as a proxy) and Sindbis genomic RNA together, the values are slightly lower than spectrophotometer values since subgenomic RNA is unaccounted for. The percentage of target RNA in the pool and the total amount of RNA in the pool being above a certain minimum for library preparation are the two most important factors for effective next-generation sequencing. Error bars represent SEM of 6 technical replicates.
Round 1 Capture
Sindbis Infected Huh 7.5 Intracellular RNA

Oligo Capture Step
- Cellular RNA
- Sindbis gRNA
The vortex washes clear about two-thirds of non-specifically bound cellular RNA relative to the control, whereas the percentage of Sindbis viral RNA in the elution is not affected at all (Figure 2.24 A). The percentage of Sindbis viral RNA of the total RNA pool is higher using our wash protocol, which is the most important factor in next-generation sequencing as long as the total RNA is above about 1 ug (Figure 2.24 B). This clearance is not as effective relative to our tests on uninfected cellular RNA (Figure 2.22), and this is perhaps due to the higher total RNA input in this experiment, or perhaps the binding of biotinylated DNA:RNA hybrids is sterically preventing the clearance of a fraction of non-specifically bound cellular RNA. As I had discussed with Dr. Jeremy Dittman, the physical properties behind the shearing of non-specific bonds during an imposed velocity would be of interest to biophysicists. On a macro-level the solution becomes a vortex; defined as a mass of whirling fluid, similar to a mini-whirlpool (perhaps why the common laboratory machine is named as such). The magnetic beads themselves likely have a level of rotation, in addition to rotating along the edges of the tube around a central point, somewhat like planets around a sun. Either these physical interactions between the moving beads and the salty bind/wash buffer liquid or the plastic sides of the semi-conical 1.5 ml tube, or both, is disrupting molecules non-specifically bound to the beads while the streptavidin-biotin and RNA:DNA hybridization bonds is unaffected. The impact of such 'pulling velocity' has been explored in the biophysics field for the biotin-streptavidin bond itself, but I
was unable to find such studies on non-specific binding of polynucleotides to streptavidin-coated beads [24].

2.14. Two Rounds of Capture Yields Final Pool of 95% Sindbis Viral RNA.

Next we wondered what if, instead of further optimizing such procedures, we simply combined the elutions from both the control and vortex protocols above, and performed a second round of the capture protocol? After all, the capture reagents are cheap, an additional day of my time isn't much more expensive, and a combination of both elution pools should be a solution of essentially pure RNA, of which about half is viral. Another capture round on this RNA pool should have no problem removing the majority of remaining cellular RNA if ethanol precipitation removed any potential contaminants. Thus we first ethanol precipitated the two output solutions from the prior experiment (Figure 2.24) and resuspended both pellets in a combined volume of 35 ul such that they can be combined into one capture experiment. We then performed another oligo capture round and ethanol precipitated the elution with all the steps according to the final written protocol, and assayed the combined input solution and the ethanol precipitated elution solution via qPCR (Figure 2.25).
Figure 2.25. A Second Round of Oligo Capture Further Clears Non-Specific Binding. An oligo capture experiment was performed on the combined elutions of the vortex wash and control experiments from figure 2.24 as input, and viral and cellular RNA was assayed via qPCR. Analysis of nanograms of each RNA (y-axis) shows that only 5% of input cellular RNA is retained in the capture, relative to 63% of input Sindbis genomic RNA. The Sindbis capture percentage was the highest thus far on combined viral and cellular RNA pools, and is likely due to the near equal ratio of each in the input pool, relative to a 10-fold excess of cellular RNA in infected intracellular RNA. Error bars represent SEM of 4 technical replicates.

The success of the second capture round confirms that there is no contaminants being pulled in the ethanol precipitation pellet of the first round's
elution that inhibits running another effective capture experiment. Combining both rounds and comparing the original input intracellular RNA before the first round to the output in the elution of the second shows a drastic reduction in cellular RNA while retaining a much higher relative fraction of the viral RNA (Figure 2.26).
Figure 2.26. Combined Two Round Capture Removes 99.9% of Cellular RNA. The input of the first round of capture (figure 2.24) was compared to the output of the second round of capture (figure 2.25). The total nanograms of Sindbis and Huh 7.5 Cellular RNA (y-axis) in each solution is shown in a linear scale with a gap (A) and a continuous log scale (B). Analysis of nanograms of each RNA shows clearance of 99.87% input cellular RNA, and retention of 20% of input Sindbis genomic RNA. Error bars represent SEM of 4 technical replicates.
Combined Rounds
Sindbis Infected Huh 7.5 Intracellular RNA

Oligo Capture Step
- Cellular RNA
- Sindbis gRNA

Total Nanograms

Input

Capture

0.13% 20%
B)

**Combined Rounds (Log Scale)**
Sindbis Infected Huh 7.5 Intracellular RNA

~ 4 Log Reduction of Cellular RNA!

---

**Oligo Capture Step**
- **Cellular RNA**
- **Sindbis gRNA**

---

**Graph Details**
- Log Total Nanograms
- Input and Capture categories
- ~100,000 for Input
- ~100 for Capture

---

0.13% 20%
Adjusting the axes and comparing each steps shows that 42% of viral RNA is captured in the first round, a fraction lost during ethanol precipitation, and 63% of that is captured in the second round, for 20% of the original input (Figure 2.27 A). In contrast, 95% of the cellular RNA is removed in the first round, a rather significant additional percentage removed during ethanol precipitation, and 95% of the remainder is removed in the second round, for a retention of only 0.13% of the original input RNA (Figure 2.27 B and Table 2.11).
Figure 2.27. Retention and Loss of Viral and Cellular RNA Per Round of Capture. Adjusted y-axes from the above figure (Figure 2.26) allows a clearer visualization of Sindbis genomic RNA (A) and Huh 7.5 cellular RNA (B) binding and retention along each step of a 2 round capture (x-axis). A) 58% of the original input Sindbis genomic RNA is lost during the first capture round before ethanol precipitation (E:P), an additional 11% is lost during ethanol precipitation, and another 11% lost during the second capture round and ethanol precipitation. B) In contrast, 93.3% of the original input cellular RNA is lost during the first capture round, an additional 2.5% lost during ethanol precipitation, and most of the remainder is lost during the second capture round and ethanol precipitation. Since we are using 18s rRNA as a proxy for cellular RNA in our qPCR assay, we cannot think of an explanation as to why a major fraction of 18s rRNA (which likely mirrors total cellular RNA) in the first round's elution is lost during ethanol precipitation. However it is of benefit for our purposes, so no complaints here. Error bars represent SEM of 4 technical replicates.
A) Sindbis RNA Only

![Bar chart showing total nanograms of Sindbis RNA for different capture steps.]

<table>
<thead>
<tr>
<th>Capture Step</th>
<th>Total Nanograms Sindbis RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>8,000</td>
</tr>
<tr>
<td>Round 1</td>
<td>3,408 (42%)</td>
</tr>
<tr>
<td>Round 1 E:P</td>
<td>2,468 (31%)</td>
</tr>
<tr>
<td>Round 2 E:P</td>
<td>1,634 (20%)</td>
</tr>
</tbody>
</table>

Oligo Capture Step

B) Cell RNA Only

![Bar chart showing total nanograms of Cell RNA for different capture steps.]

<table>
<thead>
<tr>
<th>Capture Step</th>
<th>Total Nanograms Cell RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>83,000</td>
</tr>
<tr>
<td>Round 1</td>
<td>5,613 (6.7%)</td>
</tr>
<tr>
<td>Round 1 E:P</td>
<td>1,402 (2.5%)</td>
</tr>
<tr>
<td>Round 2 E:P</td>
<td>121 (0.13%)</td>
</tr>
</tbody>
</table>

Oligo Capture Step
Table 2.11. Combined Two Round Capture Results on Infected Intracellular RNA. Sindbis virus and cellular qPCR of the two round capture experiment on Sindbis infected Huh 7.5 intracellular RNA (first round: Figure 2.24, second round: Figure 2.25, combined rounds: Figure 2.26). The mean values of total nanograms of each type of RNA are compiled and presented here for quick reference. Percentages are calculated as percent of input of the specific RNA type captured in that stated round. (*) Loss during ethanol precipitation (E:P) is the difference between the combined nanogram values of control and vortexed capture in the first round and the input in the second round (see Figure 2.27). Sindbis values match with our previous experience of 80% retention upon ethanol precipitation. In summary, around half of the target RNA, and only around 5% of the cellular RNA, is retained after each round of capture.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sindbis RNA</th>
<th>% of Input</th>
<th>Cell RNA</th>
<th>% of Input</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Round 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original Input / Rxn</td>
<td>4,100</td>
<td>100%</td>
<td>42,800</td>
<td>100%</td>
</tr>
<tr>
<td>Control Capture</td>
<td>1,730</td>
<td>42%</td>
<td>4,070</td>
<td>10%</td>
</tr>
<tr>
<td>Vortex 2x Capture</td>
<td>1,750</td>
<td>43%</td>
<td>1,700</td>
<td>4%</td>
</tr>
<tr>
<td><strong>Round 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control + Vortex E:P *</td>
<td>2,525</td>
<td>100%</td>
<td>2,110</td>
<td>100%</td>
</tr>
<tr>
<td>Round 2 Capture E:P</td>
<td>1,600</td>
<td>63%</td>
<td>110</td>
<td>5%</td>
</tr>
<tr>
<td><strong>Both Rounds</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original Input Total</td>
<td>8,200</td>
<td>100%</td>
<td>85,600</td>
<td>100%</td>
</tr>
<tr>
<td>Combined Capture E:P</td>
<td>1,600</td>
<td>20%</td>
<td>110</td>
<td>0.13%</td>
</tr>
</tbody>
</table>

Importantly, a two round procedure is an effective way, with minimal added reagent cost and an extra day of time, to enrich Sindbis viral RNA such that it is the majority species in the final RNA pool (Figure 2.28). Thus 95% of aligned reads from a next-generation sequencing run, by far the most
expensive portion, will be towards our intended target, and only a small fraction of our sequencing cost would be wasted. Additionally the RNA pool that is over 95% viral RNA is well over a microgram, allowing for efficient library preparation (Figure 2.28).
Figure 2.28. Two Rounds of Capture Enriches Sindbis Viral RNA to 95% of Total RNA Pool. Percentage composition of total RNA pool of Sindbis virus RNA and cellular RNA, from the two round capture experiment (figs 2.24, 2.25 and 2.26). Round 1 and round 2 represent post-ethanol precipitation values. The input Sindbis infected Huh 7.5 intracellular RNA contains 9% Sindbis genomic RNA when compared to cellular RNA. This increases to 54% upon the first capture round and ethanol precipitation, and to 95% upon the second capture round and ethanol precipitation. Sequencing of this pool will allow the vast majority of sequencing cost to go toward the intended target. Y-axis and numbers on graph indicate the percent of Sindbis RNA of the total RNA in the pool (quantified as Sindbis RNA + Cellular RNA). The amount of Total RNA is indicated below the x-axis. Error bars represent SEM of 4 technical replicates.
Next we prepared the elution output from the second round of capture for next-generation sequencing. This would tell us (i) if these qPCR values are accurate and reflect RNA alignment percentages, (ii) if there were any contaminants that can inhibit next-generation sequencing library prep, and (iii) if the viral RNA is full length, which would be indicated via an even coverage across the genome. If we were experiencing RNA degradation, there would be a large coverage peak at the genome location where the DNA oligo hybridizes.

We performed a CirSeq library prep as detailed in Chapter 3, and sequenced the sample on a HiSeq 2500. A coverage plot indicates that we got fairly even coverage throughout the viral genome (Figure 2.29 A). Thus the viral RNA was not being degraded in our hands. Of 43,904,297 CirSeq reads generated in the run, 41,741,229 reads (95%) aligned to the Sindbis virus genome, and only 2,163,068 reads (5%) did not, showing remarkable agreement with our qPCR assays (Figure 2.29 B). Of the reads that did not align to the virus, most (% of total reads) aligned to ribosomal RNAs; 45s rRNA (which contains 18s, 28s and 5.8s rRNAs within) and 5s rRNA. We did not align to mRNA's as it would be a minority species in our pool and alignment to all mRNA's would be many orders of magnitude more computationally intensive and did not seem worth it.
Figure 2.29. Next Generation Sequencing Confirms Composition of RNA Pool as 95% Sindbis Virus. The two round oligo capture pool (figure 2.28) was prepared as a CirSeq library and sequenced on a HiSeq 2500. A) The CirSeq reads were then aligned to Sindbis virus Toto1101 genome, and the number of reads that aligned (y-axis) at each nucleotide position of the genome (x-axis) is plotted. The dotted blue line indicates the average coverage per nucleotide of 274,260 reads. A 1-log variance in coverage is expected [25], and a lack of a distinct peak at the 40 bp site in which the DNA oligo hybridizes to the viral genome indicates the lack of RNA degradation during the capture protocol. B) The number of reads (x-axis) that aligned to the viral genome (orange) and the number of reads that did not (blue) are shown, and the percent of viral and non-viral reads are indicated. Remarkable agreement with RNA pool composition calculated using Sindbis genomic RNA qPCR and 18s rRNA qPCR as a proxy for cellular RNA (figure 2.28) confirms validity of calculations based on qPCR assays described in prior section. Of the $2000 spent on sequencing this sample on a HiSeq 2500, $1900 went towards generating coverage on our intended target genome, and only $100 did not.
A) 

Average Coverage: 274,260.8 Reads / Nucleotide

B) 

**Next-Generation Sequencing Results**

Oligo-Captured Sindbis Infected Huh 7.5 Intracellular RNA

- **95%** Sindbis-Aligned Reads
- **5%** Un-Aligned Reads
Thus in our intended downstream application, 95% of reads were against our target, Sindbis genomic RNA, and only 5% of reads were wasted on RNA we were not interested in. Compared to an input pool of 9% Sindbis RNA and 91% cellular RNA, which is what is typically sequenced in the field, this represented extensive cost savings during sequencing that allows for scalability and consequently less noisy and superior data.

2.15. Total RNA Input into Reaction Affects Percentage Viral RNA Captured.

Next we wondered how efficiently the protocol would handle a very high amount of infected intracellular RNA. We added around 270 ug of total intracellular RNA per reaction and performed three reactions in parallel for the first round, for a total of 0.8 milligrams of input RNA. The ethanol precipitated resuspensions were combined and a second capture round was performed on the total output.

Our initial input RNA was a pool of 7% Sindbis genomic RNA and 93% cellular RNA, and our output RNA pool contained 72% Sindbis genomic RNA and 28% cellular RNA (Table 2.12 A). High amounts of RNA seems to interfere with the efficiency of the reaction. Next-generation sequencing this pool on a HiSeq 2500 (Table 2.12 B) confirmed a lower viral RNA ratio than the prior
experiment. 72% viral RNA as detected by qPCR and 88% aligned to the virus genome in next-generation sequencing was fine (and superior to directly sequencing the input which is commonly done in the field), but was less efficient than our previous reaction, and consumed far more input RNA. We thus decided to stick to a maximum of 50 ug RNA per Sindbis capture reaction. In our subsequent experiments and as stated in our final protocol, we use 150 total ug of infected intracellular RNA split into three reactions for the first round. The pellets from the six ethanol precipitation tubes (since each reaction is split into two ethanol precipitations) is combined into a 35 ul volume and one reaction is performed for the second round. This has consistently yielded a pool of 95%+ Sindbis genomic RNA that can be taken to sequencing or any other downstream application.
Table 2.12. Combined Capture Results on High Input Infected Intracellular RNA. Sindbis virus and cellular qPCR of a two round capture experiment on undiluted Sindbis infected Huh 7.5 intracellular RNA as input. 3 individual captures are performed on the same input and the output is combined for the second round. A) The mean values of total nanograms of each type of RNA are compiled and presented here for quick reference. Percentages are calculated as percent of input of the specific RNA type captured in that stated round. The final output RNA pool contains 72% Sindbis genomic RNA (1.13 ug) and 28% cellular RNA (0.44 ug). B) CirSeq library preparation, next generation sequencing and alignment results on this output pool. The percent of reads aligned to the viral genome is slightly higher than expected via the qPCR results, to our benefit, but was lower than the prior experiment. Thus using extremely high amounts of infected cellular RNA input (267 ug per reaction) is less effective at producing an output pool of near-pure viral RNA than a lesser input (50 ug per reaction, table 2.11), indicating interference of high amounts of non-target RNA in input.
### A)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sindbis RNA</th>
<th>% of Input</th>
<th>Cell RNA</th>
<th>% of Input</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Round 1</strong></td>
<td>nanograms</td>
<td>nanograms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Input (3 Rxns)</td>
<td>55,970</td>
<td>100%</td>
<td>745,500</td>
<td>100%</td>
</tr>
<tr>
<td>Capture</td>
<td>2,430</td>
<td>4%</td>
<td>19,280</td>
<td>3%</td>
</tr>
<tr>
<td><strong>Round 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Input</td>
<td>2,430</td>
<td>100%</td>
<td>19,280</td>
<td>100%</td>
</tr>
<tr>
<td>Capture</td>
<td>1,130</td>
<td>47%</td>
<td>440</td>
<td>2%</td>
</tr>
<tr>
<td><strong>Both Rounds</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Input (3 Rxns)</td>
<td>55,970</td>
<td>100%</td>
<td>745,500</td>
<td>100%</td>
</tr>
<tr>
<td>Capture</td>
<td>1,130</td>
<td>2%</td>
<td>440</td>
<td>0.06%</td>
</tr>
</tbody>
</table>

### B)

<table>
<thead>
<tr>
<th>Alignment</th>
<th># CirSeq Reads</th>
<th>Percentage of Total Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Reads</td>
<td>51,841,831</td>
<td>100%</td>
</tr>
<tr>
<td>Sindbis-Aligned Reads</td>
<td>45,720,877</td>
<td>88%</td>
</tr>
<tr>
<td>Un-Aligned Reads</td>
<td>6,120,954</td>
<td>12%</td>
</tr>
</tbody>
</table>
Combining the data from all such reactions performed with varying amounts of infected intracellular RNA as input shows that the percentage of Sindbis RNA captured is affected by the total input per capture reaction. We chose our input values based on the percentage of target RNA we wanted in the final pool (95+%) and a minimum of 1 ug total RNA in the final pool. The user may adjust such input values and the number of rounds to find a balance based on their needs.

**Figure 2.30. Total RNA Input Per Reaction Determines Target RNA Capture Efficiency.** Compilation of prior oligo capture experiment efficiencies
mapped to RNA input per reaction vessel; x-axis represents the amount of total RNA input per reaction, and y-axis represents Sindbis RNA captured in the output as a percent of the specific amount of Sindbis RNA in the input (as opposed to percent Sindbis composition of RNA pool in figure 2.28 for example). The dotted orange line indicates the percent of Sindbis captured in the complete absence of cellular RNA (150x sample in figure 2.17). The total amount of Sindbis RNA captured in micrograms is also indicated below the x-axis. Sindbis genomic RNA as a percentage of total intracellular RNA varies between 5 and 15% in displayed experiments depending on infection and capture round. In our experiments, we achieve a maximum capture of 1.8 ug of Sindbis viral RNA per reaction with an input of 50 ug infected cellular RNA, which is then further refined in the second round. The amount of cellular RNA is not indicated because these data comprised of experiments performed before and after inclusion of vortexing-based washes, and thus would be highly variable. Thus the percentage of target RNA captured from the initial amount of target RNA in the input, the composition of the output RNA pool, and the total amount of RNA required in the output RNA pool must be balanced by the input RNA amount per reaction and the number of rounds. Error bars represent SEM of 4 replicates, some error bars are too small to be displayed.

2.16. Lessons from Zika: Effect of Mismatches and Annealing Conditions on Binding.

Around when Sindbis capture started working well, due to the aforementioned outbreak, Mohsan and Yingpu became interested in similarly enriching Zika virus for a variety of downstream analyses in which pure RNA would provide an advantage, and the information would be lost upon reverse transcription and PCR. We bioinformatically predicted oligos that should bind to the Cambodian strain of Zika virus using the same settings as in Figure 2.6. We
then ordered two of the predictions as biotinylated Zika Cambodian strain (ZCS) oligos using the procedure detailed in the oligo design section above:

ZCS #1:
/5BiotinTEG/CCTATCTCCATTCATACCAACAACCATCTTTAGCCCGG (+HPLC Purification). Hybridizes at nucleotides 3458 to 3497.

ZCS #2:
/5BiotinTEG/GTTGTCAATGGCCTCCTGGAATCTCTCTGTCATATGTCCT (+HPLC Purification). Hybridizes at nucleotides 6506 to 6545.

Next we tested capture against 5 ug of *in vitro* transcribed Zika Cambodian strain RNA using the same protocol as for Sindbis and received quite good results; 89% and 84% for ZCS oligos #1 and #2 respectively (*Figure 2.31 A*).

Then our Zika strain of interest in the laboratory changed from the Cambodian strain to the Puerto Rican strain. I am not sure who made this executive level decision or why it was made, but it did give us an opportunity to test oligo annealing properties against RNA targets that contain mismatches. ZCS oligo #1 and #2 had 1 and 2 mismatches, respectively, when aligned the Puerto Rican strain. Testing the Cambodian strain oligos against *in vitro* transcribed Puerto Rican strain Zika virus yielded a 15% reduction in capture efficiency for the oligo with one mismatch and a 9% reduction for the oligo with two mismatches (*Figure 2.31 B*). Thus one or two mismatches between a target RNA and a ~40 bp oligo do affect annealing and capture efficiency, but not by very much. The strain switch and this consequent experiment provided the advantage of convincing us that we can include the 40 bp region of Sindbis
that is bound by the oligo in our CirSeq analyses in chapter 3, as the variant frequencies within will only be affected slightly.

**Figure 2.31. Zika in vitro Transcribed RNA Capture with and without Mismatches.** Comparison of binding of 40 bp biotinylated probes designed for Zika Cambodian strain against (A) Zika Cambodian strain in vitro transcribed RNA which has perfect complementarity, and (B) Zika Puerto Rican strain in vitro transcribed RNA which has mismatches within the probe hybridization region, using qPCR assay. Error bars represent SEM of 4 technical replicates.

Given the fairly high capture percentage against in vitro transcripts we next tested capture against infected cellular RNA. We ordered oligo #1 with the mismatch removed against the Puerto Rico strain (ZPR #1). Flaviviruses such as Zika virus generally infect mammalian cell cultures more weakly than alphaviruses. Consequently, harvesting enough extracellular virions, which would provide a fairly pure source of viral RNA, to meet minimum RNA thresholds for many downstream analyses would require impractically large-scale infections. Thus, unlike alphaviruses, enriching viral RNA from
intracellular RNA is the only practical method of consistently retrieving a microgram or more of relatively pure flavivirus RNA. However, flaviviruses also produce less viral RNA inside the cell and we sought to test several parameters optimize the protocol further. For example, during a maximally efficient infection of supportive mammalian cells, Zika viral RNA will only compose around 1% of total intracellular RNA, compared to around 5% for Sindbis genomic RNA.

Yingpu infected Huh 7.5 cells with Zika Puerto Rico strain for 48 hours and extracted intracellular RNA. Mohsan performed one round of the oligo capture protocol using ZPR oligo #1, and I did the qPCR assay and analysis on the output. As a truly collaborative effort, we each had rotating roles for each experiment. As explained in the next section, if we had paid more attention to this rotation we could have saved months of futile efforts. We initially tested the effect of increasing hybridization times on binding of Zika viral RNA and nonspecific cellular RNA. We tested increasing each of the 10 minute incubation times (Steps F.05 and G.04) to overnight at 4°C separately and in combination (Figure 2.32 and Table 2.13). As a control, we also tested the standard protocol with in vitro transcribed Zika Puerto Rican strain RNA.
Figure 2.32 Greater Hybridization Times Increase Binding of Both Viral and Nonspecific RNA. An oligo capture experiment was performed on Zika Puerto Rico strain infected Huh 7.5 intracellular RNA samples, and viral (A) and cellular RNA (B) was assayed via qPCR. Increasing hybridization times does capture more viral RNA in the elution, but also captures far more cellular RNA in the elution, decreasing purity relative to shorter times. Protocol #0 is a control using in vitro transcribed ZPR RNA, and therefore does not contain any cellular RNA. Protocol #1 is the standard protocol with 10 min room temperature incubations, #2 contains an overnight 4°C probe + RNA annealing step, #3 contains instead an overnight 4°C binding step of the DNA:RNA complexes to magnetic beads, and #4 contains both overnight steps. Y-axes represent percentage of input, and the actual amounts of cellular RNA is 2 orders of magnitude more than Zika RNA (see table 2.13 for exact amounts). The input and elutions were assayed, and the remaining percentage that is unaccounted for could be due to loss during wash steps or degradation during long incubation times. Error bars represent SEM of 4 technical replicates.
A) Zika virus RNA

![Zika virus RNA graph](image)

% Percent of Input Zika RNA

Oligo Capture Annealing Protocol

- Yellow: Supernatant [Bad!]
- Orange: Capture [Good!]

B) Cellular RNA

![Cellular RNA graph](image)

% Percent of Input Cell RNA

Oligo Capture Annealing Protocol

- Light Blue: Supernatant [Good!]
- Blue: Capture [Bad!]
Table 2.13. Effect of Hybridization Times on Viral and Nonspecific RNA Capture. The amounts of captured Zika virus PR strain and cellular RNA, in nanograms, from the elutions of a capture experiment that tested increasing hybridization incubation times (see figure 2.32, with matching protocol numbers). The percentage of the viral RNA input into each reaction that was captured is also listed. Even the standard protocol contains very high amounts of cellular RNA, which is likely a function of the extremely high amounts of total RNA input into each reaction (see figure 2.30). In subsequent experiments, the standard protocol removed the expected ~97% of cellular RNA when ~50 ug total RNA is input into the reaction (figure 2.33).

<table>
<thead>
<tr>
<th>#</th>
<th>Annealing Protocol</th>
<th>Zika RNA (ng)</th>
<th>Zika Efficiency</th>
<th>Cellular RNA (ng)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IVT Input</td>
<td>2,231</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>#0</td>
<td>IVT Capture Control</td>
<td>734</td>
<td>33%</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Infection Input</td>
<td>1,742</td>
<td></td>
<td>465,792</td>
</tr>
<tr>
<td>#1</td>
<td>10 min Standard Protocol</td>
<td>196</td>
<td>11%</td>
<td>27,125</td>
</tr>
<tr>
<td>#2</td>
<td>Overnight Probe Hybridization</td>
<td>676</td>
<td>39%</td>
<td>239,500</td>
</tr>
<tr>
<td>#3</td>
<td>Overnight Bead Binding</td>
<td>814</td>
<td>47%</td>
<td>256,250</td>
</tr>
<tr>
<td>#4</td>
<td>Overnight Both Binding Steps</td>
<td>879</td>
<td>50%</td>
<td>346,750</td>
</tr>
</tbody>
</table>

Overnight binding protocols did increase intracellular Zika RNA binding between 3.5x and 5x as the 10 min incubations (Figure 2.32 A). However, nonspecific cellular RNA binding increased between 9x and 13x as the 10 min incubations (Figure 2.32 B). These nonspecifically bound RNAs did not decrease with the same vortexing protocol; increased incubation times somehow allows tighter bonds to form which withstand vigorous vortexing. Thus binding incubation times represent a balance; whereas specific binding
increases, nonspecific binding seems to increase more rapidly reducing the ratio of viral RNA to cellular RNA. Our goal was purity as long we met a minimum RNA threshold, and we continued with the standard 10 min incubation protocols. The user may wish adjust this balance to their specific needs, and perhaps test time intervals between 10 mins and ~16 hours. Interestingly and unexpectedly, nonspecific binding increased 9-fold when the biotinylated DNA oligo and RNA were incubated overnight and extremely slowly cooled from 68°C to 4°C, indicating that even the DNA-RNA hybridization is subject to nonspecific binding given enough time. Previously we had assumed that the magnetic streptavidin beads represented the only major source of nonspecific binding.


An astute reader with RNA extraction experience may question; why combine the two methods? Why not just use the Qiagen RNeasy kit (or any equivalent column-based kit) or why not continue the Trizol/chloroform extraction as one typically does with an ethanol precipitation? The former question is easier to answer, we surmised that phenol is more scalable and would do a better job at separating the high number of cells we were attempting to extract RNA from than the buffers included in the kits. For example, the RLT buffer in the Qiagen
RNeasy kit has a maximum per reaction recommendation of 10 million cells, and we routinely use multiples of that as input. The latter question presents quite the conundrum.

Mohsan performed a very large scale Zika Puerto Rico strain infection of Huh 7.5 cells to produce enough intracellular RNA for many optimization experiments, and qPCR assays indicated that it was a great infection, with viral RNA comprising almost 2% of total RNA. However, all subsequent oligo capture experiments kept failing miserably, no matter who performed it. An air of gloom and desperation set in after months of such failures (Figure 2.33 A), were all previous successes a fluke? Going through our notes we noticed a curious trend; during our assignment rotations the capture worked whenever I performed the infection and cellular RNA extraction, failed whenever Mohsan performed it, and sometimes worked when Yingpu performed it.

Years ago Yingpu had taught me what we called a 'lazy' cell RNA extraction protocol, whereupon mixing the cells with Trizol, instead of following it with a typical ethanol precipitation, ethanol is added to the supernatant and put through a Qiagen RNeasy column to save an hour or so and some additional pipetting. Ever since I noticed Yingpu doing this and asked him what he was doing, I had adopted it for all my cell RNA extractions, and Yingpu performs it occasionally depending on his mood. Mohsan never adopted this shortcut.
Looking through our experiments logs after months of frustration, we noticed that oligo capture worked whenever a column was used for purification, and utterly failed whenever ethanol precipitation was used. This seems inexplicable and very odd to us as ethanol precipitation has no noticeable effects in most other downstream procedures we perform. Nevertheless, we tested capture between infected cellular RNA purified via ethanol precipitation, and the same RNA passed through a column (Fig 2.33 B).
Figure 2.33. Prior RNA Column Purification Step is Critical for Successful Oligo Capture. A) Notes from my lab notebook during the months of failed oligo capture experiments expressing despair. B) An oligo capture experiment was performed on Zika Puerto Rico strain infected Huh 7.5 intracellular RNA samples with and without incorporating a Qiagen RNeasy column purification step to Trizol extracted and ethanol precipitated RNA. Viral RNA was quantified via Zika qPCR for both column (orange) and no-column (red) samples. Cellular RNA was quantified via 18s rRNA qPCR on only the column purified sample (blue). Analysis of total nanograms of RNA (y-axis) shows that (i) using columns allows an excellent recovery of the majority of viral RNA in the elution, (ii) not using a column leads to undetectable amounts of viral RNA in all oligo capture solutions, but it can still be detected in the input solution, and (iii) the vast majority of cellular RNA (97%) is being depleted in one round. Additionally, as expected when using the standard 10 min hybridization steps protocol, vortex washes eliminate a significant portion of cellular RNA that remains on the beads after the supernatant is removed, but does not remove any detectable amounts of viral RNA. Numbers on bars represent the percentage of the respective RNA input. Error bars represent SEM of 4 technical replicates. C) Segments of an electronic mail chain between Yingpu, Mohsan and I expressing excitement that we were back in business.
A) Still getting undetectable amounts of RNA.
   "maybe due to something in RNA? No DNase Treatment?"
   "buffers or beads mixed up?"
   "Start back on solid footing! Do in vitro exp, replicate first infection experiment, then push forward systematically.

B) Zika Puerto Rico Strain Capture
   RNA Extraction: Ethanol Precipitation vs RNA Columns (Post-Trizol)

   Oligo Capture Step
Pradeep Ambrose <pmambrose@gmail.com>  Tue, May 2, 2017 at 8:58 PM
To: Mohsan Saeed <msaeed@mail.rockefeller.edu>  Cc: Yingpu Yu <yyu@mail.rockefeller.edu>

Hi Mohsan,

Yes, looks like we’re still having a persistent problem with the infection. We can discuss more tomorrow, First, let’s extract infected cellular RNA using Trizol and then Qiagen columns like I do for Sindbis, using the same protocol. Perhaps there’s some interfering factor being pulled down in the glycogen. I had the same issue with PEG precipitation, where Trizol + ethanol precipitation interfered with the fragmentation, and it inexplicably went away while using columns. I’ve been using columns to purify cellular RNA ever since.

Mohsan Saeed <msaeed@mail.rockefeller.edu>  Wed, May 3, 2017 at 7:30 PM
To: Yingpu Yu <yyu@mail.rockefeller.edu>  Cc: Pradeep Ambrose <pmambrose@gmail.com>, Mohsan Saeed <msaeed@mail.rockefeller.edu>

Hi Yingpu,

Pradeep told me that you plan to run the oligo capture experiment tonight. If you plan to do so, do you think it would be nice to first pass some of the cell-extracted RNA through RNeasy mini column, and run it in parallel to see if there is something in our RNA prep that is inhibiting the oligo capture?

This is Pradeep’s idea. He was looking for you just now, but could not find you. So he asked me to convey this message.

Yingpu Yu <yyu@mail.rockefeller.edu>  Sun, May 7, 2017 at 11:33 PM
To: Mohsan Saeed <msaeed@mail.rockefeller.edu>  Cc: Pradeep Ambrose <pmambrose@gmail.com>

That’s great result. The column is the magic! Pradeep, I still don’t know what happened, but it’s great it works! I think it’s good idea to run the rRNA to confirm that we removed most of the rRNA from the Oligo capture.
Somehow when the cellular RNA is not put through a column, viral RNA is detectable in the ethanol precipitated RNA extraction input, but not detectable above trace amounts in any solution during the oligo capture procedure (Figure 2.33 B, red bars). However, viral RNA is detectable in the elution when a column is used prior to oligo capture, and we even got an excellent recovery of 61% of input (Figure 2.33 B, orange bars). Why this occurs, we still have absolutely no idea. Perhaps trace amounts of phenol or chloroform is removed by columns but not by ethanol precipitation, and has a tremendous effect on capture? It is unlikely to be trace amounts of ethanol or glycogen or sodium acetate since in vitro transcribed RNA that is ethanol precipitated is captured without issue. Us virologists pride ourselves on our phase separation based RNA extraction prowess. Despite thousands of such extractions for successful downstream procedures between the three of us, upon presenting such results several of our labmates even questioned our credentials as virologists! But no matter, we were finally back in business (Figure 2.33 C)!

Of the many strokes of luck I've had the good fortune to receive during my PhD, one of the luckiest by far was catching Yingpu in the act and learning his combined Trizol:Chloroform and RNA column protocol prior to starting any work on intracellular viral RNA capture. Hard to say, but if all of my initial Sindbis intracellular RNA oligo capture experiments produced the same results as the ethanol precipitation sample above (Figure 2.33 B, red bars), it
seems unlikely that I would have pinned down this issue prior to a single pseudo-working experiment, and perhaps more likely I would have just stopped trying. Until the precise reason why RNA columns enable oligo capture enrichment is elucidated, I would advise the reader to place their phase separation prowess ego aside and follow steps B.01 to C.04 precisely. And if the reader does discover the source of this conundrum, please update us!

2.18. Large Variability in Efficacy of Bioinformatically Predicted Hybridization Oligos.

Given that the two Cambodian strain oligos gave slightly different results, we next wondered if we can boost oligo capture efficiency by testing multiple oligos and then also by combining them. All commercial capture applications, other than PolyA tail hybridization, utilize a pool of oligos. Perhaps the shell shock of having these experiments fail for months made us more irrationally paranoid about RNA degradation as well, and if that were the case, tiling multiple oligos over the length of the genome would also enhance capture in theory. Additionally, we had yet to bioinformatically predict hybridization oligos against the Puerto Rico strain of Zika virus, and the oligo we'd used so far was predicted against the Cambodian strain and the mismatched fixed (ZPR #1). We thus ran the Oligo 7 software on Zika Puerto Rico strain using high
stringency settings to generate a list of oligo predictions, and confirmed that
ZPR #1 was also predicted for this strain. Out of 12 total predictions on this
list, we then selected at random and ordered the 3 following additional oligos
(ZPR's #2, #3 & #4):

ZPR #1:
/5BiotinTEG/CCTTATCCATTCCATAACACAGCCATCTTTAGCCCGG (+HPLC Purification). Hybridizes at nucleotides 3458 to 3497 (40 bp), in the
NS1 protein region.

ZPR #2:
/5BiotinTEG/ATGAGCACCACCAGCAATAGGAACACAACAATGAGGACAC (+HPLC Purification). Hybridizes at nucleotides 6784 to 6823 (40 bp), in the
NS4A protein region.

ZPR #3:
/5BiotinTEG/GCCCTATGAGAGATCCACACCACAAGTCTTCCCTTTT (+HPLC Purification). Hybridizes at nucleotides 10194 to 10230 (37 bp), in the
NS5 protein region.

ZPR #4:
/5BiotinTEG/GGCTTCGGCTCTCGGTGAATTGGGCGTTATCTCAACTTTTC (+HPLC Purification). Hybridizes at nucleotides 1472 to 1511 (40 bp), in the
Env protein region.

As described in Figure 2.6, each of these had very high hybridization scores
assigned by the program when using those same settings. We also created an
oligo pool by mixing the four oligos together. We then ran an oligo capture
experiment on Zika Puerto Rico strain infected Huh 7.5 intracellular RNA,
using the four oligos individually and as an equimolar mixed pool (Figure
2.34).
Figure 2.34. Bioinformatically Predicted Hybridization Oligos Are Highly Variable in Capture Efficiency. Comparison of capture efficiency between 40 bp oligos predicted to hybridize with Zika Puerto Rican strain RNA. Oligos ZPR #1-4 are described in the text, and mix indicates an equimolar pool of all four oligos combined for the capture experiment. Y-axis shows the amount in nanograms of Zika RNA assayed via qPCR. The capture experiment was conducted on 50 ug of Zika infected Huh 7.5 intracellular RNA, and the specific amount of Zika RNA input into each reaction is indicated in the leftmost bar. Percent of this RNA captured by each oligo or the mix of oligos is shown above each bar. Error bars represent SEM of 4 technical replicates.

To our surprise, there was tremendous variation in the capture efficiency of each of the oligos. One (ZPR #3) performed twice as well as what we were using previously (ZPR #1), and is even several base pairs shorter than the rest. An oligo (ZPR #4) barely captured any of the input Zika RNA, and another (ZPR #2) caught almost none at all. Therefore, within various oligos of
similar lengths, we see a range of capture efficiency between a nearly non-existent 0.2% and a terrific 89%. Using a mix of these oligos was slightly better than the calculated average, but nearly 30% less effective than ZPR #3. Thus, using one highly effective oligo is better than using a mix of great, average, and poor performance oligos. By extension, when targeting a single RNA species, we find it unlikely that a giant pool of numerous bioinformatically predicted, untested, and lengthy oligos would provide better capture than a single, short one that was tested and proven to be highly effective. Such a giant pool, as is commonly used and recommended in the literature, is guaranteed to be many orders of magnitude more expensive though.

Equipped with these tested oligos, we now were finally confident enough in our Zika capture procedure to test a second round of capture, to see if we can further decrease cellular RNA while retaining at least a half a microgram of total RNA for downstream analyses. We pooled together the elutions from the ZPR #1, ZPR #3 and mixed oligos, and performed a second round of capture. We still hadn't tested whether the captured viral RNA was full length, and still being slightly irrationally paranoid about degradation we performed this capture round using a mix of ZPR #1 and ZPR #3, which were both proven to be effective (Figure 2.35).
Figure 2.35. Two Rounds of Capture on Zika Infected Intracellular RNA Enriches Viral RNA from ~2% to ~99% of Total RNA Pool. Zika virus and cellular RNA qPCR of a two round capture experiment on Zika infected Huh 7.5 intracellular RNA, in which the elutions from ZPR #1, ZPR #1 and Mix samples shown in figure 2.34 were combined and subject to a second round of capture. Some of this sample was removed for other assays prior to the second round, and the values are adjusted accordingly. A) Analysis of the total amount viral and cellular RNA present in the input and in the elutions of each round of capture. The percentage of the input viral and cellular RNA captured in the elution is shown above each bar. After 2 rounds, half of the viral RNA from the intracellular RNA is retained, whereas 4 logs of cellular RNA is depleted. Error bars represent SEM of 6 technical replicates. B) Pie charts displaying the ratio of viral to cellular RNA in the input and each round's elution. This ratio determines the ratio of sequencing reads that can be acquired from unbiased NGS analyses of RNA (i.e., without making an amplicon of a specific region). The numbers shown indicate the amount in nanograms of each type of RNA in the pool, along with the relative percentage. The RNA in the elution from round 2 was processed using CirSeq and these results are shown in figure 2.38.
A)

Two Round Capture
Zika Infected Huh 7.5 Intracellular RNA

![Bar chart showing the total nanograms of RNA for each round and comparison between Cellular RNA and Zika RNA.]

B)

![Pie chart showing the distribution of nanograms of cellular RNA and Zika RNA throughout the process.]

- **Input**: 61,914 (98.3%)
- **Round 1**: 1,148 (61.8%)
- **Round 2 (Final)**: 516 (98.7%)

**Legend**:
- Blue: Nanograms Cellular RNA (% of RNA Pool)
- Orange: Nanograms Zika RNA (% of RNA Pool)
After two rounds of capture, in this experiment we had a retention of 47% of the Zika RNA within the input infected intracellular RNA, and cleared 99.99% (4 logs!) of cellular RNA, from 62 micrograms to 7 nanograms, as assayed via qPCR (Figure 2.35 A). There is 2 logs more cellular RNA than Zika RNA in the starting sample, and this 4 log reduction leads to 2 logs less cellular RNA than Zika RNA in the final sample. The ratio of Zika RNA in the RNA pool, which should represent the ratio of Zika-aligned reads in an NGS dataset derived from this pool, goes from 2% in the original intracellular RNA, to 38% after round one, to 99% after the second round (Figure 2.35 B). However, we still hadn't detected if the RNA is full length or was degraded along the way. And this was the best result we'd acquired thus far and seemed too good to be true - these values and ratios were calculated using qPCR assays, and what if some inexplicable problem cropped up in the machine or in the reagents? We had in hand what should be half a microgram of essentially pure Zika viral RNA, and we decided that the best way to answer this is to perform one of our intended downstream analysis; CirSeq sequencing. The details of this experiment is provided in the next section.

It was decided that further oligo capture experiments will be conducted with only ZPR #3. To ensure that the great results above were not a fluke, we reproduced the experiment for two rounds using ZPR #3. As a final optimization, we wondered if the better binding of ZPR #3 will allow us to
introduce more total RNA input into one reaction vessel without the detrimental effects on viral RNA capture percentage shown in Figure 2.30. We performed the first round of capture comparing a typical reaction input of 50 ug total intracellular RNA with increased inputs of 100 ug and 200 ug (Figure 2.36).
**Figure 2.36. Higher Total RNA Input in Reaction Increases Total RNA Retrieved but Reduces Viral RNA Ratio.** Zika virus and cellular RNA qPCR of one round of a capture experiment on Zika infected Huh 7.5 intracellular RNA, using oligo ZPR #3, in which varying amounts of total intracellular RNA was input into the reaction. A) The percentage of each respective type of RNA captured in the elution of one round of capture. Using the standard 50 ug input reproduced the ~90% viral RNA capture result in figure 2.34. As the input amount increases, the percentage of input cellular RNA in the elution increases, whereas the percentage of target Zika RNA decreases. B) Pie charts showing that this increased input amount reduces the ratio of viral RNA to cellular RNA in the elution. However, as shown in the bottom row, the total RNA in the elution increases. Certain downstream applications have high minimum total RNA thresholds, and the user can adjust their input amount accordingly with a slight compromise in purity. Additionally, a second round of capture can significantly increase purity, and both the input amounts and the number of rounds can be adjusted to meet minimum total RNA requirements.
A) Round 1 Input Amount Variation
Zika Infected Huh 7.5 IntraCellular RNA

% Percent of Input RNA

<table>
<thead>
<tr>
<th>Intracellular RNA Input Per Reaction</th>
<th>% Cellular RNA</th>
<th>% Zika RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 ug</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100 ug</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200 ug</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B) Total RNA Input: 50 ug
Per Rxn 30.4 %

Total RNA Output: 1.3 ug
Cellular RNA % of Total RNA Pool: 69.6 %
Zika RNA % of Total RNA Pool: 30.4 %

100 ug

Total RNA Output: 3.7 ug
Cellular RNA % of Total RNA Pool: 82.3 %
Zika RNA % of Total RNA Pool: 17.7 %

200 ug

Total RNA Output: 12 ug
Cellular RNA % of Total RNA Pool: 92.1 %
Zika RNA % of Total RNA Pool: 7.9 %
The ZPR #3 oligo once again captured around 90% of the Zika viral RNA from the input infected cellular RNA. As we had seen with Sindbis (Figure 2.30), varying the input RNA amounts per reaction vessel, even using a highly effective oligo, leads to a lower percentage of Zika RNA captured and a higher percentage of cellular RNA captured from the input intracellular RNA pool (Figure 2.36 A). Consequently, there is a decreased ratio of Zika RNA to cellular RNA in the capture elutions (Figure 2.36 B). However, the total RNA retrieved is higher (Figure 2.36 B, output row), and thus the input amount may be adjusted by the user depending on the minimum RNA threshold for subsequent analyses, and if a lower purity of the sample is tolerable. We combined these samples and performed a second round of capture, which purified the majority of the remaining cellular RNA, as expected. This RNA sample was then processed and analyzed for RNA methylation by Inna, and these data will be available in an upcoming publication.

2.19. Oligo Capture Enables High Coverage CirSeq of Low Titer Viruses Such as Zika Virus.

The two round capture on Zika Puerto Rico strain infected Huh 7.5 intracellular RNA shown in Figure 2.35 was processed for CirSeq. Since at the moment we were more curious about the quality of the sample more so than performing genetic analyses, we submitted it for a MiSeq run, which produces
less reads than the HiSeq runs used for Sindbis genetics (Figure 2.29). As we had expected, the vast majority of CirSeq reads aligned to the Zika genome, and an order of magnitude less reads aligned to ribosomal RNA (Figure 2.37).

![Next-Generation Sequencing Results](image)

**Next-Generation Sequencing Results**
Oligo-Captured Zika Infected Huh 7.5 Intracellular RNA

**Figure 2.37. Next Generation Sequencing on Enriched Sample Produces an Order of Magnitude more Zika Reads than Cellular RNA Reads.** 400 nanograms of the two round oligo capture pool in figure 2.35 was prepared as a CirSeq library and sequenced as part of an indexed pool on an Illumina MiSeq Sequencer. The CirSeq reads (grey, 4,435,129 reads) were then aligned to Zika virus Puerto Rico strain genome (orange, 3,486,266 reads), and to ribosomal RNAs (red, 386,922 reads); 45s rRNA (which contains 18s, 28s and 5.8s rRNAs within) and 5s rRNA. The remainder of the total reads (561,941 reads) is shown in blue; the identity of these reads is unknown and is present in numbers higher than usual, and are perhaps artifacts due to the low amount of RNA input into the CirSeq library protocol. At 400 ng this was by far
the smallest library we have attempted, since much of the sample was removed at various steps for other assays, and is under Ashley Acevedo's recommended minimum of 1 ug and optimum of 5 ug. This issue can likely be mitigated in further experiments by running two oligo captures in parallel on the intracellular RNA and combining the elutions to attain well over a microgram of essentially pure viral RNA. Nonetheless, ~80% of total reads is only slightly lower than our best libraries (eg. Figure 2.29), can provide plenty of target data, and is significantly higher than most RNAseq datasets performed directly on Zika infected intracellular RNA in which viral RNA is around only 1%.

We then performed preliminary analysis on the CirSeq dataset. Zika viral RNA within Huh 7.5 cells that were infected by passaged virions is a tremendously diverse population (Figure 2.38 A). Many variants exist at high frequencies, and the consequences and utility of this diversity is described further in chapters 1, 3 and 5. Additionally, an even coverage across the whole genome indicates that Zika viral RNA is not undergoing degradation during our oligo capture protocol, and our fears were unwarranted (Figure 2.38 B). This represents the first time the diversity of a Zika virus population, beyond the noise threshold of traditional NGS, has been revealed. Flaviviruses such as Zika virus do not produce enough virions, from which relatively pure viral RNA can be extracted, to meet the minimum RNA threshold required by the CirSeq procedure. Consequently, an oligo capture protocol presents the only way to perform a CirSeq run in which the majority of reads align to the viral RNA, enhancing data quality and minimizing sequencing cost. All such analyses with minimum RNA thresholds beyond what is easily available from enriched
sources, such as concentrated virions, will benefit from this oligo capture procedure. Additionally, though we have not attempted capture of non-viral RNA yet, there is no reason that this protocol cannot be equally effective against any RNA species.

A)

![Graph A](image)

B)

![Graph B](image)

**Figure 2.38. Oligo Capture Enrichment Enables CirSeq Revealing That Zika Viral RNA is a Diverse Population.** Subsequent analysis was performed on the Zika-aligned CirSeq reads described in figure 2.37. A) Frequency of variants (y-axis) in the diverse Zika virus RNA population are mapped to nucleotide position on the Zika virus Puerto Rico strain genome (x-axis). Each position contains up to three variants. Variants are colored based
on change from consensus base to adenosine (blue), guanosine (green), cytosine (orange), or uracil (purple). The consensus sequence masks many variants that are present in very high frequencies in the population, even approaching 50% (the threshold upon which that nucleotide is set as the consensus). B) A coverage plot shows even coverage throughout the genome and a lack of distinct peaks where ZPR oligos hybridize, indicating the lack of RNA degradation during the capture protocol. The number of reads that aligned (y-axis) at each nucleotide position of the genome (x-axis) is plotted. The dotted blue line indicates the average coverage per nucleotide of 25,531 reads. Since the primary purpose of this run was to ascertain genome quality and viral RNA reads ratio, around a tenth of the number of sequencing reads as a typical dataset used for genetics analysis (eg. figure 2.29) were acquired. A 1-log variance in coverage is expected [25].

The CirSeq procedure, briefly; CirSeq of RNA was performed as previously described [2, 25]. Fragmented RNA was circularized and converted to DNA by rolling-circle reverse transcription, yielding tandemly repeated cDNAs. The major addition to the protocol listed in Ashley Acevedo's papers was the replacement of the manual size selection step, of these cDNA's on PAGE gels, with using the Sage Science Bluepippin machine, which performs automated size selection. Size selection of 400 to 600 bp was performed on single-use BluePippin 1.5% agarose dye-free cassettes. This excellent suggestion by Dr. Christopher Mason eliminated the persistent problem of amplified adapter primer dimers. Viral RNA samples were thus prepared and 300-cycle single-end sequencing was performed on an Illumina MiSeq or HiSeq 2500 (at the Rockefeller University Genomics Core and Weill Cornell Epigenomics Core, respectively). Tandem repeat reads were converted to consensus sequences, filtering out random errors generated during library
preparation and sequencing, and then mapped to the appropriate consensus genome sequence. Further descriptions on bioinformatic pipelines are provided in chapter 4.

2.20. PEG Based Virion Enrichment: A Lesson in Manufacturer Product Variability.

Concentration of virions using polyethylene glycol is not new, and is the most convenient method to get very large amounts of virus within a given volume [26–28]. Increased virus production requires more cells, which in turn requires more growth media, representing a threshold of virions per milliliter of media that is hard to breach without subsequent concentration (Figure 2.39).

![Figure 2.39. Infection Growth Curve of Sindbis Virus Toto1101 in Huh 7.5 Cells. Sindbis virus Toto1101 strain was used to infect a 500 cm² plate with a](image-url)

<table>
<thead>
<tr>
<th>Hour p.i.</th>
<th># 500cm² plates For 5 µg RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1,197,605</td>
</tr>
<tr>
<td>3</td>
<td>232,288</td>
</tr>
<tr>
<td>6</td>
<td>2,105</td>
</tr>
<tr>
<td>9</td>
<td>299</td>
</tr>
<tr>
<td>12</td>
<td>53</td>
</tr>
<tr>
<td>15</td>
<td>12.5</td>
</tr>
<tr>
<td>18</td>
<td>9.4</td>
</tr>
<tr>
<td>21</td>
<td>5.9</td>
</tr>
<tr>
<td>24</td>
<td>4.6</td>
</tr>
</tbody>
</table>
monolayer of Huh 7.5 cells at an MOI of 0.1. A small aliquot of media containing virions output from cells was removed every 3 hours for 24 hours, viral RNA was extracted and assayed via qPCR. The number of 500 cm² plate infections required to produce the 5 micrograms of viral RNA requisite for CirSeq is calculated and shown. Even after 24 hours of infection and past the exponential growth phase with a highly replicating virus and a highly permissive cell line, about five 500 cm² plate infections are required to acquire enough virions to extract 5 micrograms of viral RNA, representing 500 ml's of media volume. An effective virion concentration procedure is required to minimize the volume which can be put through RNA extraction protocols. Scaling up phenol or column-based extractions to 500 ml's of volume is not practical, to say the least. Yes, I'd stayed up for 24 hours straight with intermittent naps in a lab office couch to collect all the timepoints for this experiment, only had to do that once though. Error bars represent SEM of three technical replicates, some error bars are too small to be shown.

PEG precipitation scales extremely well, and allows for thousand-fold or more concentration of liters of media volume. The concentrated virus can then be used to infect viruses at very high MOI's, or for viral RNA extraction. It is far easier and cheaper to do a Trizol-based RNA extraction from a set number of virions in one milliliter than in one liter. To extract the optimal 5 ug of viral RNA for the CirSeq protocol, we had to extract RNA from all virions produced from multiple 500 cm² dishes full of cell monolayers, each of which contain 100 ml's of cell culture media. PEG precipitation allows most of these virions to be concentrated within a ml, such that RNA extraction can be done in microcentrifuge tubes and using RNA columns.

The story here will continue the theme of having things work fine until they
suddenly stop working completely for reasons unknown at the time. For my concentrations I was using Charlie's 1985 bottle of PEG 6000, made in a factory in Cherry Hill, New Jersey and supplied by a company called EM Sciences (Figure 2.40 A). This stock recovered ~99% of plaque-forming units (pfu's) in the PEG pellet, and less than 2% of pfu's maximum remained in the supernatant. This was the only PEG I had used thus far, and such efficacy was my expectation as long as the PEG was of the 6000 molecular weight variety, which was the only important parameter I was told to look for. Due to the large amounts of volume I was concentrating, this bottle ran out and I switched to another bottle of PEG 6000 in our lab's stock room (Figure 2.40 B), purchased in 2014, and thought nothing of it. Suddenly, odd things started to happen. Only 35% to 50% of pfu's were being recovered from the PEG pellet. Additionally, around this time, the first fragmentation step in the CirSeq library protocol stopped working. Turned out that when Trizol and ethanol precipitation was used on the PEG precipitated virions, there was something in this PEG that was not in the 1985 PEG that interfered with magnesium ion and heating based RNA fragmentation. Note that the labels of both bottles only list one ingredient; Polyethylene Glycol 6,000 (Figure 2.40 A & B). Whatever this was, it was purified away by subsequently running the RNA through a Qiagen RNeasy column, and this allowed the fragmentation reaction to work properly again (Figure 2.33 C).
**Figure 2.40. Polyethylene Glycol Brands Behave Differently During Virion Concentration.** Various bottles of PEG that was available in the lab and that were purchased for testing. A) Charlie’s original 1985 stock of chemical synthesis grade PEG 6000 that captured ~99% of Sindbis virions, which is unmatched by all newer ones tested (EM Sciences). Unfortunately, this stock ran out. B) Our lab stockroom PEG 6000 that captured 35-50% of virions, and contained something that blocked RNA fragmentation if a column purification step was not implemented (Alfa Aesar). C) A crystallography grade PEG 6000 used by structural biology lab members that only captured less than 10% of Sindbis infectious virions, and drastically reduced its specific infectivity (Fluka).
Chemicals). A variety of other PEG 6000's available in the lab performed as poorly and are not shown. D) A molecular biology grade PEG 8000 available in the lab that was fairly effective at capturing Sindbis virions, though most literature advises PEG 6000 for this purpose (Sigma-Aldrich). E) An acquired pharmaceutical excipient grade (i.e., an inactive drug carrier appropriate for injection into humans) PEG 6000 that was quite effective for capturing Sindbis virions and was a fraction of the price per kg as the others (EMD Millipore). F) An acquired molecular biology grade PEG 6000 that was also quite effective at capturing Sindbis virions (EMD Millipore). Do not assume that all PEG 6000's are created alike. Further descriptions are provided in the text.

I thought this was too ridiculous to put up with, and I borrowed some PEG 6000 from a lab member who uses it to make crystals for his structural protein studies (Figure 2.40 C). This was stated to be highly pure, and I figured it must be if it's used to make protein crystals. However, things got even weirder. The cell culture growth media that contains the virions, which is usually red, turns yellow upon the addition of PEG solution, likely due to the phenol red indicator detecting a change in pH. There was no change in color when this crystal grade PEG 6000 was added. Plaque assays showed that recovery was even worse than the 2014 lab stock, and less than 10% of pfu's were recovered in the pellet as shown by plaque assay. However, more than 10% of the virion RNA was being recovered as per qPCR, indicating an effect on specific infectivity (Figure 2.41, Fluka 6k bar). Quantification revealed that specific infectivity went from 1:3 before PEG precipitation to 1:85 after PEG precipitation. Thus, this crystal grade PEG 6000 was, at the very least, modifying some part of the viral glycoproteins, and consequently inhibiting its
ability to form plaques. This was not a good sign, as it is preferable if the concentrating agent did not destroy various parts of the virus prior to downstream uses and analyses. Charlie then recalled that prior to selecting his bottle of PEG 6000 from 1985 and buying enough of it to last 30+ years until I showed up, he had run into similar issues and had to test various manufacturers to find the best one, and suggested I do the same.

I continued to borrow and try different bottles of PEG that was available within the lab, as many lab members had personal stocks for a variety of purposes. Most of them similarly had some problem or another, but one bottle of a different molecular weight, PEG 8000 (Figure 2.40 D), worked okay (Figure 2.41, Sigma 8k bar). The manufacturer of Charlie's 1985 bottle, EM Sciences, had since merged into the current company EMD Millipore, and I contacted them to ask if the factory in Cherry Hill still manufactured PEG 6000 that we could purchase. Unfortunately, I was told that the factory has since shut down, but that several departments within EMD Millipore offered various versions of PEG 6000, made in different factories around the world for different purposes. Of these, in 2016, their molecular biology grade (Figure 2.40 F, EMD Cat #528877) cost $149 per kilogram, and their European pharmaceutical excipient grade (Figure 2.40 E, EMD Cat #817007) cost about the same for 5 kilograms, for $30.60 per kilogram. I explained our protocol and asked the company for a recommendation, and a sales manager replied; "Since the
EMPROVE (European pharmaceutical excipient) version would be the highest quality and have the most regulatory documentation supporting it that would be the one I would recommend”. Surprisingly, their recommendation for the purest and highest quality version was 1/5th the cost as the others. I asked the company what the difference between that and the molecular biology grade was, which was presumably the basis for the cost difference. A senior scientist at the company replied, "Molecular Biology Grade simply means the material is undetectable for proteases, RNAses and DNases, which is less of an issue unless you are working on proteins on the surface of the viruses". A panel of these simple tests, perhaps in addition to economies of scale, result in a 5x cost multiple! Additionally, we presumed that it was highly unlikely that pharmaceutical excipient grade products which are meant to be injected into people would contain such contaminants. We ordered and tested both (Figure 2.41), and while not as great as the 1985 stock, they were both fairly effective. Thus, both EMD Millipore’s pharma and biology grade PEG 6000's are fine for use, and we chose to purchase more of the pharma grade product due to cost.
Figure 2.41. Polyethylene Glycol Brands Vary in Recovery of Arbovirus Virions. Huh 7.5 cells monolayers were infected with Sindbis virus Toto1101 strain and supernatant containing virions was harvested from each at 21 hrs p.i. These supes were combined and split into 100 ml fractions representing the output of one 500 cm² plate. PEG precipitation using the indicated brands was conducted as per the protocol in the next section. Viral RNA was then extracted from the resuspensions and assayed via Sindbis qPCR, and y-axis represents the amount of viral RNA in nanograms acquired from 100 ml fractions. The numbers indicate the percentage of virions in the original supe that was present in the PEG pellet resuspensions. Plaque assays (not shown) for the Fluka 6k sample was far lower than 32% of input indicating a detrimental effect on specific infectivity, whereas the other samples did not see a change in specific infectivity. Each respective bottle is shown in Figure 2.40. Error bars represent SEM of 2 technical replicates.

Many papers published a long time ago explicitly state the manufacturer of
important reagents used, and sometimes even the same reagent is split in
tables and they are listed individually with their respective manufacturer (eg.
**Table 1.1**). That is rarely done in recent decades, and I was not expecting
manufacturer variability of the same stated chemical product to be a major
factor, which led to many failed experiments and some worry. The paper that
our lab's protocol is based on mentioned a variety of parameters, such as the
ionic strength of the resuspension solution, which they found to be important,
but the manufacturer of the PEG they used is not listed, nor is there any
mention of manufacturer variability [26]. I know better now. I thought to share
this lesson because manufacturer was unexpectedly the most important
parameter for us, and is not mentioned in recent literature, whereas
parameters which did not matter for us, such as precise molecular weight and
molecular biology grade, are often described. This practical section hopefully
serves to prevent the reader from having to stumble through the same issues
for months as I did when ordering or switching PEG bottles for virion
concentration.

Kits such as ABCAM PEG kit are used frequently in the field for both virus
studies [29–32] and for lentivirus concentration prior to transduction in other
studies [33]. However, ordering a bottle of PEG crystals and following the
protocol in the next section is far more scalable, and orders of magnitude
cheaper, for large volumes of infection supernatant (**Table 2.1**). The best
practice is to acquire various bottles and test their virion pelleting ability via plaque assay or similar method, and then buy a large amount of the same lot to last a number of years before potentially having to do it again. Immediately upon acquiring the results in Figure 2.41, we purchased all remaining PEG 6000 from the same manufacture lot that EMD Millipore had available, which was around 20 kilograms. This amount can concentrate virions from 200 liters of media, which should last us a while, and was acquired for the same price as one and a half ABCAM PEG kits which only concentrate a half liter of media each. I highly recommend trying PEG manufactured by EMD Millipore as a first pass, as all three lots acquired from them spanning over 30 years - Charlie’s 1985 stock, the current molecular biology grade and the European pharmaceutical grade - gave the best results for virion concentration out of all manufacturers.


This protocol is a slightly modified version of the protocol taught to me by Dr. Melody Li, who is now at the University of California, Los Angeles. Her protocol was largely based on one from the Strauss laboratory [26]. Using the following protocol, we routinely successfully concentrate arbovirus virions from 200-500 ml's of cell culture supernatant into 200-500 ul of PEG resuspension solution.
Materials:

Nalgene transparent polycarbonate centrifuge bottles, Style 3140, Capacity 250 mL (rated for RCF 27,500 xg, Sigma Aldrich Cat # B1908-4EA)

Note: It is highly suggested that one purchase fully transparent bottles that is rated above the required RCF. Our common lab stock bottles did not have ratings indicated, and either due to age or original lower rating, I had the unfortunate experience of having the bottles break and cave in during centrifugation, which led to the loss of virus produced upon much effort, and also required a long cleaning procedure. Fully transparent bottles make visualizing the PEG pellet far easier, and I've noticed that common bottles stocked in many labs are translucent. Additionally, it is perhaps best not to use the same bottles used for maxipreps, as even upon autoclaving many bottles still retain small streaks of bacterial pellets. The product indicated above is the exact model we bought and used for these experiments and it works well for PEG precipitation, but other products may also suffice. A centrifuge and rotor attachment capable of spinning the bottles are the indicated RCF’s is also required, most laboratories should already have access to them.

Reagents:

Polyethylene Glycol 6000 molecular weight, European pharmaceutical excipient grade (EMD Millipore Product #817007, shown in Figure 2.39 E)

Sodium Chloride, 5M, molecular biology grade (Promega Product # V4221)
Tris pH 7.5, 1M (Invitrogen Product # 15567027)
EDTA pH 8, 0.5M, Fluka Bioultra Grade (Sigma Adlrich Product # 03690)
Distilled deonized water (ddH2O) was purified using a Millipore Milli-Q
Synthesis system and Q-gard 1 cartridges (Millipore Sigma #QGARD00R1).

Solutions (And Preparation Using Reagents Listed Above):
1. PEG Solution: 40% PEG in 2M NaCl.
Preparation of 500 ml's: To 200 grams of PEG crystals, add 200 ml's of 5M NaCl and 200 ml's of ddH2O. Mix well and let the solution sit in a 56°C water bath for 3 hours, at which point there should be minimal visible crystals remaining. Then dilute the volume to 500 ml using ddH2O. Add a magnetic stir bar and stir the solution for 2 hours on a heated magnetic plate set to 60°C. Finally, filter the solution using a 0.22 um filter in a cell culture hood.
2. Resuspension Solution: 50 mM Tris pH 7.5, 1 mM EDTA, 0.2M NaCl.

Key:
+ I Add Reagent
- I Remove Supernatant
* I Centrifuge or Shake Step
> I Transfer to Indicated Tube
= I Incubation Step
L I Pelleting Virions Using PEG

To cell culture supernatant;

L.01 | + 25% Volume of 40% PEG in 2M NaCl Solution
L.02 | * Swirl and Mix well
L.03 | = Overnight @ 4°C incubation or 1 hour @ ice incubation
L.04 | * 15 min 4°C spin @ 20,000 g
L.05 | - Pour out supernatant (save 5 ml for next step)
L.06 | + Resuspend pellet in ~5 ml Supernatant and respin
L.07 | * 15 min 4°C spin @ 20,000 g
L.08 | - Remove supernatant, avoid disturbing pellet

M I Resuspension and Clarification of PEG Pellet

M.01 | + Desired volume of PEG resuspension solution. 1/500 or 1/1000 of original volume is common.
M.02 | * Vigorously rinse pellet for several minutes until fully dissolved.
M.03 | > Resuspension -> 1.5 ml eppendorf tube
M.04 | * 10 min RT spin @ 400 g (clarification)
M.05 | > Supernatant -> New 1.5 ml eppendorf tube

This resuspension can now be used for infections, or RNA can be extracted from the virions in this resuspension using the same steps as in the cellular RNA extraction protocol near the beginning of this chapter.
2.22. Chapter II References.


8. 5’ Biotin-TEG - Integrated DNA Technologies. https://www.idtdna.com/site/Catalog/Modifications/Product/8859


Chapter III: Accurate Sequencing of RNA Virus Evolution Enables High Throughput Viral Genetics

3.1. Introduction.

RNA viruses exist as genetically diverse populations due to high error rates of the viral RNA-dependent RNA polymerase. Each replication cycle introduces thousands of mutants into each population, which are then subject to selection and determine the adaptive capacity of the virus. Accurate profiling of these viral populations and monitoring the dynamics of viral variants through various infection conditions can reveal how the population adapts to various bottlenecks and also potentially allow for high-throughput viral genetics.

Next generation sequencing provides the sequencing depth to profile these viral populations. However, most RNA virus mutants are generated in frequencies below the error rates of current next generation sequencing platforms, resulting in PCR, library preparation and sequencing errors being indistinguishable from these mutants. CirSeq allows for identification of mutations which are too infrequent to be detected by conventional RNAseq [1]. This method relies on generation and sequencing of tandem cDNA repeats from the same RNA molecule. These sequences are then bioinformatically parsed to correct the sequencing errors and generate a consensus sequence (Figure 3.1). This allows for accurate detection of low-frequency variants and
Figure 3.1 The CirSeq Method. CirSeq improves next-generation sequencing accuracy by using rolling reverse transcription to produce biologically tandem repeats which are then sequenced and bioinformatically parsed to produce error-corrected accurate sequencing reads. From [1].

Sindbis virus is one of the best-studied members of the alphavirus genus. The viral mechanisms described in the introduction chapter were primarily elucidated using mutagenesis studies on identified sites of interest.

Historically, the mutations that conferred temperature sensitivity or other desired phenotypes were chosen for detailed analysis. An unbiased genome-wide analysis of mutations that arise during virus propagation in mammalian
cells has not been performed. In this study we used CirSeq to examine how the virus population adapts to various bottlenecks encountered during infection of Human hepatoma 7.5 cells. We then utilized the dynamics of single nucleotide variants to expand virus genetics beyond the limits of the previous methods and allow for genome-wide analyses in a high-throughput manner.

3.2. *in vitro* Transcribed RNA Used to Generate Virus is a Mixture of Mutants.

To understand how the viral population evolves in various infection conditions, it is critical to understand if the viral RNA diversity is generated *de novo* during infection of cell culture or if it pre-exists in the *in vitro* transcript population. As the complexity of viral RNA populations in a biological sample is beyond the capabilities of conventional RNAseq, we set out to use CirSeq to obtain a more accurate picture of Sindbis viral RNA complexity [1]. We first *in vitro* transcribed Sindbis Toto1101 RNA using the SP6 RNA polymerase enzyme and typical reaction conditions which include sub-saturating amounts of rGTP to promote cap analog incorporation. Next we employed CirSeq to examine the heterogeneity of the *in vitro* transcribed RNA population. The *in vitro* transcription generated a mutant swarm of RNA, with a mutation rate of $2.7 \times 10^{-4}$ (*Figure 3.2*). The specific mutation rates (rate of conversion from one particular nucleotide to another) vary over a 30-fold range (*Figure 3.3*).
Figure 3.2. SP6 *in vitro* transcribed RNA is a diverse population. Frequency of variants (y-axis) generated in SP6 *in vitro* transcription reaction using sub-saturating rGTP, detected using CirSeq, are mapped to nucleotide position on Sindbis Toto1101 genome (x-axis). Each position contains up to three variants; the frequencies of substitutions from the consensus base to the other three nucleotides. Dot size is scaled with frequency for visualization. The dotted line represents the detection limit of traditional next-generation sequencing.
Figure 3.3. Specific Mutation Rates of SP6 *in vitro* transcription reaction. Determination of specific mutation rate estimates (y-axis) for each type of nucleotide mutation (x-axis) generated by the SP6 enzyme in the above reaction. Selection is not a factor in the *in vitro* reaction, and rates are determined using maximum likelihood estimates as detailed in materials and methods. Error bars represent SEM.

The rate of mutations from G to other nucleotides, particularly G → A mutations, were high relative to other specific mutations. We wondered if this was due to sub-saturating rGTP concentration in the SP6 *in vitro* reaction. To test this, we transcribed Sindbis Toto1101 RNA using equimolar rNTPs while all other reaction conditions were held constant, and noticed that it doubled the production of RNA relative to the typical reaction conditions using sub-
saturating rGTP (**Figure 3.4**). We then performed CirSeq on the RNA transcribed in equimolar rNTP concentrations and analyzed specific mutation rates. Compared to sub-saturating rGTP conditions, equimolar rNTP’s had reduced G → A mutation rates, whilst most specific mutation rates were relatively unaffected (**Figure 3.5**).

![Sindbis RNA Synthesis by SP6](image)

**Figure 3.4: RNA Synthesis Rates Are Dependent on Reaction Conditions.** RNA production measured using nanodrop in a 30 minute *in vitro* transcription reaction of Sindbis Toto1101 with sub-saturating (2 mM) and equimolar (10 mM) rGTP amounts, with all other reaction conditions being held equal. More than double the amount of RNA is synthesized when the rGTP starvation condition is relieved.
Figure 3.5: The Diversity of SP6 *in vitro* Transcribed RNA is Affected by Reaction Conditions. A comparison of specific mutation rates of the reaction performed in 2mM rGTP (x-axis) and 10 mM rGTP (y-axis), with mutations from G to another nucleotide highlighted in green. Error bars represent SEM.

Collectively, these data demonstrate that SP6 generates a diverse pool of RNA molecules, and the extent of this diversity is at least partly affected by the ribonucleotide concentrations present in the reaction. While rNTP concentrations affect both the RNA synthesis rate and polymerase fidelity, not all specific mutation rates are significantly affected when a particular rNTP is sub-saturating. This complex mixture of viral RNA is what is transfected into cells to initiate infection, and may provide the virus with an adaptive advantage.
during infection.

*Thus, the in vitro transcribed RNA is a diverse population:*

RNA viruses used for study are typically generated as DNA plasmids containing the viral sequence (infectious clones), which are *in vitro* transcribed as RNA and transfected into cells to initiate infection and produce virions. Bacterial DNA polymerases have a low mutation rate due to multiple error correction mechanisms, therefore plasmid amplification in bacteria yields a homogenous population [2]. However, whether *in vitro* transcription of this DNA yields a homogenous RNA population remains unknown. Often *in vitro* transcribed RNA is treated as a homogenous population that mirrors the homogeneity of its DNA template [3]. At times, mutants that modify a viral phenotype early in the infection cycle are thought to have arisen during SP6-driven transcription, albeit without direct measurement [4].

We show via direct measurement using CirSeq that *in vitro* transcribed RNA is a diverse population (Figure 3.2). Our data clearly indicates heterogeneity of viral RNA prior to infection, and contradict with the existing notion that *in vitro* transcription yields homogenous RNA. Our finding is not completely unexpected as several DNA-dependent RNA polymerases, including phage T7 and eukaryotic RNA polymerases, have been shown to possess low fidelity
[5]. Since SP6 is a phage-derived DNA-dependent RNA polymerase, it is not surprising that the fidelity of SP6 falls within the range of other such polymerases [6].

Ribonucleotide concentrations affect the rate of polymerization by SP6, with rNTP starvation conditions slowing down the enzymatic activity [6]. Since SINV RNA has a 5’ cap, a cap analog must be included in the in vitro transcription reaction to obtain functional viral RNA that can infect cells [7, 8]. To promote cap analog incorporation, a typical in vitro transcription reaction uses sub-saturating rGTP conditions [9, 10]. In our experiments, such conditions produced half the amount of RNA as the reaction with equimolar rNTP concentrations and specifically increased G → A mutation rates (Figures 3.4 & 3.5). Such an effect of rNTP starvation on RNA synthesis rates have been shown to be common to many RNA polymerases, including eukaryotic DNA-dependent RNA polymerases [11]. Historically, increased polymerase synthesis activity has been associated with lower fidelity due to decreased polymerase pausing times during mismatch incorporation [12]. Traditionally, due to the difficulty of directly measuring mutation rates, differences in RNA synthesis rates are often used as a proxy for measuring fidelity in the study of various RNA polymerases [11, 13, 14]. As our data of the complexity of SP6-transcribed Sindbis RNA shows, CirSeq can be used to directly measure the mutation frequencies in synthesized RNA populations,
and may benefit such studies of various RNA polymerases. The complex mixture generated by SP6 may also be advantageous for the virus as upon transfection into cells, it might give the virus a head start in terms of evolution and lay the foundation for adaptation during infection.

3.3. Characteristics of Virus Population Dynamics During Electroporation and Passaging.

Next, we introduced the *in vitro* transcribed RNA into Huh 7.5 cells by electroporation and monitored the complexity of the viral RNA in packaged virions in the culture medium at 20 hours post-electroporation (P.E.). We performed CirSeq on the RNA in virions secreted by the electroporated cells (virus stock) and compared it to the input SP6 *in vitro* transcribed RNA. We observed a significantly decreased frequency for 28.3 % of variants, while the frequency of 12.4 % of variants increased significantly, with nonsynonymous mutations experiencing greater selection than synonymous mutations (*Figure 3.6*).
Relative enrichment of statistically significant ($p < 0.05$) synonymous (blue) and nonsynonymous (orange) mutations in the coding region of viruses recovered from electroporated Huh-7.5 cells. The SP6 \textit{in vitro} transcribed RNA served as a denominator for these calculations. Y-axis represents density, where the areas under the synonymous and nonsynonymous density curves total to 1.

To remove any residual input RNA from the culture medium, we extensively washed the electroporated cells at 4.5 hours post-electroporation, which reduced the input RNA to undetectable levels (\textbf{Figure 3.7}).
Figure 3.7. Removal of Input *in vitro* Transcribed RNA in Cell Culture Media. A) qPCR of electroporation media with RNA copies per ml of media (y-axis) per hours post-electroporation (x-axis). High amounts of input *in vitro* transcribed RNA is detected initially, but it becomes undetectable after washing the cells multiple times with a phosphate-buffered saline solution. 4.5 hours was chosen for washing because it gives the electroporated cells time to attach to the bottom of the plate, and is before viral RNA resulting from infection is detectable in the media. B) qPCR of Sindbis Toto1101 *in vitro* transcribed RNA with various amounts of added Huh 7.5 cellular RNA per well shows reliable detection above $10^2$ RNA copies per well, yielding 29 cycles of dynamic range using our qPCR Taqman assay protocol (material and methods). The standard curve is not masked by presence of cellular RNA in the sample. Error bars represent SEM.
A)

RNA Copies in Electroporation Media

RNA Copies / ml

1x10^6
1x10^5
1x10^4
1x10^3
1x10^2
1x10^1
1x10^0

0 3.5 4.5 12 15 18 21 24

Pre-Wash  Post-Wash

Hours P.E.
B)

Toto1101 qPCR Taqman Assay

Detection Limit (Water Sample Average)

Huh 7.5 Cellular RNA
Added Per Well:
- 0 ng
- 250 ng
- 500 ng
- 1000 ng

Sindbis Toto1101 RNA Copies per Well
Next, we wondered how the variant population continues to change during viral passaging in mammalian cells. To test this, we passaged the released virions five times on Huh 7.5 cells and monitored the complexity of the virion RNA using CirSeq after each passage (Figure 3.8). We passaged the virus at a low multiplicity of infection of 0.1 (MOI; ratio of virus particles capable of initiating infection to target cells). The total variant frequency remained around $2 \times 10^{-4}$ (Figure 3.9), as previously determined by classical biochemical assays [15].

Figure 3.8. Experiment Schematic. SP6 in vitro transcribed RNA was electroporated into Huh 7.5 cells, and virions were harvested at 20h p.e. for CirSeq library preparation and sequencing. The electroporated virions were then serially passaged at low MOI for 5 passages and similarly prepared and sequenced.
We then examined the frequency of mutations to stop codons and other potentially deleterious mutations, such as those present in the viral polymerase active site, which should be impacted by selection. We found an immediate post-electroporation reduction in the frequency of polymerase active site mutations, with a 7-fold mean frequency reduction between the in vitro and passage 5.
vitro transcribed population and the first passage, which likely represents selection effects (Figure 3.10 A). The frequency of polymerase active site mutations reaches a relative equilibrium for the remaining passages. However, the stop codon mutants only reached a relative equilibrium by the second passage (Figure 3.10 B). Since stop codon mutants and viral polymerase active site mutants are not expected to be viable, this equilibrium likely represents the frequency at which such variants are newly generated by the low-fidelity viral polymerase during each passage [1].
Figure 3.10. Mean Mutation Frequency in Polymerase Active Site and Mutations to Stop Codons. A) Mean frequency of nonsynonymous single nucleotide mutations in the viral RNA-dependent RNA polymerase active site. B) Mean frequency of single nucleotide mutations that yield a stop codon. Error bars in B and C represent SEM.
A)

GDD Domain Mutations
(Polymerase Active Site)

Mean Frequency (n = 22)

Virus Passage

in vitro, electro, passage 1, passage 2, passage 3, passage 4, passage 5
Stop Codon Mutations

Mean Frequency (n = 1,430)

Virus Passage

in vitro, electro, passage 1, passage 2, passage 3, passage 4, passage 5
The high-throughput nature of these data allowed us to similarly assess the relative change in mean frequency between the *in vitro* transcribed population and post-equilibrium passages 2 to 5 for mutations to all amino acids and codons (Figures 3.11 and 3.12).

**Figure 3.11. Tolerability of Amino Acid Substitutions Across the Whole Genome.** The mean relative enrichment, between *in vitro* transcribed RNA and low MOI passages 2 to 5, was calculated for all nonsynonymous single-nucleotide mutations to a particular amino acid (orange) or a stop codon (red). In general, mutations to a stop codon are the most detrimental as we expected, followed by mutations to a proline and asparagine. Error bars represent SEM.
Figure 3.12. Relative Enrichment of All Possible Single Nucleotide Substitutions. Heatmap of relative enrichment, between in vitro transcribed RNA and low MOI passages 2 to 5, for all possible codon changes arising from single-nucleotide mutations. X-axis represents codons in the consensus Sindbis Toto1101 genome, and Y-axis represents the mutant codons, in alphabetical order of the corresponding amino acids. Each codon can mutate to 9 possible other codons as a result of single-nucleotide mutations, and boxes that represent mutations that are not possible are shaded in gray. Log enrichment values are color coded in a gradient from blue (adaptive) to white (neutral) to red (detrimental).

We wondered how the mutation frequency was distributed between
synonymous, nonsynonymous and untranslated mutations. The 11,703 nucleotides in Sindbis Toto1101 before the PolyA tract allows for 35,109 potential total single nucleotide mutations. If we assume an equal mutation frequency for every mutant, the 26,033 nonsynonymous mutations, 7,789 synonymous mutations, and 1,287 non-coding region mutations would contribute 74.1%, 22.2% and 3.6% of the total mutation frequency respectively (Figure 3.13, equal dist. column). The initial in vitro transcript population has a higher proportion of mutation frequency in untranslated regions. During electroporation and passaging, the proportion of mutation frequency in untranslated regions increased initially, with a corresponding decrease in nonsynonymous mutations, and with synonymous mutations relatively similar to the starting population (Figure 3.13).
Figure 3.13. Proportion of Mutation Frequency In Coding and Noncoding Regions of the Viral Genome. Proportion of total mutation frequency per dataset of nonsynonymous codon variants (orange), synonymous codon variants (blue) and untranslated region variants (green). Equal dist represents the hypothetical percentage contribution of each type of mutation if all individual
single nucleotide mutations had an equal mutation rate.

We then analyzed the viral coding region and examined the distribution of mutations at each nucleotide position within a codon. As mutations at the wobble position are mostly silent and therefore generally more tolerable, we expected that the cell-derived virus might have more mutations at this position as compared to the first and second nucleotide position. In contrast, *in vitro* transcribed RNA should have a distribution of mutations across the whole genome that is not biased by codon context. This indeed turned out to be the case in early passages; the mutation frequency at all three positions was significantly less in the passaged virus compared to the *in vitro* transcribed RNA (Figure 3.14). However, in later passages the overall nonsynonymous mutation frequency increased due to several adaptive nonsynonymous mutants, predominantly in the second codon position.
Collectively, these data suggest that the composition of virus population changes upon passaging when compared to the parental in vitro transcribed RNA. The frequency of most mutations present in the starting population are significantly de-enriched upon electroporation and passaging. Overall,
nonsynonymous mutants are more detrimental than synonymous mutants, resulting in a higher proportion of mutation frequency in the third wobble position of codons. Analysis of mutants expected to be very detrimental shows a sharp initial decrease in frequency when faced with selection during electroporation and passaging, followed by a relative equilibrium where the mutation frequency is likely due to errors newly generated by the low fidelity viral polymerase.

Thus, electroporation and low multiplicity passaging changes the viral mutational distribution:

While the diversity of the *in vitro* transcribed RNA population itself is underappreciated, what happens to the diversity of this RNA population upon electroporation into cells is unknown. The population encounters the bottleneck of selection; though entry is facilitated, the genomes must complete the remaining steps of the life cycle to establish infection. As many of the viral RNA variants generated during *in vitro* transcription are expected to be defective in replication, only a subset of the electroporated RNA should be able to replicate.

Indeed, the mutant population is dynamic during electroporation. We observed that the frequency of the majority of single-nucleotide variants were either
significantly de-enriched or remained unchanged, with a small percentage of variants enriched upon selection. Nonsynonymous variants experienced greater overall de-enrichment than synonymous variants. This indicates that most variants that are generated by the SP6 polymerase are either detrimental or neutral, with a few being adaptive, and nonsynonymous mutations on average being more detrimental than synonymous mutations (Figure 3.6).

Passaging the virus further reveals the population dynamics that occur during experimental cell culture. Due to lack of selection during in vitro transcription, mutations are distributed in accordance with the SP6 enzyme’s specific mutation rates at sites that are heavily detrimental to the virus during infection, such as premature stop codons and enzyme active sites. Upon electroporation, many of these variants experience negative selection due to their inability to complete replication steps (Figure 3.10 A, B). There is a large ratio of in vitro transcribed RNA per cell during electroporation and multiple viral RNAs likely enter each cell, allowing complementation between genomes. These variants initially experience further negative selection during passaging in a low MOI setting where most cells only receive one genome capable of infection, as the viral RNA must now be capable of every infection step including entry. Variation in these highly detrimental sites reach a relative equilibrium by the second passage, and since these variants should be incapable of replication this equilibrium likely represents the rate at which the
error-prone viral RNA-dependent RNA polymerase generates these variants \textit{de novo} during replication (\textbf{Figure 3.27}). Thus both the incoming variants in the \textit{in vitro} transcribed population and newly generated errors by the viral polymerase shape the mutational landscape upon which evolution acts.

How these variants are distributed among the various codon positions within translated regions and in untranslated regions before and after experiencing selection has not been demonstrated. The total mutation frequency distributed amongst synonymous codons, nonsynonymous codons and untranslated regions in the SP6 \textit{in vitro} transcribed population is largely similar to the frequencies expected if every single nucleotide mutation had an equal probability of generation, with a slightly higher percentage in untranslated regions, perhaps due to more SP6 polymerase errors in these highly structured regions (\textbf{Figure 3.13}). The nonsynonymous variant frequency in all three codon positions decreases heavily upon electroporation and in the early passages (\textbf{Figure 3.14}). The synonymous variant frequency, primarily localized to the third codon position, experiences a smaller decrease due to lesser selection effects. Consequently, the mutation frequency at synonymous sites and in the third codon position encompasses a larger percentage of the total mutation frequency in these passages. Interestingly, the nonsynonymous variant frequency begins to increase in later passages due to adaptive variants primarily in the second codon position, indicating that the population continues
to evolve and adapt using amino acid substitutions throughout passaging.

3.4. The Presence of Insertions and Deletions in Viral Populations.

Next we characterized the distribution of insertions and deletions (indels) generated by the polymerases in the \textit{in vitro} transcribed population and in the passaged virion populations. In the \textit{in vitro} transcribed RNA population, on average we detected an insertion every third genome copied and a deletion every tenth genome copied by the SP6 polymerase (Figure 3.15). In the low MOI virions, on average we detected an insertion once every 17 genome copies and a deletion every 22 genome copies. Our detection counts are likely an underestimate due to the strict screening criteria we employed to filter out alignment errors. All of the populations consistently contain a higher number of insertions than deletions. The indel frequency in the passaged populations is lower than in the \textit{in vitro} transcribed population, which could be due to a lower indel generation rate by the viral polymerase \textit{in vivo} than the SP6 polymerase \textit{in vitro}, or the presence of negative selection against indels, or a combination.
Figure 3.15. Number of Insertions and Deletions Per Genome Copied. The mean number of insertions and deletions detected per genome copied in each dataset, calculated as the number of individual insertion and deletion events divided by mean whole genome coverage.

Interestingly, even though the number of insertions is consistently higher than the number of deletions in the virion populations, the frequency of inserted bases is consistently lower than the frequency of deleted bases (Figure 3.16). This indicates that deletions in these populations are of significantly longer lengths on average than insertions.
Figure 3.16. The Frequency of Nucleotides Inserted or Deleted. The frequency of bases inserted or deleted by each respective polymerase, calculated as the total number of bases in insertions or deletions divided by total base coverage. Insertions are indicated in blue and deletions in red, with the *in vitro* transcribed RNA population and passaged virion populations on the x-axis.

Indeed, the number of both insertions and deletions is inversely correlated with their length in both the *in vitro* RNA and virion populations (Figure 3.17 A & B). In the virion populations, the mean deletion is larger than the mean insertion, and deletions outnumber insertions for all indel lengths above 1 bp.
Figure 3.17. Size Dependence of Generated Insertions and Deletions. The length dependence of generated insertions and deletions by the SP6 polymerase (A) and the viral polymerase (B). The length of the InDels are on the x-axis, and the number of detected indels of that length per genome copied is on the y-axis. Insertions are in blue and deletions in red, and in (B) indels that keep the reading frame are shaded darker than indels that cause frameshifts.
B)

![Graph showing InDels per genome copy vs. length of InDel (bp). The x-axis represents the length of InDel (bp), and the y-axis represents InDels per genome copy. Two bars are shown: one for deletions (pink) and one for insertions (blue). The graph illustrates the distribution of InDels across different lengths.](image-url)
Indels of a particular length are always present in the population in a higher frequency than indels of longer lengths in the \textit{in vitro} transcribed RNA population (\textbf{Figure 3.17 A}). This is not the case in the passaged virion populations, as indels of lengths that preserve the reading frame are sometimes present in the population at higher frequencies than indels of slightly shorter lengths that cause frameshift mutations, indicating selection effects (\textbf{Figure 3.17 B}). Indel lengths mapped to Sindbis genome position reveals a distribution of short indels throughout the genome, but also certain 'hotspots' that contain indels of multiple lengths in the same position (\textbf{Figure 3.18 A, B}). These data indicate that the presence of longer indels at a particular genome position in a population is often accompanied by the presence of shorter indels occurring in the same location. The RNA fragment lengths used to generate the CirSeq library preps preclude detection of indels longer than 80-100 bp, and the presence of significant numbers of larger indels in the population that we cannot detect is a remote possibility.
Figure 3.18. Lengths of Insertions and Deletions Across Viral Genome. The size distribution of insertions (A) and deletions (B) by genome position in Low MOI virion passages 5. The genome position in the Sindbis Toto1101 genome is on the x-axis, and all sizes in base pairs of detected InDels at that position on y-axis. Each point represents the detection of at least one InDel of that particular length at that position, and frequency is not plotted. Many InDels of varying sizes occur at certain genome positions.

Thus, a significant number of insertions and deletions are generated by SP6 and viral polymerases:
Polymerases are also known to make insertion and deletion errors during transcription, and this observation was used to establish the triplet code for protein translation [16]. However, their presence in an RNA virus population has not been thoroughly studied due to the difficulty of differentiating true indels from reverse transcription, PCR, base-calling, and alignment errors. Since CirSeq allows us to overcome these difficulties, we were able to characterize the distribution of indels generated by the SP6 polymerase in vitro and the viral polymerase in vivo for the first time. Interestingly, the number of both insertions and deletions is inversely correlated with their length (Figure 3.17 A, B). Short indels that cause frameshift mutations form the majority of such errors in the population, and since the majority of these should be incapable of replication, most indels seem to be generated de novo by the polymerase at each passage. Whereas short indels are produced throughout the genome, there are certain positions that experience 'stutters' where the polymerases tend to be more error prone and produce many insertions or deletions of multiple lengths at same genome position. Analysis of flanking regions will be of interest for further study to better understand the nucleotide contexts that contribute to this phenomenon.
3.5. Viral Genetics: Congruence of High Throughput Population Dynamics and Previous Studies.

We next sought to determine if the evolution of the population upon infection and passaging can be used to analyze single coding sites on the genome. In the SINV field, several sites that are evolutionarily conserved among alphaviruses have been targeted for mutagenesis studies, where multiple single-site amino acid mutants were constructed and assayed for relative viability. Of these, two of the best studied sites include the penultimate glycine in the cleavage site between nsP1 and nsP2, and the opal stop codon between nsP3 and nsP4 [17].

As described above, the mutation rates of stop codon mutants and the polymerase GDD domain stabilize upon passage 2, and thus we calculated the relative enrichment between the frequency of mutants in the *in vitro* transcribed population and in low MOI passages 2 to 5. Our comparison was based on an average read depth of $5 \times 10^5$ and $1.2 \times 10^6$ reads per nucleotide position for the *in vitro* transcribed RNA and low MOI passages 2 to 5 respectively (Figure 3.19).
**Figure 3.19. CirSeq Sequencing Depth.** Sequencing Coverage Density Plots for (A) *in vitro* transcribed RNA and (B) Low MOI virion passages 2 to 5. X-axis values represent of total reads at each nucleotide of the Sindbis genome, and y-axis are calculated density values. Blue bars represent binned total reads. Orange line is a parametric density and green line is a non-parametric density of total counts per nucleotide position. Legend shows the mode of each type of calculation, with parametric mode equaling the average counts per nucleotide position.
In the P1/2 cleavage site (nsP1-539-G), our data shows that a synonymous mutation to a glycine is tolerated, and a mutation to a stop codon is not tolerated (Figure 3.20 A). Of the nonsynonymous mutations, alanine shows no significant difference, whereas valine and glutamate at this position are detrimental. These data mirror previous observations on the relative viability of nonsynonymous mutants constructed as infectious clones [18]. Serine and threonine are detrimental at both alanine positions that are immediately adjacent to the cleavage site (nsP1-540-A and nsP2-1-A), which are not fully conserved among alphaviruses and consequently were not previously studied (Figure 3.20 B, C).
Figure 3.20. Viral Infection Population Dynamics Reveals Amino Acid Substitution Tolerances at P1/2 Cleavage Site. (A) Amino acid mutants of the penultimate glycine in the P1/2 cleavage site, which is highly conserved among alphaviruses and consequently well-studied. (C, D) Amino acid mutants of the sites immediately 5' (C, nsP1-540-A) and 3' (D, nsP2-1-A) to the P1/2 cleavage site, which are both alanines in Sindbis Toto1101. These sites are not fully conserved among alphaviruses and consequently were not targeted in previous studies. X-axis represents amino acid substitutions and y-axis represents log relative enrichment of mutation frequency in the *in vitro* transcribed population compared to low MOI passages 2 to 5. Bars are color coded as follows - blue = synonymous mutants, orange = nonsynonymous mutants, and red = stop codon mutants. P-values and confidence intervals calculated using Fisher exact test. Error bars represent 95% confidence intervals, and p-values are summarized as follows: n.s. (P > 0.05), * (P < 0.05), ** (P < 0.01), *** (P < 0.001), and **** (P < 0.0001).
nspl Amino Acid 539
Gly at nucleotide positions 1674-1676

Enrichment (log_e)

Amino Acid Mutant

Gly  Ala  Val  Arg  Stp  Glu
B)

nspl1 Amino Acid 540
Ala at nucleotide positions 1677-1679

Enrichment (loge)

Glu  Val  Ala  Pro  Gly  Ser  Thr

Amino Acid Mutant
c)

nsp2 Amino Acid 1
Ala at nucleotide positions 1680-1682

Enrichment (log_e)

<table>
<thead>
<tr>
<th>Amino Acid Mutant</th>
<th>Gly</th>
<th>Ala</th>
<th>Val</th>
<th>Pro</th>
<th>Glu</th>
<th>Ser</th>
<th>Thr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

****  ****
In the opal stop codon site (nsP3-550-Opal), our data shows that nonsynonymous mutations to tryptophan, serine and arginine show no significant enrichment difference, whereas a mutation to an ochre stop codon (UAA) is detrimental during low MOI passaging (Figure 3.21). These data also mirror previous observations of the relative viral titers of these constructed mutants [19].
Figure 3.21. Amino Acid and Alternative Stop Codon Substitution Tolerances at P3/4 Opal Stop Codon. Amino acid mutants of the P3/4 opal stop codon site are on the x-axis. Log relative enrichment of mutation frequency in the *in vitro* transcribed population compared to low MOI passages 2 to 5 is on the y-axis. The stop codon mutant in nsP3-550 represents an opal stop codon to ochre stop codon mutation. Bars are color coded and error bars and
p-value summaries are calculated and plotted as in figure 3.20.

For a particular mutation, a positive enrichment difference will not be detected if the mutation frequency is already high in the *in vitro* transcribed population, and similarly a negative enrichment difference will remain undetected if the starting mutation frequency is already low. Additionally, despite having high sequencing coverage, ~7% and ~3% of mutants in the *in vitro* transcribed RNA and low MOI passages 2 to 5 datasets respectively had undetectable counts precluding enrichment calculations, perhaps representing the most detrimental mutations in the passage series (*Figure 3.22 A, B*). Nonetheless, given the congruence between our data and previous studies on historically well studied sites, we conclude that these variant population dynamics can be harnessed for site-specific genetics.
Figure 3.22. Counts of Viral Variants in CirSeq Datasets. Discrete raw counts of detected mutants in the (A) *in vitro* transcribed RNA and (B) Low MOI virion passages 2 to 5 datasets. X-axis represents the edges of the bins used for mutant counts. Y-axis represents the number of all possible single nucleotide variants (~35,000) that fall within the bins. Blue bars represent detected variants, and the purple bar represents undetected variants. Also listed is the percentage of undetected variants (Percent Zeroes) and the average raw counts of detected variants (Non-Zero Counts Average). An undetected variant in either or both datasets precludes enrichment calculations, due to a divide by zero error. However it does not preclude calculation of Fisher exact test p-values if the variant is undetected in one dataset but detected in the other.
Passage 2 -> 5 Low MOI Supes

Variant Counts

Number of Variants

Percent of Zeros: 3.21 %  |  Non-Zero Counts Average: 91.42
These data on evolution dynamics reflects previous observations on amino acid substitutions and reveals information on previously unstudied sites:

Historically, sites of interest in the Sindbis genome were primarily limited to (i) identified sites that confer temperature sensitivity (ii) sites that are macroevolutionarily conserved among all alphaviruses and (iii) forward genetics on sites that confer a specific phenotype upon screening (e.g. noncytopathic variants). Mutagenesis studies of such identified sites have revealed many aspects of the Sindbis life cycle and interesting viral infection mechanisms [17]. We wondered if the dynamics of the population can be harnessed for viral genetics in a high-throughput manner. We first compared our data to established Sindbis genetics, and found remarkable agreement between our data and previous studies.

The penultimate position of the nonstructural protein cleavage sites and the opal stop codon between nsP3 and nsP4 are highly conserved among alphaviruses and consequently the best studied sites where most single-nucleotide variants have been constructed and assayed [18, 19]. In the penultimate glycine at the P1/2 cleavage site, assays using constructed infectious clones with valine or glutamate substitutions abolished polyprotein processing and severely impaired replication and also produced tiny plaques.
in plaque assays on chicken embryo fibroblast cells [18, 20]. Occasionally large plaques appeared with reversions to glycine or a substitution to alanine, indicating the viability of an alanine mutation at this position [20]. Comparison of the virion populations from the low MOI infection series that have undergone selection with the in vitro transcribed population in our data mirrors relative viability observations of all historically studied mutants on this position (Figure 3.20). Similarly, at the opal stop codon position, substitutions to tryptophan, serine and arginine were tolerated when assayed in CEFs, whereas a mutation to an ochre stop codon grew more slowly in a low MOI setting [19]. Our data also mirrors these viability observations; sense mutations to tryptophan, serine and arginine had no significant difference whereas the ochre stop codon variant was highly significantly de-enriched (Figure 3.21). Given the congruence between our data and previous studies, we conclude that these dynamics of the variant population can be harnessed for site-specific genetics. However, whereas in previous studies each mutant had to be constructed individually and then assayed, our data allows a genome-wide analysis in a high-throughput manner across various infection conditions (Figure 3.23).
Figure 3.23. Highly Accurate Next-Generation Sequencing Allows High Throughput, Whole Genome Virus Genetics. In previous studies, each mutant had to be constructed individually and assayed for viability information. Highly accurate next-generation sequencing provides the ability to genotype (sequence) and phenotype (frequency of sequence in population) every single nucleotide variant at once and compare different reaction conditions across the whole genome. As described in the text, there is a high level of congruence between both methods in sites that were thoroughly investigated previously.

Interestingly, mutations at the above sites influence both the rate of RNA production and the ratio of different viral RNA forms, which is an exciting mechanism that is tightly controlled by the virus. Sindbis virus ceases production of minus-strand (−) RNA several hours post-infection while plus-strand genomic (+) and sub-genomic (sg) RNAs are continually produced.
throughout the infection [21–24]. This transcriptional regulation is thought to be controlled by the sequential cleavage of the nonstructural polyprotein by the viral protease in the C-terminal region of nsP2 [25, 26]. The protease is active in the mature nsP2 protein and also in uncleaved polyproteins in which the nsP2 domain is present, albeit with slightly different cleavage site preferences [27]. In early infection, cleavage at the P3/4 site occurs cotranslationally, followed by cleavage at the P1/2 site and then at the P2/3 site, both in trans as the concentration of the protease increases [18]. Thus the initial replication complexes are composed of P123 + nsP4, which are gradually processed into nsP1 + nsP2 + nsP3 + nsP4 replication complexes. The early replication complexes preferentially transcribe minus strands, whereas the later replication complexes primarily transcribe plus strands [28, 29].

The conserved opal stop codon between nsP3 and nsP4 has 5-20% readthrough which allows for a relatively lower production of P1234 than P123. Sense codons at the opal stop codon position overproduce uncleaved P34 and reduce and delay the amount of nsP3, which also affects +/−/sgRNA ratios [19]. Our genome-wide data includes information on previously unstudied variants at these sites and variants at all the cleavage positions and nearby the opal stop codon, which allow further fine-tuning of RNA ratios. We are currently following up on these variants with RNA-imaging studies utilizing
RNA immuno-FISH to analyze the effects on RNA ratios and how they affect alphavirus replication and spread. A better understanding of how protein cleavage differentially affects production of various RNAs may also provide useful tools for synthetic biology.


Next we focused on the dynamics of synonymous mutations. No lethal synonymous mutation in SINV genome has previously been reported. Several codon wobble positions that are evolutionarily conserved among alphaviruses have been studied to discover critical RNA elements, but discovery of novel elements in non-conserved sites have thus far proved elusive. We ranked the enrichment between the *in vitro* transcript population and passages 2 → 5 from low to high, to focus on the most detrimental mutations, for all statistically significant (p < 0.05) synonymous mutations (*Table 3.1*).

The two most detrimental synonymous mutations were found at the 2nd and 3rd codon of the SINV genome (nsP1-2 and nsP1-3), immediately after the methionine start codon, which cannot have a synonymous mutation (*Figure 3.24*). Neither mutant codon (GAG → GAA in the nsP1-2 and AAG → AAA in nsP-3) are rare codons in the human genome (*Table 3.2*). These mutations lie
between the previously identified conserved sequence element (CSE) in the 5' UTR (nucleotides 1-44) and the 51-nt CSE in nsP1 (nucleotides 155-205). In fact, 12 of the 25 most de-enriched mutations are within the first 260 nucleotides of nsP1. Within the 10th percentile of all de-enriched mutations, 60% of synonymous mutations lie in nsP1, indicating the importance of the viral RNA sequence in the 5' end of the coding sequence (Table 3.1).
Figure 3.24. Viral Infection Population Dynamics Reveal Silent Mutations Adjacent to Nonstructural Polyprotein Start Codon Are Highly Detrimental. Codon mutants and amino acids coded for in nsP1 codon position 2 (A) and 3 (B) are on the x-axis and log relative enrichment on the y-axis. Bars are color coded as follows - blue = synonymous mutants, orange = nonsynonymous mutants, and red = stop codon mutants. P-values and confidence intervals calculated using Fisher exact test. Error bars represent 95% confidence intervals, and p-values are summarized as follows: n.s. (P > 0.05), * (P < 0.05), ** (P < 0.01), *** (P < 0.001), and **** (P < 0.0001).
A)

nsp1 Codon 2
GAG [Glu] at nucleotide positions 63-65

Enrichment (log2)

CAG   GTG   GAC   GGG   GCG   GAT   TAG   AAG   GAA
Gln   Val   Asp   Gly   Ala   Asp   Stp   Lys   Glu
B)

nspl Codon 3
AAG [Lys] at nucleotide positions 66-68

Enrichment (log)

<table>
<thead>
<tr>
<th>Codon Mutant</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGG</td>
</tr>
<tr>
<td>AAC</td>
</tr>
<tr>
<td>GAG</td>
</tr>
<tr>
<td>ATG</td>
</tr>
<tr>
<td>ACG</td>
</tr>
<tr>
<td>TAG</td>
</tr>
<tr>
<td>CAG</td>
</tr>
<tr>
<td>AAT</td>
</tr>
<tr>
<td>AAA</td>
</tr>
<tr>
<td>Arg</td>
</tr>
<tr>
<td>Asn</td>
</tr>
<tr>
<td>Glu</td>
</tr>
<tr>
<td>Met</td>
</tr>
<tr>
<td>Thr</td>
</tr>
<tr>
<td>Stp</td>
</tr>
<tr>
<td>Gln</td>
</tr>
<tr>
<td>Asn</td>
</tr>
<tr>
<td>Lys</td>
</tr>
</tbody>
</table>

****  n.s.  ****  ****  **  ****  ****  ****  ****  ****
After nsP1-2 and nsP1-3, the next most detrimental synonymous mutations were similarly in adjacent codons nsP4-608 and nsP4-609 (Figure 3.25 A, B). All synonymous mutations arising from single nucleotide substitutions at both these codons were very detrimental and none of the mutant codons were rare codons in the human genome. In fact, both mutant codons ATT and ATC in nsP4-608 have double codon usage frequency in humans as the original ATA codon (Table 3.2). Both these codons lie within the previously identified subgenomic promoter region, and immediately adjacent to the minimal promoter region necessary for subgenomic RNA transcription [30].
Figure 3.25. Low Tolerance for Silent Mutations in Codons Within Sindbis Subgenomic Promoter. Codon mutants of nsP4-608 (A) and nsP4-609 (B), which both lie in the previously identified subgenomic promoter region. X-axis represents codon mutants and amino acids coded for, and y-axis represents log relative enrichment of mutation frequency in the \textit{in vitro} transcribed population compared to low MOI passages 2 to 5. Bars are color coded and error bars and p-value summaries are calculated and plotted as in figure 3.24.
nsp4 Codon 608

ATA [Ile] at nucleotide positions 7572-7574

Enrichment (log₂)

GTA   ACA   CTA   AGA   ATG   TTA   AAA   ATT   ATC
Val   Thr   Leu   Arg   Met   Leu   Lys   Ile   Ile

Codon Mutant

---

378
B)

**nsp4 Codon 609**

AAG [Lys] at nucleotide positions 7575-7577
The G → A mutation at nucleotide 158 was the most detrimental synonymous mutation within the 51-nucleotide CSE (Figure 3.26). In previous studies, a virus with this mutation produced two distinct plaque morphologies which both retained the original mutation and can be plaque purified. However, the possibility of compensatory mutations outside this region were not investigated.

Figure 3.26. Codon Context of G158A Silent Mutant Within The 51-Nucleotide Conserved Sequence Element Copromoter. Relative enrichment
of mutation frequency in the *in vitro* transcribed population compared to low MOI passages 2 to 5. Codon mutants and amino acids coded for in nsP1 codon position 33 are on the x-axis and log relative enrichment on the y-axis. The G158A synonymous mutant (blue column) produced weird plaque phenotypes in previous studies, as described in the text. Bars are color coded and error bars and p-value summaries are calculated and plotted as in figure 3.24.

*Thus, some of the most detrimental silent mutations genome-wide occur in known RNA elements and also reveal novel sites:*

Identification of sites of interest on the viral RNA itself have been largely limited to studies on macroevolutionarily conserved sites. Phenotype-based screening has rarely yielded discovery of sites in which synonymous variants affect virus replication. Additionally, synonymous single mutants that have a lethal effect on the virus have not been identified, in line with our data that shows lesser negative selection effects on synonymous variants than nonsynonymous ones. Previous studies of nucleotide conservation across alphaviruses have revealed the following aptly named conserved sequence elements (CSE’s); a promoter in the 5’ UTR, a copromoter in nsP1 coding region (51-nt CSE), a subgenomic RNA promoter primarily in the intergenic UTR region, and a promoter in the 3' UTR. Mutagenesis studies have identified a downstream loop hairpin structure in the subgenomic mRNA that aids in translation initiation [31]. More recent computational analyses of alphavirus nucleotide conservation have identified several potential RNA
features, which have not yet been experimentally tested [32]. RNA structural analyses of the Sindbis genome using RNA-SHAPE have identified an additional structured region in nsP1 near the packaging signal region which reduces specific infectivity when disrupted. Additional structured regions discovered did not have a detectable impact on virus viability.

Synonymous variants in the subgenomic promoter region and 51-CSE copromoter are amongst the most detrimental synonymous variants in our analysis, highlighting agreement between previous studies on macroevolutionary nucleotide conservation and our genome-wide viability data on synonymous mutants (Figures 3.25 & 3.26). The 51nt-CSE lies near the 5' end of the genome within the coding region of nsP1 and is the most highly conserved coding region nucleotide sequence in the alphavirus family. It forms two hairpin loops as determined by computational folding analysis and experimental RNA-SHAPE, and is believed to be a copromoter for minus strand synthesis.

The most detrimental synonymous variant within the 51nt-CSE in our data, G158A, provided an interesting phenotype in a previous study. When transfected into CEFs, the virus produced two different plaque phenotypes which could be isolated using plaque purification; (i) plaques smaller than wild type which grew even better than wild type, and (ii) plaques five times the size
of wild type which had a fivefold decrease in titer relative to the small plaque variant. These plaque variants retained the original sequence within the 51-nt CSE region and the rest of the genome was not sequenced.

However, the most detrimental synonymous mutants lie in the second and third codons of the virus, adjacent to the methionine start codon which does not have synonymous codons. This site, between the 5' CSE and 51-nt CSE, does not have significant nucleotide conservation across alphaviruses. Along with the start codon, these codons were found to be part of a larger stem loop structure adjacent to the 51-nt CSE. Recent studies have revealed that the first several codons downstream of the start codon impacts translational elongation efficiency for cellular mRNA's, in what is termed a translational ramp [33]. We are currently investigating this possibility for SINV viral RNA.

Interestingly, RNA-SHAPE analysis of VEEV in comparison with Sindbis revealed that the majority of RNA structures are unique to each virus and cannot be identified with nucleotide conservation analysis. Our analysis expands the study of potential functional sites on the RNA itself that affect virus viability beyond macroevolutionarily conserved nucleotide regions. 40% of highly detrimental synonymous mutations are in the nsP1-coding region, indicating the presence of additional functional RNA elements near the 5' end of the genome. It would be interesting to further compile these genome-wide
viability data and experimental structure probing to characterize the role of novel RNA elements in viral replication mechanisms.

3.7. Relative Fitness: Another Estimate for the Distribution of Viral Variant Viability.

Next we wondered how the viral population continues to adapt during passaging, as the mutants generated by the virus's own low-fidelity polymerase encounter cellular selection and attempt to establish infection. Relative fitness using classical population genetics is a commonly used measure to estimate variant fitness during passaging [1, 34]. Accurate estimates of relative fitness, which measure the effects of selection on a variant, relies upon accurate in vivo estimates of the de novo rates at which that variant is introduced into the population before selection acts. We calculated these de novo specific mutation rates as the specific mutation frequencies of expected lethal mutations, such as those that produce stop codons, which should be produced anew in each generation (Figure 3.27).
Figure 3.27. Estimates of *in vivo* Mutation Rates of Sindbis Virus Polymerase. Mutation rates can be calculated as mutation frequencies at lethal sites [1]. Specific mutation rates (y-axis) for each type of nucleotide mutation (x-axis) were calculated as frequencies in stop codon mutations, and for specific mutations that cannot produce a stop codon, as frequencies in highly detrimental mutations. Error bars represent SEM.

Using our estimates for the *de novo* specific mutation rates and the virion population frequencies over the five low MOI passages, we calculated relative fitness for all variants across the Sindbis genome using a classical genetics model (*Figure 3.28*). The distribution of mutational fitness effects is consistent
with previous studies on poliovirus and small-scale studies on RNA viruses [1, 34–36]. The mutational distribution for nonsynonymous mutations is primarily deleterious and for synonymous mutations is more neutral (Figure 3.29).

$$\frac{V_p}{WT_p} = \frac{V_{p-1}}{WT_{p-1}} \times F_{rel} + \mu_{p-1}$$

**Figure 3.28. Classical Genetics Based Model to Estimate Relative Fitness.** Each variant at every passage is fit to the above model to determine the relative fitness value, where V and WT are counts of variant and wild type respectively, \(F_{rel}\) is relative fitness and \(p\) is passage number (time in generations). \(\mu\) is the de novo specific mutation rate from the WT nucleotide to the V nucleotide, which represents the rate at which de novo mutations are introduced into the population via polymerase error, upon which selection then acts. Our experimental conditions of using a low MOI for passaging and transferring a large population size for subsequent passage infections are assumed to minimize genetic drift and complementation [1].
Figure 3.29. Relative Fitness Distribution for All Single Nucleotide Variants of Sindbis Virus. Distributions of relative fitness (y-axis) for synonymous (blue) and nonsynonymous (orange) mutations. Dotted blue line indicates neutral relative fitness. Lower values indicate less fit variants that are selected against during passaging, with zero being the most detrimental variants detected.

Additionally, the fitness values of nonsynonymous variants were mapped onto three-dimensional structures of the Sindbis virion and uncleaved P23 to explore the effect of amino acid substitutions on structurally contiguous
functional regions which may not be continuous on the viral RNA, highlighting overall substitution tolerances (Figure 3.30) [37]. These data can be similarly mapped onto various Sindbis viral protein structures [37–40].
Figure 3.30. Relative Fitness of Variants Highlight Structure-Function Relationships. Fitness landscape of the nonsynonymous variant with the highest fitness value mapped onto the structure of the virion and the P23 protein to highlight structure-function relationships. Fitness value scale is represented as a color gradient from red (detrimental) to white (neutral) to blue (beneficial). A) Fitness values mapped onto a cryo-EM structure of the whole Sindbis virion, showing the envelope proteins [41]. Most substitutions are either detrimental or neutral, while a few are beneficial. B) Fitness values mapped onto a crystal structure of uncleaved P23 which stretches from the nsP2 protease domain to the nsP3 zinc-binding domain [37]. The penultimate glycine in the P2/3 cleavage site is highlighted, along with the N-terminal and C-terminal residues in the structure.
A) Sindbis Virion
B) P23 Protein
Thus, viral polymerase de novo mutation rates during passaging can be harnessed for estimating relative fitness:

The comparison of variant frequencies in the passages in which the lethal variants reach an equilibrium to the initial frequency in the *in vitro* population allows us to estimate the selection effects that each detectable variant faces. Such analysis works best for variants in which there is a large difference in frequency between the *in vitro* transcribed population and in the later passages after facing selection. For example, G→A mutations occur frequently in the *in vitro* transcribed population due to reaction conditions with sub-saturating rGTP described above, and thus detrimental mutations resulting from G→A changes can be readily assessed. However, such analyses is less clear for variants which by chance happen to have a similar frequency in both pre- and post-electroporation populations.

Relative fitness using classical population genetics is another estimation we used for variant dynamics during passaging, in which the *de novo* mutation rates of the viral polymerase is harnessed for fitness values of variants as they are newly generated by the polymerase and face the various selection bottlenecks during infection passages. Both estimates are complementary; in some cases, relative fitness estimates provide the advantage that more subtle
variations across passages in similar infection conditions can be compared, however it requires higher individual variant coverage at each passage precluding assessment of a portion of detected variants. Additionally, the frequency of highly detrimental mutants can be masked by the rate in which they are input into the population de novo by the polymerase, and consequently a significant percentage of detrimental variants share the lower bound value of zero precluding further relative comparison. Thus the in vitro mutation rates of SP6 polymerase and the in vivo mutation rates of the viral RdRp polymerase can both be utilized for estimating the effects of selection on viral variants.

Both types of estimates reveal a similar distribution of selection effects; the majority of mutations are detrimental or neutral and a minority are adaptive, with nonsynonymous mutants overall more detrimental than synonymous ones (Figures 3.6 & 3.29). We expect the neutral nonsynonymous variants to include amino acid substitutions that allow the protein to function similarly to the wild type, and neutral synonymous variants to include sites that primarily function for amino acid coding. At the edges of the distributions is likely where the most interesting biology lies; highly detrimental and adaptive synonymous variants may include RNA sequence elements important for RNA structure and RNA-protein binding, or suppress or enhance codon usage bias and gene expression. We show that highly detrimental nonsynonymous mutants include
premature stop codons and macroevolutionarily conserved sites such as the polymerase active site, and likely include many functionally important amino acids (Figure 3.10). We expect highly detrimental and adaptive nonsynonymous mutations to also suppress or enhance protein-protein interfaces, protein function and immune antagonism. Further studies in cell lines with intact innate immune systems can yield insights into sites on the viral genome that contribute to overcoming the immune response.

Additionally 3D protein structures for many viral proteins have been obtained via X-ray crystallography or cryo-EM, and RNA secondary structures for several structured RNA regions have been elucidated using RNA-SHAPe [42]. Spatial proximity of coded amino acids in the protein structure or base-pairing in the RNA structure often involve sites that are not linearly contiguous on the genome. As such, our variant viability data can be overlaid onto protein or RNA structures to add another dimension of information and correlate the importance of particular amino acids or base-pairing within its structural context respectively (Figure 3.30).

Next we wondered if the intracellular viral population encounters bottlenecks when it gets packaged into virions and exits the cell membrane, and how the population adapts to such bottlenecks. A Sindbis infection can produce ~300,000 genomic viral RNA inside a cell, but only ~10,000 are released as virions by the cell [17]. Within a highly infected cell, genomic RNA is ~5% of total RNA, and ~20% of polyA-tailed RNA. To achieve high sequencing coverage, we developed a scalable capture assay that utilizes biotinylated nucleotides complementary to Sindbis genomic RNA (Figure 3.31). The oligo-capture method selectively and significantly enriches the intracellular Sindbis genomic RNA to 98% of the RNA that is submitted to high-throughput sequencing (Figure 3.32).
Figure 3.31. Schematic of Oligonucleotide Hybridization Based Capture Method. A biotinylated 40-bp probe complementary to the structural region of Sindbis viral RNA is first bound to the intracellular viral RNA and then bound to streptavidin beads. After the unbound RNA is washed, the RNA-probe hybrid is released from the beads and then the probe is digested away, yielding relatively pure viral RNA (see materials and methods).
Figure 3.32. Oligo-Capture of Infected Cell RNA Yields Pure Sindbis Viral RNA. Percentages of viral (orange) and non-viral RNA (blue) in the intracellular RNA from the first passage before and after the oligo capture procedure, detected via qPCR. Numbers indicate percent of viral RNA in sample. Error bars represent SEM.
We performed oligo-capture and CirSeq sequencing on the intracellular genomic RNA from the five low MOI passages from which the virions were sequenced. We then compared the extracellular population to the intracellular population for passages 2 to 5 (Figure 3.33), for which we obtained an average read depth of $1.3 \times 10^6$ reads per nucleotide position and detected 98.6% of all possible variants (Figure 3.34 A, B). We observed a significantly decreased frequency for 40% of variants, and of these 21% and 71% were synonymous and nonsynonymous variants respectively. The frequency of only 2.5% of variants increased significantly, of which 37% and 63% were synonymous and nonsynonymous variants respectively. Thus variants in the intracellular population experience far more negative selection during packaging and egress than positive selection, indicating that the wild-type virus is relatively well adapted for these processes. However, the variants overall experience less frequency enrichment and de-enrichment, and hence less selection effects, relative to infection initiation (Figure 3.6).
Figure 3.33. Intracellular Viral RNA Enrichment and Sequencing Reveals Selection During Packaging and Egress. Relative enrichment of statistically significant (p < 0.05) synonymous (blue) and nonsynonymous (orange) mutations in the coding region of viral variants in passaged extracellular virions. The intracellular viral RNA served as a denominator for these calculations. Y-axis represents density, where the areas under the synonymous and nonsynonymous density curves total to 1, and x-axis is kept the same as in figure 3.6 for comparison.
Figure 3.34. Sequencing Coverage for Hybridization-Enriched Sindbis Intracellular RNA passages 2 to 5. A) Coverage density plot of total reads at each nucleotide of the Sindbis genome (x-axis) and calculated density values (y-axis). Blue bars represent binned total reads. Orange line is a parametric density and green line is a non-parametric density of total counts per nucleotide position. Legend shows the mode of each type of calculation, with parametric mode equaling the average counts per nucleotide position. B) Discrete raw counts of detected mutants. X-axis represents the edges of the bins used for mutant counts. Y-axis represents the number of all possible single nucleotide variants (~35,000) that fall within the bins. Blue bars represent detected variants, and the purple bar represents undetected variants. Also listed is the percentage of undetected variants (Percent Zeroes) and the average raw counts of detected variants (Non-Zero Counts Average).
We mapped the synonymous mutants that experienced significant changes to detect if there were any hotspots in which the RNA sequence is particularly affected during packaging and egress (Figure 3.35). The synonymous variants that were most enriched and de-enriched between the intracellular and extracellular populations lie in the nsP2 region of the genome, outside of the purported packaging signal region in nsP1 (Figure 3.36).

Figure 3.35. Selection During Packaging and Egress Occurs in Packaging Signal and Across Whole Genome. Relative enrichment of statistically significant (p < 0.05) synonymous mutations in the coding region of viral variants in passaged extracellular virions mapped to nucleotide position across the whole Sindbis genome. The intracellular viral RNA served as a denominator for these calculations. Point size is scaled to absolute value of log enrichment for visualization. Area between blue dotted lines represents the proposed packaging signal region.
Figure 3.36. The Most Adaptive and Detrimental Variants Detected During Packaging and Egress Lie In nsP2 Gene Region. Codon mutants of codons containing the most significantly enriched (A, G3995A) and de-enriched (B, C3326G) synonymous variants between the extracellular and intracellular populations, both of which lie in the nsP2 region. X-axis represents codon mutants and amino acids coded for, and y-axis represents log relative enrichment of mutation frequency when comparing variants in the extracellular (numerator) and intracellular (denominator) populations in low MOI passages 2 to 5. Bars are color coded as follows - blue = synonymous mutants, orange = nonsynonymous mutants, and red = stop codon mutants. P-values and confidence intervals calculated using Fisher exact test. Error bars represent 95% confidence intervals, and p-values are summarized as follows: n.s. (P > 0.05), * (P < 0.05), ** (P < 0.01), *** (P < 0.001), and **** (P < 0.0001).
nsp2 Codon 772
CTG [Leu] at nucleotide positions 3993-3995
B)

\textbf{nsp2 Codon 549}

CCC [Pro] at nucleotide positions 3324-3326
However, 5 of the 25 most de-enriched variants lay within the packaging signal (Table 3.3). In these codons, almost every possible synonymous mutation is significantly de-enriched (Figure 3.37). Within the 10th percentile of all de-enriched mutations, 8.2% of synonymous mutations lie in the packaging region, which comprises 4.4% of the total Sindbis coding region. The packaging signal is among regions, though not the only region, in which de-enriched synonymous mutants experience strong selection during packaging and egress (Figure 3.38).
Figure 3.37. Codon Mutants of Codons Containing Highly Detrimental Synonymous Mutations Within the Packaging Signal Region. G902C (A), G902T (A), C1103G (B), G1139A (C) and C1256G (D) are present within the 25 most significantly de-enriched mutations across the whole genome (Table 3.3). The majority of synonymous mutants within these codons are significantly de-enriched. X-axis represents codon mutants and amino acids coded for, and y-axis represents log relative enrichment of mutation frequency when comparing variants in the extracellular (numerator) and intracellular (denominator) populations in low MOI passages 2 to 5. Bars are color coded and error bars and p-value summaries are calculated and plotted as in figure 3.36.
nspl Codon 281
GTG [Val] at nucleotide positions 900-902

Enrichment (log_e)

ATG  GCG  GAG  GTA  GGG  CTG  GTT  GTC  TTG
Met  Ala  Glu  Val  Gly  Leu  Val  Val  Leu

Codon Mutant
B) nspl Codon 348
GCC [Ala] at nucleotide positions 1101-1103
**C**

**nspl Codon 360**

CTG [Leu] at nucleotide positions 1137-1139

![Bar chart showing enrichment (log scale) for different codon mutants comparing CAG, CTA, CCG, CGG, ATG, TTG, GTG, CTC, and CTT (Leu).](chart)

Enrichment (log)

<table>
<thead>
<tr>
<th>Codon Mutant</th>
<th>CAG</th>
<th>CTA</th>
<th>CCG</th>
<th>CGG</th>
<th>ATG</th>
<th>TTG</th>
<th>GTG</th>
<th>CTC</th>
<th>CTT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gln</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pro</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

n.s. indicates non-significant enrichment; '*' represents significant enrichment.
nspl Codon 399
CGC [Arg] at nucleotide positions 1254-1256

Enrichment (log_e)

n.s. * n.s. *** **** * **** **** ****

CAC TGC CCC CGT CGA GGC AGC CTC CGG
His Cys Pro Arg Arg Gly Ser Leu Arg

Codon Mutant
Figure 3.38. Mean Enrichment of Binned Sindbis Genome Regions During Packaging and Egress. Mean enrichment of Sindbis genome regions calculated in 300 nucleotide bins for significant (p < 0.05) positively enriched (A) and negatively enriched (B) synonymous variants. Bars represent log relative enrichment of mutation frequency between the extracellular (numerator) and intracellular (denominator) populations in low MOI passages 2 to 5. The bar containing the majority of the packaging signal region is indicated in purple. The packaging signal is among regions, though not the only region, in which de-enriched synonymous mutants experience strong selection during packaging and egress. Error bars represent SEM.
We then compared the relative fitness values of nonsynonymous mutants between the intracellular and extracellular populations and mapped this comparison onto the C-terminus of the capsid protein, arranged in a nucleocapsid shell (Figure 2.39). The structure of the C-terminal region of Sindbis capsid protein has been resolved, and cryo-EM allowed characterization of the whole nucleocapsid [39]. Previous studies have been unable to resolve the N-terminal 114 aa's of the capsid protein, perhaps due to the flexible nature of this region. As was expected from the enrichment analyses, the majority of nonsynonymous mutations show no difference in fitness between the intracellular and extracellular populations (Figure 2.39, white regions). However, the RNA-facing side of the capsid protein reveals several localized surface regions which face strong selection during egress (Figure 2.39, dark red regions). These may represent surface regions of the capsid protein that interface with the viral RNA during packaging, and warrants further characterization.
Figure 3.39. Intracellular vs Extracellular Fitness Map of Nucleocapsid Substitutions Reveals Regions Experiencing Strong Selection During Packaging and Egress. The relative fitness values, calculated as described in figure 3.28, for nonsynonymous variants in the extracellular population were normalized to their fitness values in the intracellular population and mapped onto a cryo-EM structure of the nucleocapsid shell [39]. These comparative values are displayed in a gradient from dark red (highly detrimental during packaging and assembly) to white (no difference between intracellular and extracellular populations) to dark blue (highly beneficial during packaging and assembly). Only the C-terminus region of the capsid protein is shown, excluding the N-terminal 114 residues for which structures are not available. Shown is the interior RNA facing side of the capsid protein when in a virion.
Collectively these data show that our oligo based viral RNA capture protocol isolates pure intracellular RNA which allows sequencing to high coverage, equivalent to coverage of extracellular RNA from purified virions. The CirSeq sequencing results revealed that not all variants in the intracellular RNA population are equally probable to be packaged into virions and exit the cell. Most variants are neutral or experience negative selection, and relatively few variants experience positive selection during packaging and egress. Synonymous variants that experience highly negative selection are slightly concentrated in the packaging signal region, but are present throughout the genome. The RNA facing side of the capsid protein, when assembled in a virion, contains localized surface regions of nonsynonymous variants that experience highly negative selection during packaging and egress.

Thus, the intracellular viral population adapts during packaging and egress:

During replication in mammalian cells, Sindbis virus genomic RNA forms around 5-10% of total cellular RNA. However, after packaging and egress, the virions contain almost exclusively the viral genomic RNA. Thus the virus utilizes mechanisms to both exclude the viral subgenomic mRNA which is several fold in abundance as the genomic RNA, and also other host RNA in the cytoplasm. Several lines of evidence have led to a proposed interaction between a packaging signal region on the genomic RNA, which lies in the nsP1 region of the genome, and the capsid protein as facilitating such RNA
exclusivity in virions [43].

These studies indicate that packaging has significant redundancy and it is highly unlikely that the facilitating interaction is restricted to one or a couple sites on the Sindbis genomic RNA. As such, we thought it was an ideal question to investigate by applying our genome-wide viability analysis. First, we investigated the selection that the viral variants face during packaging and egress by comparing the viral intracellular RNA population with the extracellular virions using CirSeq. Since the Sindbis genomic RNA is a minority population within the cell, we developed an oligonucleotide-based enrichment method which allows us to attain an intracellular genomic RNA pool that is as pure as that extracted from virions. This method allowed us to get equivalent genomic coverage during sequencing for both datasets, a requirement for such analysis. Additionally, CirSeq requires a large input pool of RNA to produce good sequencing libraries as several of the steps are inefficient, and the enrichment method was optimized to be cost-effective and scalable. We have already expanded the method to several other viruses including those that do not produce enough virions that can be processed into next-generation sequencing libraries. As such, we believe our enrichment method can be useful for a wide range of studies that would benefit from having pure microgram quantities of viral RNA.
Surprisingly, we did find significant positive and negative selection on many synonymous and nonsynonymous variants during packaging and egress. Thus not every viral variant in the intracellular RNA population is equally probable to be packaged and released as virions, and the process does seem to be influenced by the nucleotide context of the genome. While there is some enrichment of highly detrimental synonymous sites in the proposed packaging signal region, viral variants over the whole genome experienced positive and negative selection during packaging and egress.

In previous studies, the introduction of a multitude of synonymous mutations within the packaging signal region had a modest effect and still allowed viral replication and virion production. The introduction of more than 100 synonymous mutations in a GFP-tagged strain reduced viral titers ~3 logs in BHK-21 cells at early timepoints between 4 and 10 hrs p.i., and later timepoints were not assayed [44]. And the introduction of 70 synonymous mutations in this region reduced viral titers by ~1-2 logs between 12 and 24 hrs p.i. in Vero81 cells, but the titers become equivalent to wild type by 48 hrs p.i. [42]. Further profiling the dynamics of such viral variants during packaging and egress would provide further insight into the genome determinants of the RNA-protein interactions essential for this process.

Previous studies have also found regions within the N-terminus of the capsid
protein that aid in packaging efficiency and the specificity of packaging viral genomic RNA. A portion was shown to specifically interact with the RNA packaging signal \textit{in vitro}, and another segment when deleted led to loss of specificity and allowed the packaging of subgenomic RNA within nucleocapsids [43]. However, elucidating the involvement in packaging of the C-terminal region of the capsid protein, which contains the autoprotease domain, has been more elusive. Mapping the extracellular versus intracellular relative fitness values onto the nucleocapsid structure revealed that surface regions on the C-terminus of the capsid protein, on the RNA-facing side when assembled into a nucleocapsid, may also interface with viral RNA and be involved in packaging. Presumably, viral genomes with codon mutations that render its capsid unable to package will get generated by polymerase errors during replication, and can get packaged by capsids translated from subgenomic RNAs templated from the parental strand. During the next round of infection at a low MOI however, these variants will likely replicate normally upon entering a cell but cannot exit the cell efficiently. This will lead to an underrepresentation in the extracellular variant population compared to the intracellular population. However, such effects could also be on the RNA level, and the nucleotide change could be causing this underrepresentation, which just so happens to code for an amino acid change that is irrelevant to packaging. Further characterization is necessary to differentiate between these possibilities for such detected nonsynonymous variants.

Next we overlaid our variant viability data onto individual viral protein structures to assess the importance of particular amino acids within their structural contexts. Of the viral proteins, nsP1 does not have a crystal or cryo-EM structure available, and the lack of structures of proteins similar to nsP1 preclude homology modeling. However, significant portions of all remaining viral proteins either have crystal or cryo-EM structures available, or have high homology to such structures for similar proteins which allows for homology modeling using the Phyre2 software [45]. Fitness values of the SINV in Huh 7.5 passage series were mapped onto these structures using Chimera, and all mapped 3D structures shown are available as chimera session files for further analysis [46]. As in figure 3.30, CirSeq fitness values are displayed in a gradient from dark red indicating that all detected amino acid substitutions were heavily detrimental, to white (neutral) to dark blue, indicating that at least one amino acid substitution was highly adaptive [1].

A full length structure of nsP2 was acquired by incorporating the protease and methyltransferase-like domains from the crystal structure of P23 (Figure 3.30 B), and a homology model of the RNA helicase domain based on a crystal structure of chikungunya nsP2, using Phyre2 [37, 47]. Mapped fitness values reveal that nonsynonymous mutations in the papain-like protease active site residues, Cys-481 and His-558, are highly detrimental to the virus, as
expected (Figure 3.40, top left panel). Solvent-soluble surface imaging reveals that substitutions in a large surface area in the methyltransferase domain, near the C-terminal of the protein, are highly detrimental, implicating this region as a potential protein-protein interaction interface critical to the viral life cycle (Figure 3.40, top right panel). Additionally, these data show low tolerance for substitutions in sites known to contain temperature sensitive mutations and mutations which render the virus noncytopathic or unable to induce cellular transcriptional shutoff (Figure 3.40, bottom panels). Mutagenizing highly detrimental residues nearby these known sites may affect the same processes and allow for further fine-tuning of the resulting phenotypes. Similarly, fitness values were mapped to the macrodomain and the zinc-binding domain of nsP3 from the crystal structure of uncleaved P23 (Figure 3.41), with known sites detailed in the top left panel [37]. Any substitution to the four residues which coordinate the zinc ion are highly detrimental, indicating the requirement of retaining cysteine at these positions (Figure 3.41, top right panel). Two of the cysteines (Cys-288 and Cys-306) are on loops where mutations in adjacent residues are largely neutral. However, our data indicates that Pro-262 and Leu-264, which form a 4-residue stretch of extremely low mutational tolerance together with the other two cysteines (Cys-263 and Cys-265), are critical for replication. These residues may be in involved in proper positioning of the adjacent cysteines for metal ion binding. The site of the temperature sensitive mutation in ts4, Ala-268, is also adjacent to this site, and has fairly low
tolerance for substitutions. A surface diagram shows that the surface region around sites known to affect ADP-ribose binding activity in the macrodomain, asparagines Asn-10 and Asn-24, has somewhat low substitution tolerance indicated by the light red patch, and there is a surface region in which substitutions are extremely detrimental adjacent the macrodomain and zinc-binding domain linker, implicating this region in critical viral functions (Figure 3.41, bottom panels) [48].

Whereas Sindbis nsP4 does not have a crystal or cryo-EM structure available, conservation with other viral RNA-dependent RNA polymerases for which structures do exist allowed a predicted homology model for the region between nsP4 residues 185 and 569, which includes the active site, representing about two-thirds of the full length-protein (Figure 3.42). The nsP4 structure was homology modeled based on crystal structures for RNA-dependent RNA polymerases from foot-and-mouth disease virus, enterovirus 71, porcine Aichi virus, chicken Sicinivirus A and human rhinovirus serotypes 1B, 14 and 16 [45, 49–52]. The nsP4 structure model predicts that the catalytic active site GDD domain is inserted into a pocket via a β-sheet turn (Figure 3.42, top left panel). Zooming in on this pocket, CirSeq data predicts extremely low substitution tolerance for multiple residues - Phe-345, Asp-370, Phe-374, Asp-375, Leu-435, Asn-438 and Cys-505 - predicted to be 3-dimensionally adjacent the GDD domain in the active site pocket (Figure 3.42, top right
Further study of the effect of CirSeq-predicted mutations in these residues on polymerase activity can serve to validate the predicted homology model. The surface diagram shows a deep red pocket containing the GDD domain residues in the middle, indicating that all residues surrounding the GDD residues have very low nonsynonymous mutation tolerance (Figure 3.42, bottom left panel). Additionally, the protein model contains a smaller region of low tolerance near a site which contains the temperature-sensitive mutation in ts24, Gln-191, and also contains several highly detrimental and highly adaptive novel surface regions which warrant further characterization (Figure 3.42, bottom right panel).
Figure 3.40. Nonsynonymous Substitution Fitness Map of nsP2. Fitness values of the most tolerated mutation at each amino acid position are mapped onto a structure of nsP2, the viral protease, in a gradient from dark red (heavily detrimental) to white (neutral) to dark blue (highly adaptive). Top left panel is a zoom of the protease active site in a P23 crystal structure [37]. Residue numbers shown in the figure use a cumulative numbering system for the P1234 polyprotein, and are indicated within parentheses in this description, "A" refers to an arbitrarily designated chain number in the structure file and is not biologically relevant. Highlighted in green are the papain-like protease active site residues Cys-481(1021) and His-558(1098), along with the adjacent Trp-559(1099) which has been shown to be critical for protease activity. Also highlighted is the site of the temperature-sensitive mutation in ts17, Ala-517(1057). Top right panel shows a solvent-soluble surface diagram of a homology model of full-length nsP2, which incorporates the protease and methyltransferase-like domains from the aforementioned crystal structure, and a homology model of the RNA helicase domain based on a crystal structure of chikungunya nsP2 using Phyre2 [45, 47]. In addition to the active site residues, highlighted are the N-terminal Ala-1 (from P1/2 cleavage site) and C-terminal Ala-807 (from P2/3 cleavage site). Bottom left panel shows a ribbon diagram of the full length nsP2 with residues of interest highlighted, and bottom right panel is the same image rotated by 180°. Sites of temperature-sensitive mutations are highlighted in green; Asp-522 (ts7), Glu-163 (ts14), Val-275 (ts16), Ala-517 (ts17), Phe-509 (ts18), Leu-416 (ts19), Cys-304 (ts21), and Gly-736 (ts24). Highlighted in yellow, in addition to the active site, N-terminal and C-terminal residues, are Pro-726 where certain substitutions render the virus noncytopathic, and Pro-683 where certain substitutions render the virus incapable of inducing cellular transcriptional shutoff [53].
**Figure 3.41. Nonsynonymous Substitution Fitness Map of nsP3.** Fitness values of the most tolerated mutation at each amino acid position are mapped onto the crystal structure of nsP3 macro and zinc binding domains in a gradient from dark red (heavily detrimental) to white (neutral) to dark blue (highly adaptive) [37]. Residue numbers shown in the figure use a cumulative numbering system for the P1234 polyprotein, and are indicated within parentheses in this description, "A" refers to an arbitrarily designated chain number in the structure file and is not biologically relevant. Top left panel shows a ribbon diagram with residues of interest, with the macrodomain on the left and the zinc-binding domain on the right, attached by the linker region which traverses around a portion of nsP2, in which Leu-165(1512) coordinates with nsP2-726. Yellow highlighted residues Asn-10(1357) and Asn-24(1371) are involved in ADP-ribose hydrolase activity, and Cys-263(1610), Cys-265(1612), Cys-288(1635) and Cys-306(1653) coordinate a zinc ion [48]. Thr-286(1633) was shown to bind with the 51-nt CSE, and Tyr-267(1614), Arg-273(1620) and Arg-276(1623) coordinate a sulfate ion. Sites of temperature-sensitive mutations ts4, Ala-268(1615), and ts7, Phe-312(1659) are highlighted in green. Top right panel shows the zinc binding domain, with the four zinc-coordinating cysteines in yellow. The identified highly detrimental residues adjacent to one of the cysteines, Pro-262(1609) and Leu-264(1611) are shown in green, along with the nearby ts4 site. Bottom left panel shows the solvent-soluble surface diagram and bottom right panel is the same image rotated by 180°. The two ADP-ribose involved residues are highlighted, along with the N-terminal cleavage site Ala-1(1348) and the C-terminal residue in the crystal structure, Ile-326(1673).
Figure 3.42. Nonsynonymous Substitution Fitness Map of nsP4. Fitness values of the most tolerated mutation at each amino acid position are mapped onto a homology modeled structure of nsP4, the viral RNA-dependent RNA polymerase, in a gradient from dark red (heavily detrimental) to white (neutral) to dark blue (highly adaptive). A region of high homology allowed modeling of the stretch from N-terminal Ala-185 to C-terminal Val-569 using the Phyre2 software [45]. Top left panel shows a ribbon diagram with the conserved catalytic active site of the polymerase, the GDD domain, highlighted, along with the site of one of the temperature-sensitive mutations in ts24, Gln-191. Top right panel is zoomed in on the β-sheet turn which contains the GDD domain residues, highlighted in yellow. Highlighted in green are identified residues which do not tolerate substitutions that are predicted by the homology model to be near the GDD domain; Phe-345, Asp-370, Phe-374, Asp-375, Leu-435, Asn-438 and Cys-505. The CirSeq data indicate that they are involved in critical protein function, and may be involved in assisting polymerase activity. Bottom left panel shows the surface diagram with the highly detrimental pocket containing the GDD domain highlighted. Bottom right panel is the same image rotated by 180°, with the ts24 site highlighted, which is also in a surface region with low tolerance for substitutions.
A crystal structure of the C-terminal region of the capsid protein between residues 114 and 264, the autoprotease domain, was acquired and cryo-EM of the whole virion allowed placement of the individual protein structure within its nucleocapsid arrangement context [39, 54]. Fitness values were mapped onto the whole nucleocapsid (Figure 3.43 A) and the individual protein (Figure 3.43 B). A surface diagram shows several large contiguous regions with very low tolerance for substitutions on the N-terminal side of the autoprotease domain (Figure 3.43 B, top panels). Of the serine protease catalytic active site residues, our data confirms previous observations that Ser-215 cannot tolerate any substitutions, whereas substitutions in His-141 are more tolerated but still detrimental to the virus (Figure 3.43 B, bottom panels) [55]. All original capsid protein temperature-sensitive mutations occurred three residues away from these catalytic residues. CirSeq data indicates that Lys-138, mutated in ts13, has similarly low tolerance for substitutions as His-141. However, select substitutions in Pro-218, mutated in both ts2 and ts6, can be neutral in our infection context. A change of Pro-218 to serine abrogates autoprotease activity and infection at the non-permissive temperature is characterized by presence of the uncleaved structural polyprotein, the "ts2 protein" [17].

The full virion structure shown in figure 3.30 A contains the structures of the Sindbis envelope proteins E1, E2 and E3 [41]. Analysis of fitness values
mapped onto one unit of the E1-E2 heterodimer reveals that substitutions in E2 His-291, the temperature sensitive site in ts20, are fairly detrimental and the site lies within a stretch of 6 residues with very low mutational tolerance (Figure 3.44, top panels). Similarly, substitutions in Ala-106 and Arg-267 in the E1 protein, both temperature sensitive sites in ts23, are highly detrimental and both lie within stretches of very low mutational tolerance. In contrast, our data indicates that substitutions are better tolerated at E1 Lys-176 and its surrounding regions, which is mutated in ts10 to confer its temperature sensitivity phenotype. Of all the mapped glycosylation sites, Asn-139 in E1 has the least mutational tolerance in our conditions (Figure 3.44, bottom panels).

Solvent-soluble surface analysis indicates that contiguous surface regions (i) around E1 glycosylation site Asn-139, (ii) in one of ts23's E1 mutation sites Arg-267, and (iii) along a portion of the heterodimer's cytoplasmic tail have very low tolerance for amino acid substitutions, implicating them as important protein interfaces for viral functions. However, E1 and E2 also contain many of the most adaptive individual sites across the whole viral genome, indicating the importance of this heterodimer in driving viral evolution, and providing novel sites that warrant further characterization.

Interestingly, E3 contains a large surface region which contains fairly adaptive substitutions in all residues, nearby one of the conserved cysteines, Cys-19 (Figure 3.45 A, left panel). The majority of the opposite face of the protein has
low tolerance for substitutions (**Figure 3.45 A**, right panel). Finally, the 6K protein is a transmembrane protein which multimerizes to form ion channels [56]. 6K is the sole Sindbis structural protein which does not have a crystal or cryo-EM structure available. However, fairly good homology to a cryo-EM structure of vertebrate CALHM1/CALHM2, a calcium homeostasis modulator protein which forms large-pore membrane ion channels, and to an NMR structure of a portion of the hepacivirus GB Virus B NS2 protein allowed a homology structure prediction between residues 5 and 43, representing 70% of the full length protein [45, 57, 58]. Mapped fitness values revealed that substitutions in the three palmitoylated cysteines are detrimental, along with substitutions anywhere in the surface region where these cysteines reside (**Figure 3.45 B**). Further characterization of these identified detrimental residues or the several highly adaptive residues may help elucidate the functions of the 6K protein, which still remains largely mysterious.

Collectively, CirSeq enables genome-wide variant viability information to be combined with structural information to provide high-definition and bias-free insights into structure-function relationships, at particular sites and in continuous protein regions that are discontinuous on the viral RNA [1]. These data reveal additional information on the structural contexts of known functional sites and additionally reveal many novel sites and regions that are tuned by evolution, which are of interest for further studies. The 3D models of
all described structures with mapped viability data from both the passage series (Figures 3.30, 3.40 to 3.45) and the intracellular vs extracellular comparison (Figure 3.39) are available for researchers to further investigate their regions of interest [46]. Such structure-function studies highlighting structural elements that are tuned by evolution can assist in revealing drug targets and developing treatments for emerging viruses [1].
Figure 3.43. Nonsynonymous Substitution Fitness Map of Capsid Protein. Fitness values of the most tolerated mutation at each amino acid position are mapped onto structures of the capsid's autoprotease domain in a gradient from dark red (heavily detrimental) to white (neutral) to dark blue (highly adaptive) [54]. A) A fitness map of the nucleocapsid in a whole virion, which was acquired by placement of the crystal structure within a cryo-EM map [39]. B) Top left panel shows a surface diagram of the capsid protein with the N-terminal residue of the structure, Arg-114, highlighted. Top right panel is a 180° rotation with the C-terminal residue of the protein, Trp-264, highlighted. Bottom left panel shows a ribbon diagram with the serine protease catalytic active site residues, Ser-215 and His-141, highlighted in yellow and sites of temperature-sensitive mutations, Pro-218 (ts2 & ts6) and Lys-138 (ts13) highlighted in green. Bottom right panel is a zoom of the protease active site highlighting the catalytic residues and nearby temperature-sensitive mutation sites. "A" in the residue labels refers to an arbitrarily designated chain number in the structure file and is not biologically relevant.
Figure 3.44. Nonsynonymous Substitution Fitness Map of E1-E2 Envelope Heterodimer. Fitness values of the most tolerated mutation at each amino acid position are mapped onto a cryo-EM structure of the E1-E2 envelope protein heterodimer in a gradient from dark red (heavily detrimental) to white (neutral) to dark blue (highly adaptive) [41]. ".N" labels refer to E1 and ".P" labels refer to E2, and these letters are arbitrarily designated chain numbers in the structure file and are not biologically relevant. The N-terminal and C-terminal residues of each protein are labeled - the E1 structure is from Glu-2 to Thr-435. The E2 structure stretches from Phe-6 to Thr-398, with a prominent deletion of the region between Ser-178 and Thr-234. E1 residue numbers align with the viral genome, whereas E2 residue numbers do not, and the alternate numbers shown are listed within parentheses in this caption. Top left panel shows a ribbon diagram with residues of interest labeled. Sites of temperature-sensitive mutations in E1, Ala-106 and Arg-267 from ts23 and Lys-176 from ts10, are highlighted in green. Also highlighted in green is the site of the sole temperature sensitive mutation originally mapped to E2, His-291(256.P), from ts20. Glycosylation sites in E1, Asn-139 and Asn-245, and in E2, Asn-318(283.P), are highlighted in cyan. Top right panel is the ribbon diagram rotated by 180°, prominently showing the cytoplasmic tails of each envelope protein. Conserved cysteine palmitoylation sites in these cytoplasmic tails - Cys-430 in E1 and Cys-396(361.P) in E2 - are highlighted in pink. Several envelope protein glycosylation and palmitoylation sites lie outside the regions for which structural information exists and are not displayed. Bottom panels display the solvent-soluble surface of the E1-E2 heterodimer, with the same view angle as the corresponding top panel, and with the visible glycosylated and palmitoylated residues highlighted.
Figure 3.45. Nonsynonymous Substitution Fitness Map of E3 and 6K Proteins. Fitness values of the most tolerated mutation at each amino acid position are mapped onto the structures of each protein in a gradient from dark red (heavily detrimental) to white (neutral) to dark blue (highly adaptive). The left and right panels in A and B are rotated 180° from each other. A) Fitness values are mapped onto a cryo-EM structure of the E3 protein [41]. “.R” in the residue labels refers to an arbitrarily designated chain number in the structure file and is not biologically relevant. The N-terminal and C-terminal residues in the structure, and two conserved cysteines that are linked by a disulfide bond and may be important to viral function, are highlighted in green [59]. Cys-19 is part of a fairly adaptive (blue) surface region. B) Fitness values are mapped onto a homology modeled structure of 6K, which is a transmembrane protein that multimerizes to form an ion channel [56]. The region of fairly high homology allowed modeling of the stretch from N-terminal Glu-5 to C-terminal Leu-43 using the Phyre2 software [45]. The three palmitoylation sites on 6K, Cys-35, Cys-36 and Cys-39, are highlighted in green.
3.10. Conclusion.

The goal of these studies was to investigate SINV adaptation to various infection bottlenecks and utilize the dynamics of viral variants to study viral genetics in a high-throughput manner. To this end, we used CirSeq to profile SINV populations in mammalian cell culture. We analyzed how the starting \textit{in vitro} transcribed RNA population adapts to various bottlenecks encountered during electroporation and subsequent passaging, and during packaging and egress. Then we compared these data to previous studies of critical genome sites and expanded our study to new sites of interest. These data were then mapped to 3D protein structures to discover novel structure-function relationships. We posit that such unbiased high-throughput genetics pushes the envelope beyond the previous limits on discovery of viral functional elements.
Table 3.1. The Most De-enriched Synonymous Mutants Within The 10th percentile of all De-enriched Single Nucleotide Mutations During Infection. 89 synonymous mutants fall within the 10th percentile, representing 2.6% of mutants within the 10th percentile, with the remaining 97.4% being nonsynonymous mutants. Enrichment (Loge) is the relative enrichment of mutation frequency in the *in vitro* transcribed population compared to low MOI passages 2 to 5. Percentile is the percentile of enrichment value relative to all single-nucleotide mutations. Nucleotide position refers to the position of the first nucleotide of each codon in the Sindbis Toto1101 genome. P-values are summarized as follows: n.s. (P > 0.05), * (P < 0.05), ** (P < 0.01), *** (P < 0.001), and **** (P < 0.0001).
<table>
<thead>
<tr>
<th>Nucleotide Position</th>
<th>Protein</th>
<th>Amino Acid Number</th>
<th>Consensus Codon</th>
<th>Mutant Codon</th>
<th>Enrichment (Loge)</th>
<th>P Value Summary</th>
<th>Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>66</td>
<td>nsP1</td>
<td>3</td>
<td>AAG</td>
<td>AAA</td>
<td>-3.676</td>
<td>****</td>
<td>1.484</td>
</tr>
<tr>
<td>63</td>
<td>nsP1</td>
<td>2</td>
<td>GAG</td>
<td>GAA</td>
<td>-3.613</td>
<td>****</td>
<td>1.600</td>
</tr>
<tr>
<td>7575</td>
<td>nsP4</td>
<td>609</td>
<td>AAG</td>
<td>AAA</td>
<td>-3.207</td>
<td>****</td>
<td>2.688</td>
</tr>
<tr>
<td>4515</td>
<td>nsP3</td>
<td>139</td>
<td>GTA</td>
<td>GTC</td>
<td>-3.075</td>
<td>****</td>
<td>3.075</td>
</tr>
<tr>
<td>7572</td>
<td>nsP4</td>
<td>608</td>
<td>ATA</td>
<td>ATC</td>
<td>-2.956</td>
<td>***</td>
<td>3.468</td>
</tr>
<tr>
<td>522</td>
<td>nsP1</td>
<td>155</td>
<td>GTG</td>
<td>GTT</td>
<td>-2.929</td>
<td>****</td>
<td>3.569</td>
</tr>
<tr>
<td>219</td>
<td>nsP1</td>
<td>54</td>
<td>GAG</td>
<td>GAA</td>
<td>-2.912</td>
<td>****</td>
<td>3.616</td>
</tr>
<tr>
<td>102</td>
<td>nsP1</td>
<td>15</td>
<td>CCG</td>
<td>CCT</td>
<td>-2.883</td>
<td>****</td>
<td>3.722</td>
</tr>
<tr>
<td>156</td>
<td>nsP1</td>
<td>33</td>
<td>CAG</td>
<td>CAA</td>
<td>-2.864</td>
<td>****</td>
<td>3.793</td>
</tr>
<tr>
<td>243</td>
<td>nsP1</td>
<td>62</td>
<td>ACG</td>
<td>ACA</td>
<td>-2.833</td>
<td>****</td>
<td>3.921</td>
</tr>
<tr>
<td>3612</td>
<td>nsP2</td>
<td>645</td>
<td>GTA</td>
<td>GTC</td>
<td>-2.815</td>
<td>***</td>
<td>4.006</td>
</tr>
<tr>
<td>3867</td>
<td>nsP2</td>
<td>730</td>
<td>CTC</td>
<td>CTG</td>
<td>-2.686</td>
<td>****</td>
<td>4.512</td>
</tr>
<tr>
<td>243</td>
<td>nsP1</td>
<td>62</td>
<td>ACG</td>
<td>ACT</td>
<td>-2.668</td>
<td>****</td>
<td>4.604</td>
</tr>
<tr>
<td>288</td>
<td>nsP1</td>
<td>77</td>
<td>GAG</td>
<td>GAA</td>
<td>-2.634</td>
<td>****</td>
<td>4.742</td>
</tr>
<tr>
<td>Position</td>
<td>Type</td>
<td>Start</td>
<td>End</td>
<td>Nucleotide 1</td>
<td>Nucleotide 2</td>
<td>Change</td>
<td>Significance</td>
</tr>
<tr>
<td>----------</td>
<td>------</td>
<td>-------</td>
<td>-----</td>
<td>--------------</td>
<td>--------------</td>
<td>--------</td>
<td>--------------</td>
</tr>
<tr>
<td>3324</td>
<td>nsP2</td>
<td>549</td>
<td></td>
<td>CCC</td>
<td>CCG</td>
<td>-2.590</td>
<td>**</td>
</tr>
<tr>
<td>8772</td>
<td>e2</td>
<td>48</td>
<td></td>
<td>TCC</td>
<td>TCG</td>
<td>-2.505</td>
<td>*</td>
</tr>
<tr>
<td>195</td>
<td>nsP1</td>
<td>46</td>
<td></td>
<td>TCG</td>
<td>TCA</td>
<td>-2.496</td>
<td>****</td>
</tr>
<tr>
<td>7572</td>
<td>nsP4</td>
<td>608</td>
<td></td>
<td>ATA</td>
<td>ATT</td>
<td>-2.486</td>
<td>**</td>
</tr>
<tr>
<td>738</td>
<td>nsP1</td>
<td>227</td>
<td></td>
<td>TCG</td>
<td>TCA</td>
<td>-2.441</td>
<td>****</td>
</tr>
<tr>
<td>240</td>
<td>nsP1</td>
<td>61</td>
<td></td>
<td>GCG</td>
<td>GCA</td>
<td>-2.396</td>
<td>****</td>
</tr>
<tr>
<td>255</td>
<td>nsP1</td>
<td>66</td>
<td></td>
<td>ATA</td>
<td>ATT</td>
<td>-2.387</td>
<td>***</td>
</tr>
<tr>
<td>132</td>
<td>nsP1</td>
<td>25</td>
<td></td>
<td>CCG</td>
<td>CCT</td>
<td>-2.362</td>
<td>****</td>
</tr>
<tr>
<td>1707</td>
<td>nsP2</td>
<td>10</td>
<td></td>
<td>GTA</td>
<td>GTC</td>
<td>-2.361</td>
<td>***</td>
</tr>
<tr>
<td>7254</td>
<td>nsP4</td>
<td>502</td>
<td></td>
<td>GTC</td>
<td>GTA</td>
<td>-2.359</td>
<td>*</td>
</tr>
<tr>
<td>825</td>
<td>nsP1</td>
<td>256</td>
<td></td>
<td>TTG</td>
<td>TTA</td>
<td>-2.359</td>
<td>****</td>
</tr>
<tr>
<td>141</td>
<td>nsP1</td>
<td>28</td>
<td></td>
<td>GAG</td>
<td>GAA</td>
<td>-2.351</td>
<td>****</td>
</tr>
<tr>
<td>102</td>
<td>nsP1</td>
<td>15</td>
<td></td>
<td>CCG</td>
<td>CCA</td>
<td>-2.339</td>
<td>****</td>
</tr>
<tr>
<td>795</td>
<td>nsP1</td>
<td>246</td>
<td></td>
<td>TCG</td>
<td>TCA</td>
<td>-2.338</td>
<td>****</td>
</tr>
<tr>
<td>10053</td>
<td>6k</td>
<td>52</td>
<td></td>
<td>AAG</td>
<td>AAA</td>
<td>-2.313</td>
<td>****</td>
</tr>
<tr>
<td>792</td>
<td>nsP1</td>
<td>245</td>
<td></td>
<td>GGA</td>
<td>GGC</td>
<td>-2.283</td>
<td>*</td>
</tr>
<tr>
<td>900</td>
<td>nsP1</td>
<td>281</td>
<td></td>
<td>GTG</td>
<td>GTT</td>
<td>-2.276</td>
<td>***</td>
</tr>
<tr>
<td>Position</td>
<td>Protein</td>
<td>Gene</td>
<td>Codon</td>
<td>Codon</td>
<td>Score</td>
<td>Significance</td>
<td>Length</td>
</tr>
<tr>
<td>----------</td>
<td>---------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>--------------</td>
<td>--------</td>
</tr>
<tr>
<td>2229</td>
<td>nsP2</td>
<td></td>
<td>GTG</td>
<td>GTT</td>
<td>-2.274</td>
<td>****</td>
<td>6.570</td>
</tr>
<tr>
<td>237</td>
<td>nsP1</td>
<td></td>
<td>ACA</td>
<td>ACT</td>
<td>-2.254</td>
<td>****</td>
<td>6.700</td>
</tr>
<tr>
<td>870</td>
<td>nsP1</td>
<td></td>
<td>CAG</td>
<td>CAA</td>
<td>-2.224</td>
<td>****</td>
<td>6.892</td>
</tr>
<tr>
<td>849</td>
<td>nsP1</td>
<td></td>
<td>GTG</td>
<td>GTA</td>
<td>-2.181</td>
<td>****</td>
<td>7.128</td>
</tr>
<tr>
<td>708</td>
<td>nsP1</td>
<td></td>
<td>AAG</td>
<td>AAA</td>
<td>-2.180</td>
<td>****</td>
<td>7.146</td>
</tr>
<tr>
<td>5748</td>
<td>nsP3</td>
<td></td>
<td>TGA</td>
<td>TAA</td>
<td>-2.167</td>
<td>****</td>
<td>7.211</td>
</tr>
<tr>
<td>10086</td>
<td>e1</td>
<td></td>
<td>CCA</td>
<td>CCT</td>
<td>-2.143</td>
<td>****</td>
<td>7.371</td>
</tr>
<tr>
<td>2922</td>
<td>nsP2</td>
<td></td>
<td>GTG</td>
<td>GTA</td>
<td>-2.141</td>
<td>****</td>
<td>7.398</td>
</tr>
<tr>
<td>738</td>
<td>nsP1</td>
<td></td>
<td>TCG</td>
<td>TCT</td>
<td>-2.126</td>
<td>****</td>
<td>7.539</td>
</tr>
<tr>
<td>3957</td>
<td>nsP2</td>
<td></td>
<td>AGA</td>
<td>CGA</td>
<td>-2.118</td>
<td>***</td>
<td>7.587</td>
</tr>
<tr>
<td>858</td>
<td>nsP1</td>
<td></td>
<td>TTG</td>
<td>TTA</td>
<td>-2.109</td>
<td>****</td>
<td>7.634</td>
</tr>
<tr>
<td>762</td>
<td>nsP1</td>
<td></td>
<td>AAG</td>
<td>AAA</td>
<td>-2.085</td>
<td>****</td>
<td>7.791</td>
</tr>
<tr>
<td>681</td>
<td>nsP1</td>
<td></td>
<td>GCG</td>
<td>GCT</td>
<td>-2.076</td>
<td>****</td>
<td>7.838</td>
</tr>
<tr>
<td>225</td>
<td>nsP1</td>
<td></td>
<td>GAG</td>
<td>GAA</td>
<td>-2.062</td>
<td>****</td>
<td>7.924</td>
</tr>
<tr>
<td>522</td>
<td>nsP1</td>
<td></td>
<td>GTG</td>
<td>GTA</td>
<td>-2.062</td>
<td>****</td>
<td>7.927</td>
</tr>
<tr>
<td>828</td>
<td>nsP1</td>
<td></td>
<td>CAG</td>
<td>CAA</td>
<td>-2.058</td>
<td>****</td>
<td>7.947</td>
</tr>
<tr>
<td>441</td>
<td>nsP1</td>
<td></td>
<td>CCG</td>
<td>CCT</td>
<td>-2.056</td>
<td>****</td>
<td>7.968</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td>201</td>
<td>nsP1</td>
<td>48</td>
<td>CTG</td>
<td>CTA</td>
<td>-2.055</td>
<td>****</td>
<td>7.977</td>
</tr>
<tr>
<td>570</td>
<td>nsP1</td>
<td>171</td>
<td>GTG</td>
<td>GTA</td>
<td>-2.052</td>
<td>****</td>
<td>7.989</td>
</tr>
<tr>
<td>1512</td>
<td>nsP1</td>
<td>485</td>
<td>TCG</td>
<td>TCA</td>
<td>-2.051</td>
<td>****</td>
<td>8.007</td>
</tr>
<tr>
<td>900</td>
<td>nsP1</td>
<td>281</td>
<td>GTG</td>
<td>GTA</td>
<td>-2.036</td>
<td>****</td>
<td>8.125</td>
</tr>
<tr>
<td>96</td>
<td>nsP1</td>
<td>13</td>
<td>CAG</td>
<td>CAA</td>
<td>-2.035</td>
<td>****</td>
<td>8.134</td>
</tr>
<tr>
<td>942</td>
<td>nsP1</td>
<td>295</td>
<td>CCC</td>
<td>CCG</td>
<td>-2.033</td>
<td>*</td>
<td>8.151</td>
</tr>
<tr>
<td>1932</td>
<td>nsP2</td>
<td>85</td>
<td>CTG</td>
<td>CTA</td>
<td>-2.030</td>
<td>****</td>
<td>8.178</td>
</tr>
<tr>
<td>10056</td>
<td>6k</td>
<td>53</td>
<td>GTA</td>
<td>GTT</td>
<td>-2.029</td>
<td>**</td>
<td>8.193</td>
</tr>
<tr>
<td>753</td>
<td>nsP1</td>
<td>232</td>
<td>AAG</td>
<td>AAA</td>
<td>-2.020</td>
<td>****</td>
<td>8.249</td>
</tr>
<tr>
<td>3276</td>
<td>nsP2</td>
<td>533</td>
<td>TTG</td>
<td>TTA</td>
<td>-2.010</td>
<td>****</td>
<td>8.311</td>
</tr>
<tr>
<td>4287</td>
<td>nsP3</td>
<td>63</td>
<td>AAG</td>
<td>AAA</td>
<td>-2.007</td>
<td>****</td>
<td>8.329</td>
</tr>
<tr>
<td>10104</td>
<td>e1</td>
<td>14</td>
<td>CCG</td>
<td>CCT</td>
<td>-1.989</td>
<td>****</td>
<td>8.441</td>
</tr>
<tr>
<td>132</td>
<td>nsP1</td>
<td>25</td>
<td>CCG</td>
<td>CCA</td>
<td>-1.987</td>
<td>****</td>
<td>8.462</td>
</tr>
<tr>
<td>240</td>
<td>nsP1</td>
<td>61</td>
<td>GCG</td>
<td>GCT</td>
<td>-1.985</td>
<td>****</td>
<td>8.489</td>
</tr>
<tr>
<td>10047</td>
<td>6k</td>
<td>50</td>
<td>CTG</td>
<td>CTA</td>
<td>-1.983</td>
<td>****</td>
<td>8.497</td>
</tr>
<tr>
<td>642</td>
<td>nsP1</td>
<td>195</td>
<td>GCG</td>
<td>GCT</td>
<td>-1.981</td>
<td>****</td>
<td>8.515</td>
</tr>
<tr>
<td>666</td>
<td>nsP1</td>
<td>203</td>
<td>GAG</td>
<td>GAA</td>
<td>-1.968</td>
<td>****</td>
<td>8.633</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>522</td>
<td>nsP1</td>
<td>155</td>
<td>GTG</td>
<td>GTC</td>
<td>-1.820</td>
<td>****</td>
<td>9.831</td>
</tr>
<tr>
<td>642</td>
<td>nsP1</td>
<td>195</td>
<td>GCG</td>
<td>GCA</td>
<td>-1.817</td>
<td>****</td>
<td>9.843</td>
</tr>
<tr>
<td>1056</td>
<td>nsP1</td>
<td>333</td>
<td>ACG</td>
<td>ACT</td>
<td>-1.816</td>
<td>****</td>
<td>9.857</td>
</tr>
<tr>
<td>2331</td>
<td>nsP2</td>
<td>218</td>
<td>GAG</td>
<td>GAA</td>
<td>-1.811</td>
<td>****</td>
<td>9.902</td>
</tr>
<tr>
<td>405</td>
<td>nsP1</td>
<td>116</td>
<td>GAG</td>
<td>GAA</td>
<td>-1.809</td>
<td>****</td>
<td>9.925</td>
</tr>
<tr>
<td>2187</td>
<td>nsP2</td>
<td>170</td>
<td>AAG</td>
<td>AAA</td>
<td>-1.806</td>
<td>****</td>
<td>9.955</td>
</tr>
</tbody>
</table>
Table 3.2. Human Codon Usage of Selected Sindbis Synonymous Variants. Codon Usage in the Human Genome of Selected Sindbis Virus Consensus and Synonymous Mutant Codons Shown in Figures 3.24 to 3.26. Codon usage was calculated as frequency per 1000 codons using HIVE Codon/Codon Pair Usage Tables. Adapted from [60].

<table>
<thead>
<tr>
<th>Codon Position</th>
<th>Amino Acid</th>
<th>Codon Type</th>
<th>Codon</th>
<th>Frequency Per 1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>nsP1-2</td>
<td>Glu</td>
<td>Original</td>
<td>GAG</td>
<td>39.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mutant</td>
<td>GAA</td>
<td>33.72</td>
</tr>
<tr>
<td>nsP1-3</td>
<td>Lys</td>
<td>Original</td>
<td>AAG</td>
<td>31.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mutant</td>
<td>AAA</td>
<td>27.65</td>
</tr>
<tr>
<td>nsP1-33</td>
<td>Gln</td>
<td>Original</td>
<td>CAG</td>
<td>35.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mutant</td>
<td>CAA</td>
<td>14.15</td>
</tr>
<tr>
<td>nsP4-608</td>
<td>Ile</td>
<td>Original</td>
<td>ATA</td>
<td>8.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mutant</td>
<td>ATT</td>
<td>16.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mutant</td>
<td>ATC</td>
<td>18.67</td>
</tr>
<tr>
<td>nsP4-609</td>
<td>Lys</td>
<td>Original</td>
<td>AAG</td>
<td>31.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mutant</td>
<td>AAA</td>
<td>27.65</td>
</tr>
</tbody>
</table>
Table 3.3. The 25 Most Significantly De-enriched Synonymous Mutants of All Single Nucleotide Mutations During Packaging and Egress. Enrichment (Loge) is the relative enrichment of mutation frequency between the extracellular population (numerator) and intracellular population (denominator) in low MOI passages 2 to 5. Sites within the proposed packaging signal region is highlighted in blue, and are plotted in Figs E, F & Supplementary Figs E, F. Percentile is the percentile of enrichment value relative to all single-nucleotide mutations. Nucleotide position refers to the position of the first nucleotide of each codon in the Sindbis Toto1101 genome. P-values are summarized as follows: n.s. (P > 0.05), * (P < 0.05), ** (P < 0.01), *** (P < 0.001), and **** (P < 0.0001).
<table>
<thead>
<tr>
<th>Nucleotide Position</th>
<th>Protein</th>
<th>Amino Acid Number</th>
<th>Consensus Codon</th>
<th>Mutant Codon</th>
<th>Enrichment (Loge)</th>
<th>P Value Summary</th>
<th>Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>3324</td>
<td>nsP2</td>
<td>549</td>
<td>CCC</td>
<td>CCG</td>
<td>-3.376</td>
<td>****</td>
<td>0.018</td>
</tr>
<tr>
<td>10176</td>
<td>e1</td>
<td>38</td>
<td>GTT</td>
<td>GTG</td>
<td>-2.994</td>
<td>****</td>
<td>0.077</td>
</tr>
<tr>
<td>4830</td>
<td>nsP3</td>
<td>244</td>
<td>CGC</td>
<td>CGG</td>
<td>-2.958</td>
<td>****</td>
<td>0.086</td>
</tr>
<tr>
<td>1254</td>
<td>nsP1</td>
<td>399</td>
<td>CGC</td>
<td>CGG</td>
<td>-2.846</td>
<td>****</td>
<td>0.127</td>
</tr>
<tr>
<td>8772</td>
<td>e2</td>
<td>48</td>
<td>TCC</td>
<td>TCG</td>
<td>-2.636</td>
<td>**</td>
<td>0.239</td>
</tr>
<tr>
<td>1101</td>
<td>nsP1</td>
<td>348</td>
<td>GCC</td>
<td>GCG</td>
<td>-2.632</td>
<td>***</td>
<td>0.248</td>
</tr>
<tr>
<td>2010</td>
<td>nsP2</td>
<td>111</td>
<td>CCC</td>
<td>CCG</td>
<td>-2.623</td>
<td>***</td>
<td>0.257</td>
</tr>
<tr>
<td>10206</td>
<td>e1</td>
<td>48</td>
<td>ACC</td>
<td>ACG</td>
<td>-2.620</td>
<td>**</td>
<td>0.263</td>
</tr>
<tr>
<td>9252</td>
<td>e2</td>
<td>208</td>
<td>ACC</td>
<td>ACG</td>
<td>-2.582</td>
<td>**</td>
<td>0.310</td>
</tr>
<tr>
<td>1137</td>
<td>nsP1</td>
<td>360</td>
<td>CTG</td>
<td>CTT</td>
<td>-2.541</td>
<td>****</td>
<td>0.378</td>
</tr>
<tr>
<td>2127</td>
<td>nsP2</td>
<td>150</td>
<td>CTG</td>
<td>CTT</td>
<td>-2.533</td>
<td>****</td>
<td>0.387</td>
</tr>
<tr>
<td>5580</td>
<td>nsP3</td>
<td>494</td>
<td>TCC</td>
<td>TCG</td>
<td>-2.514</td>
<td>**</td>
<td>0.414</td>
</tr>
<tr>
<td>2574</td>
<td>nsP2</td>
<td>299</td>
<td>CCT</td>
<td>CCG</td>
<td>-2.500</td>
<td>**</td>
<td>0.426</td>
</tr>
<tr>
<td>Species</td>
<td>Domain</td>
<td>Position</td>
<td>Mismatch</td>
<td>NTM [a]</td>
<td>P-value</td>
<td>q-value</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>--------</td>
<td>----------</td>
<td>----------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>nsP2</td>
<td>112</td>
<td>GCC</td>
<td>GCG</td>
<td>-2.497</td>
<td>**</td>
<td>0.432</td>
</tr>
<tr>
<td>9258</td>
<td>e2</td>
<td>210</td>
<td>ACC</td>
<td>ACG</td>
<td>-2.444</td>
<td>****</td>
<td>0.497</td>
</tr>
<tr>
<td>3507</td>
<td>nsP2</td>
<td>610</td>
<td>CTG</td>
<td>CTT</td>
<td>-2.431</td>
<td>****</td>
<td>0.514</td>
</tr>
<tr>
<td>900</td>
<td>nsP1</td>
<td>281</td>
<td>GTG</td>
<td>GTC</td>
<td>-2.415</td>
<td>****</td>
<td>0.541</td>
</tr>
<tr>
<td>3612</td>
<td>nsP2</td>
<td>645</td>
<td>GTA</td>
<td>GTC</td>
<td>-2.395</td>
<td>**</td>
<td>0.571</td>
</tr>
<tr>
<td>9636</td>
<td>e2</td>
<td>336</td>
<td>GTG</td>
<td>GTT</td>
<td>-2.384</td>
<td>****</td>
<td>0.597</td>
</tr>
<tr>
<td>8094</td>
<td>capsid</td>
<td>150</td>
<td>GTG</td>
<td>GTT</td>
<td>-2.375</td>
<td>****</td>
<td>0.618</td>
</tr>
<tr>
<td>2592</td>
<td>nsP2</td>
<td>305</td>
<td>ACC</td>
<td>ACG</td>
<td>-2.373</td>
<td>**</td>
<td>0.621</td>
</tr>
<tr>
<td>9513</td>
<td>e2</td>
<td>295</td>
<td>CTC</td>
<td>CTG</td>
<td>-2.328</td>
<td>****</td>
<td>0.710</td>
</tr>
<tr>
<td>900</td>
<td>nsP1</td>
<td>281</td>
<td>GTG</td>
<td>GTT</td>
<td>-2.318</td>
<td>****</td>
<td>0.721</td>
</tr>
<tr>
<td>6927</td>
<td>nsP4</td>
<td>393</td>
<td>CTG</td>
<td>CTT</td>
<td>-2.313</td>
<td>****</td>
<td>0.736</td>
</tr>
<tr>
<td>4458</td>
<td>nsP3</td>
<td>120</td>
<td>CGC</td>
<td>CGG</td>
<td>-2.312</td>
<td>****</td>
<td>0.739</td>
</tr>
</tbody>
</table>
3.11. Chapter III References.


59. Parrott MM, Sitarski SA, Arnold RJ, Picton LK, Hill RB, Mukhopadhyay S

Chapter IV: Development of Bioinformatic Analyses Software for Genome-Wide Virus Genetics.

4.1. Introduction to Bioinformatics Pipeline for CirSeq Data.

Herein is a brief summary of the bioinformatic methods pipeline that was used for all analyses presented in the prior chapter. CirSeq is a rather unique protocol that produces reads with built-in tandem repeats, and with far higher accuracy than typical protocols. Consequently, the downstream bioinformatic pipelines are CirSeq-specific, and quite different than commonly performed RNAseq analyses.

Ashley Acevedo developed the protocol and a bioinformatic pipeline as a graduate student in Dr. Raul Andino's laboratory at the University of California, San Francisco [1]. These programs, collectively called CirSeqScripts, were built using R and Python 2.7, and are available on GitHub [2]. CirSeqScripts can generate plots such as every figure in her paper, as well as similar figures in the previous chapter of this thesis (Figures 3.2 & 3.3), and is described in further detail in Ashley's methods publication in Nature Protocols [3]. I provide a brief overview only as it relates to downstream pipelines below.

I learned R as part of my first year course, Quantitative Biology, taught by Jason Banfelder within the Department of Physiology, Biophysics and
Systems Biology at Weill Cornell. To be able to better understand and manipulate these scripts, I learned python 2 in an elective course during my fourth year, Programming for the Life Sciences, at Rockefeller University. Ashley Acevedo joined the Rice lab as a postdoc, and we built several updates to CirSeqScripts that operate directly on the files output by Illumina next-generation sequencing machines, and are briefly described below (Figure 4.1). The most prominent of these are scripts which process information on insertions and deletions – shunted into a separate raw sequence file in the original CirSeqScripts – into easily readable tables. Under the guidance of Ashley and others, I then wrote a python program that operates on tables output by CirSeqScripts and generates all remaining analyses and tables for plots shown in the prior chapter. Plotting is then performed within python or using R and R's ggplot2 package (Figure 4.1). Due to the unique nature of these files and our analyses, this program was written from scratch to maximize efficiency and necessary features and minimize bugs. This was often necessary even for common programs; codon translation scripts are common but most treat all stop codons similarly, whereas we had to treat them individually since the opal stop codon between nsP3 and nsP4 in Sindbis virus does not tolerate the various stop codons similarly (Figure 1.16 B). Bugs within code written by others seems to be the bane of existence for my cousins who are professional software developers. Confronted with the necessity of providing a name which differentiates it from
CirSeqScripts, I have tentatively named this collection of programs as PushpaScripts, after Telford Work's pet leopard (Figure 1.5 H). This name may change in the future.

Figure 4.1. CirSeq Bioinformatics Pipeline Summary. Data from an Illumina sequencer run of CirSeq libraries are first processed and aligned using CirSeqScripts, which outputs mutation counts and insertion and deletion information. The mutation data from multiple samples is then processed using PushpaScripts, which primarily consists of a database and analyses tools in python, which is then plotted using R and ggplot2. Mutation frequency data can be analyzed using a sliding window analyses MATLAB application written by Zak Singer. Fitness or enrichment values output by the python software can also be mapped onto 3D structures using UCSF Chimera, which was developed in the same building in which Ashley Acevedo developed CirSeq [4]. An example of fitness values mapped onto Sindbis P23 and a list of all Sindbis
proteins for which 3D structures are available is also shown. All of these tools are available separately or within the CircSeqLinuxHD operating system distribution.

PushpaScripts automatically generates a panoply of what we hope are useful plots when run on appropriate input files. Additionally, PushpaScripts was consciously designed to allow – with just a few lines of code – the user to perform any specific analyses he or she desires and filter the data in any which way. A very brief overview of PushpaScripts is described below; just enough to describe expectations, folder structure, required inputs, and expected outputs. It should be taken as a brief summary for a potential user, who can gain far more descriptive information - hundreds of pages worth describing every aspect that is larger than this entire thesis dissertation - within the program files themselves. Using the scripts to produce default outputs would benefit from basic knowledge of command line terminal and the ability to install software from within the terminal. Manipulating the scripts requires basic programming ability in python 2, especially pertaining to objects and classes, and/or R based on which aspects the user would like to harness. I developed these after a one semester introductory class in R and in python, and a similar background gained via class or self-study would equip any potential user with the requisite skillset. Since downloading the scripts is a prerequisite for usage, we decided that it was a more suitable home for such information than this printed document. Every program has an accompanying file, named as the program + '_manual.txt', that describe every feature and
function within the program. The manual can also be accessed within each program by using the .man() function. Additionally, every set of lines within the python and R code itself has comment lines describing them immediately above.

Finally, I briefly describe best practices to install the software dependencies for CirSeqScripts and PushpaScripts. Unfortunately, porting these from Ashley's computer to mine took a while to get everything working properly. Advancements in package managers such as Conda have perhaps made such porting easier now. To prevent any porting issues for potential users, I have built a Linux-based deployable operating system (OS) with all CirSeq-related analyses pipelines that we have used and appropriate versions of all prerequisite software packages pre-installed and working properly. It also comes with popular R and python 2 integrated development environments (IDE's) such that the user can manipulate code and run custom analyses easily. All the analyses presented within this thesis dissertation was performed on this operating system, named 'CircSeqLinuxHD'. The easiest way to begin using these bioinformatic pipelines would be to acquire an empty drive and copy CircSeqLinuxHD onto it and connect it to a computer. Everything should be good to go upon booting into this drive. Brief instructions and best practices as to choosing a drive and a computer is described in this section.
4.2. **CirSeqScripts: Summary, Expected Inputs and Outputs.**

CirSeqScripts as developed by Ashley during her graduate studies is available on her GitHub page [2]. The version that was expanded by us in the Rice lab (version 4.3) is available through the lab either separately or as part of CircSeqLinuxHD OS. The programs require at minimum two input files: (i) one fasta format file with the consensus sequence of the RNA to be analyzed, and (ii) one or more '.fastq.gz' raw sequence files, that are output from Illumina sequencers, containing sequencing reads of CirSeq libraries. If the run was indexed the raw sequence files must be pre-sorted by index. The best data is acquired by notifying the service used for sequencing to turn off quality-score binning prior to sample submission. Once the fasta, fastq.gz and CirSeqScripts folder are within the same directory, the program can be run by simply navigating to the directory in terminal and running the following command:

```
bash ./circseqscripts/run.sh subfolder *.fasta *.fastq.gz
```

Wherein 'subfolder' can be named as anything, and a folder with that name will be generated containing all the output files. On a computer equipped with an Intel i7-4790K processor, running CirSeqScripts takes several hours to process data from a full MiSeq run, and a day or two to process data from 1/5th of an indexed HiSeq 2500 run. Many plots, including many of the ones in
Ashley's publication and several quality control ones, are generated automatically within the /plots folder. Three types of tables are produced upon analysis of insertions and deletions, and are within the /indels folder: a file with each individual insertion or deletion and the flanking regions which defaults to 12 bp (Table 4.1 A), a file with a tally of insertions or deletions per genome position (Table 4.1 B), and a size distribution of insertions or deletions (Table 4.1 C). We have only performed basic plotting of these tables, available within the /plots folder and in Figures 3.15-3.18. The flanking regions have potential for further analyses in regards to motifs that increase the frequency of insertions and deletions.
Table 4.1. CirSeqScripts Insertion and Deletion Analysis Tables. The first few rows of an indel and flanking site information file (A), an indel tally per nucleotide position file (B), and a tally of lengths of indels file (C). Such tables are generated by CirSeqScripts version 4.3 separately for insertions and deletions and plotted within the /plots subfolder. These indel data tables are plotted directly without further processing within PushpaScripts, and examples of plots are shown in figures 3.15 to 3.18.

A)

<table>
<thead>
<tr>
<th>position</th>
<th>length</th>
<th>bases</th>
<th>5p_flank</th>
<th>3p_flank</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>1</td>
<td>C</td>
<td>CGGCGTAGTACA</td>
<td>ACTATTGAATCA</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>C</td>
<td>GCGTAGTACACA</td>
<td>TATTTGAATCAAAA</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>TA</td>
<td>CGTAGTACACAC</td>
<td>TTGAATCAAACACA</td>
</tr>
<tr>
<td>21</td>
<td>8</td>
<td>TATTAAT</td>
<td>CGTAGTACACAC</td>
<td>CAAACAGCCGAC</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>TAT</td>
<td>CGTAGTACACAC</td>
<td>TGAATCAACACAG</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>T</td>
<td>TAGTACACACTA</td>
<td>TGAATCAACACAG</td>
</tr>
</tbody>
</table>

B)

<table>
<thead>
<tr>
<th>position</th>
<th>total_indels</th>
<th>total_bases</th>
<th>average_bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>1</td>
<td>1</td>
<td>1.00000000</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>1</td>
<td>1.00000000</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>13</td>
<td>4.33333333</td>
</tr>
<tr>
<td>23</td>
<td>2</td>
<td>2</td>
<td>1.00000000</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
<td>4</td>
<td>1.00000000</td>
</tr>
<tr>
<td>26</td>
<td>2</td>
<td>2</td>
<td>1.00000000</td>
</tr>
</tbody>
</table>

C)

<table>
<thead>
<tr>
<th>length</th>
<th>reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>7819</td>
</tr>
<tr>
<td>2</td>
<td>1364</td>
</tr>
<tr>
<td>3</td>
<td>891</td>
</tr>
<tr>
<td>4</td>
<td>415</td>
</tr>
<tr>
<td>5</td>
<td>352</td>
</tr>
</tbody>
</table>
Two output files in the main subfolder are used as the sole inputs, per sample, for PushpaScripts. One contains the counts of each nucleotide at each genome position of the genome indicated in the fasta file, along with the consensus base (Table 4.2 A). The other contains mutation rates for each specific base-to-base mutation, using maximum likelihood estimates (Table 4.2 B) [1]. These can be converted within PushpaScripts to mutation rates based on mutations to stop codons and other heavily detrimental codons.
Table 4.2. CirSeqScripts Mutation Counts and Rates Tables. The first few rows of a mutation counts file (A) and a mutation rates file (B) are shown. These data are from CirSeq of in vitro transcribed Sindbis Toto1101 RNA. These files are generated for each sample that CirSeqScripts is run on, and serve as the input for PushpaScripts. Position and consensus are the nucleotide number and nucleotide, respectively, in the .fasta sequence that the NGS reads are aligned to, MLE is the maximum likelihood estimate and SE is standard error.

A)

<table>
<thead>
<tr>
<th>position</th>
<th>consensus</th>
<th>A_count</th>
<th>C_count</th>
<th>G_count</th>
<th>T_count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>44254</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>102997</td>
</tr>
<tr>
<td>3</td>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>164247</td>
</tr>
<tr>
<td>4</td>
<td>G</td>
<td>0</td>
<td>1</td>
<td>209372</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>247211</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>C</td>
<td>1</td>
<td>281868</td>
<td>0</td>
<td>13</td>
</tr>
</tbody>
</table>

B)

<table>
<thead>
<tr>
<th>Type</th>
<th>MLE</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>1.74616196846568E-05</td>
<td>1.80378528317035E-07</td>
</tr>
<tr>
<td>AG</td>
<td>8.16391496859936E-06</td>
<td>8.13133888633433E-08</td>
</tr>
<tr>
<td>AT</td>
<td>9.1057442733305E-06</td>
<td>1.15909138267513E-07</td>
</tr>
<tr>
<td>CA</td>
<td>1.31400364920296E-05</td>
<td>1.11734258269176E-07</td>
</tr>
<tr>
<td>CG</td>
<td>2.42541456851843E-06</td>
<td>4.08229093646525E-08</td>
</tr>
</tbody>
</table>
By default these are calculated using a quality score threshold of 20, as explained in Ashley's publication, and thus these files are named 'Q20threshold.txt' and 'MutationRates_Q20MLE.txt', respectively [1]. For PushpaScripts these two files must be renamed as the sample with '_mu' added to the mutation rates file and added to PushpaScript's /data folder. For example, if the user wishes to refer to a dataset as 'sinv_passage_1' within PushpaScripts, the files should be named 'sinv_passage_1.txt' and 'sinv_passage_1_mu.txt' respectively.

4.3. PushpaScripts: Summary, Expected Inputs and Outputs.

At its core, PushpaScripts is a database built in python that performs, indexes and stores all necessary calculations and information derived from the two files shown in Table 4.2, for multiple CirSeq samples. These data within the database, which ascribe to the philosophies described below, are then harnessed for further analyses. PushpaScripts contains a variety of functions that allows the user to filter the data however he or she should desire, and generates tables with these data that can be read and plotted in Microsoft Excel, Graphpad Prism, R, or any similar application. Since R and its plotting package ggplot2 are free and powerful, these are used for most built-in plotting functions. An example of several built-in filtering functions within the Compare program module is shown in Figure 4.2 A. The inputs and outputs
for all filter functions are matched, so that these functions can be strung together and applied in any order. An example analysis that requires just a few short lines of code is shown in Figure 4.2 B, the output of which is shown in Table 3.1, showing 8 of the 30 total columns.
Figure 4.2. PushpaScripts Database and Filter Functions. A) Summary of the three databases available per sample (orange) and various filter functions for these databases in the Compare program module which can be used in any order. B) An example analysis which generates the data table partially shown in table 3.1 and the plots shown in figure 4.3 D. The codon database is loaded for both pre-selection in vitro transcribed Sindbis RNA and post-selection viral RNA passaged in cells, and various filter functions are strung together to produce the specific data desires. (*) Indicates a built-in addition function which can combine multiple samples. The actual code used to generate this analysis is shown in the text - creating a graphical user interface (GUI) for analysis as in this figure is beyond my current skillset, but perhaps one day.
**Compare Data:** ________________ vs ________________

<table>
<thead>
<tr>
<th>Nucleotide Bases Table</th>
<th>Codons Table</th>
<th>Amino Acids Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sindbis Genome Region: ______</td>
<td>Synonymous</td>
<td>Non-Synonymous</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Remove Zeros</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column Values</td>
<td>Columns &gt; No.</td>
<td>Nucleotide Bases</td>
</tr>
<tr>
<td>• Column:_____</td>
<td>• Columns:_____</td>
<td>• Consensus:_____</td>
</tr>
<tr>
<td>• Values:_____</td>
<td>• Threshold:_____</td>
<td>• and/or</td>
</tr>
<tr>
<td>• Hierarchize All</td>
<td></td>
<td>• Variants:_____</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Hierarchize All</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sort Enrichment</td>
<td>Sort Enrichment</td>
<td>Recalculate Percentile</td>
</tr>
<tr>
<td>Highest → Lowest</td>
<td>Lowest → Highest</td>
<td>Maximum Enrichments</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Per Codon</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Fisher Exact Test
- P-Values
- Default: < 0.05

**Complete Tables** | **Specific Filter Functions** | **General Filter Functions** | **Sort and Crop Functions**
Question: What are the most de-enriched (detrimental) codon mutations which don’t change the viral protein amino acid upon passaging of Sindbis virus in vitro transcribed RNA?

Compare Data: \textit{in vitro} transcribed RNA vs Passage 2 + 3 + 4 + 5*
Within PushpaScripts, the following lines of code (functions in blue, user-specified function inputs in green) performs the analysis shown in Figure 4.2 B. It outputs the table partially shown in Table 3.1, that is readable in Microsoft Excel, GraphPad Prism, R, etc., for further analysis. It also generates the ordered plots shown below in Figure 4.3 D:

```python
analysis_one = Compare("ivt", "pass2345")  # points index to datasets
analysis_one.fill_sieve("codons")  # loads database of codon mutants
analysis_one.sift_synonymous("synonymous")  # gets synonymous mutants only
analysis_one.sift_p_value(0.05)  # sets statistical significance threshold
analysis_one.sort_lowest_first()  # sorts to most de-enriched at top of table
analysis_one.crop(100)  # crops table to top 100 (hence, "billboard")
analysis_one.write_sieve("synonymous_most_deenriched")  # outputs files
```

PushpaScripts is provided with all the sample data that was analyzed in chapter 3. The Builder program module serves as an umbrella program that appropriately processes all the data in sequence. When Builder is run, databases of nucleotide mutations, codon mutations, and amino acid mutations are generated for each sample individually, and then for comparisons between samples. When run for the first time, generating the database took around 1 hr 40 mins per sample on a 2016 15" Apple Macbook Pro, and less than a second to load in all subsequent analyses. Summaries of these databases are provided under the Process and Compare module descriptions below. For each set of comparisons, enrichment plots for all codon mutants within a codon (example in Figure 3.24), and for all amino acid mutants within a codon (example in Figure 3.20), are generated for every codon, ordered from 5' to 3' of the genome (Figure 4.3 A & C).
Figure 4.3. PushpaScripts Generated Default Outputs. Example output plot files automatically generated when PushpaScripts is run on a comparison between Sindbis in vitro transcribed RNA versus the virus from passages 2 to 5. Plots for amino acid mutants and codon mutants for every site ordered from 5' to 3' of the genome (A) and ordered as billboards (B) are produced as a folder with individual site PDFs, a zip file containing all individual site PDFs, and as singular PDFs with each site as a separate page (C and D). The compiled PDFs that are ordered by codon number (C) allows the user to conveniently scroll to and browse their genome region of interest. The compiled billboard PDFs, on the other hand, allow the user to identify specific sites that best answer one of the questions described in the text (D). The example shown results from the analysis pipeline described in figure 4.2 B and shows the most detrimental synonymous mutations in order, that allowed us to quickly identify that synonymous mutations in the codon right next to the nonstructural protein start codon were the most detrimental of such mutations throughout the genome in the aforementioned comparison.
Billboards are lists that answer the following six questions using only statistically significant mutants, upon selection (comparison between the pre-selection and post-selection samples), available as both ordered codon plots (Figure 4.3 B & C) and as tables with all available columns of information. Selection during packaging, i.e., comparison between the intracellular and extracellular viral RNAs, is used as an example. Question 1: What are the most enriched (adaptive/tolerated) codon changes that change the viral protein amino acid upon selection? Eg., what are the most enriched non-synonymous mutations outside, which are selected for while packaging? Question 2: What are the most de-enriched (detrimental) codon changes that result in an amino acid substitution upon selection? Eg., what are the most enriched non-synonymous mutations inside, which are selected against while packaging? Question 3: What are the most enriched (adaptive/tolerated) codon changes upon selection that do not change the viral protein amino acid? Question 4: What are the most de-enriched (detrimental) codon changes that retain the viral protein amino acid upon selection (Figure 4.2 B)? Eg., what are the most enriched synonymous mutations inside, that are selected against while packaging, perhaps indicating important RNA nucleotides for packaging (Table 3.3)?

The above four questions look at specific codon-to-codon mutations.
Questions 5 and 6 look for evolutionary "Do Not Disturb" sites, as Chris Mason named them, in which any change is detrimental, indicating importance of retaining the consensus codon or amino acid. Question 5: At which positions are there no other tolerated codons that change the amino acid upon selection (all mutant codons are most de-enriched)? These are where even the most enriched of all possible amino acid mutations are heavily detrimental, and for example would include enzyme active sites that cannot tolerate any other residue. One must be cautious here, however. Positions with mutants that yielded zero counts in the post-selection sample even at high coverage, indicating that it was likely detrimental, nonetheless does not allow a ratio to be calculated and this codon will be excluded from the "Do Not Disturb" list. There are also sites where a somewhat tolerated mutation to a similar amino acid will mask detrimental effects of the majority of other mutants. As such, it is prone to bias and fundamentally an incomplete list, but perhaps a starting point to delve in further. Finally, question 6: In which positions are there no other tolerated nucleotide substitutions when the amino acid is held constant upon selection (all synonymous mutations are most de-enriched)? This analysis is also subject to the problems listed for the previous question, and additionally positions that only have one synonymous mutation possibility will be heavily represented. Questions 5 and 6 should be treated as starting points to subsequent analyses.
All codon plots and amino acid plots ordered by codon number (Figure 4.3 A) and all billboard plots (Figure 4.3 B) are generated in three formats: (i) a folder containing individual PDF plots in the respective order such that the user can quickly pull out and share data for a single site, (ii) a zip file that compresses and contains all the individual files within the aforementioned folder, and (iii) singular PDF files with every plot ordered as pages (Figure 4.3 C & D) such that the user can open one file and scroll through without the annoyance of having to individually open multiple PDFs. If the user is archiving the data onto an external spinning disk hard drive, it is recommended that the folders with individual plot PDFs are deleted, as thousands of PDF files significantly slows down copying and adds wear and tear on the drive. These folders can be recreated by extracting the zip files at any time.

By default, PushpaScripts performs analyses on Sindbis virus, Toto1101 strain. To perform new analyses on this strain, the user may simply replace the data in the /data folder and assign the new sample names and desired comparisons at the top of Builder module (Figure 4.4 A). Analysis of a different strain or virus would require the user to modify the nucleotide sequence and coding regions within the Sindbis module (Figure 4.4 B). The various modules with the ‘x_figure’ prefix provide examples of analyses performed for generating the various plots in chapter 3. These modules may be used to generate similar plots, or can be used as a guide for the user to
develop any type of analyses, filtered data tables or plots that he or she may desire.
Figure 4.4. Builder and Sindbis Program Inputs. A) The user adjustable area of the Builder program module is shown. In this example, the in vitro transcribed RNA dataset and virion RNA datasets from passages 1 to 5 are computed and added to the database. Then the data from passages 2 to 5 are combined and compared with the in vitro transcribed RNA data, for which statistics are computed and added to the database. The nicknames input should refer to the renamed CirSeqScripts output files and should be used consistently throughout PushpaScripts. B) A part of the user adjustable area of the Sindbis program module. Nucleotide ranges must have the starting nucleotide and the first nucleotide of the last codon within the brackets. These can be changed for other strains of Sindbis or other viruses. The individual regions and features are not necessary for program function, and can be added or subtracted as desired, but allow quick access to analyses within certain regions without having to look up the nucleotide numbers each time. Shown is the python 2.7 code opened within the free PyCharm Community Edition software, which provides a clear programming and bug identification environment and is included within CircSeqLinuxHD.
A)

```python
# Individual datasets to compute and add to database
passages_to_process = ["ivt", "pass1", "pass2", "pass3", "pass4", "pass5"]

# Datasets to combine into a single dataset. Eg. Combine passages 2 -> 5
combinations = [["pass2", "pass3", "pass4", "pass5"], "pass2345"]

# Performs and generates database with statistics for indicated comparison
comparisons = [["ivt", "pass2345"]]

# Nicknames used for convenience, and the full datasets they represent
name_dictionary = {}
name_dictionary["ivt"] = "Sindbis Totoll01 In Vitro Transcribed RNA"
name_dictionary["pass1"] = "Sindbis Passage 1 Low MOI Supernatant Virion RNA"
name_dictionary["pass2"] = "Sindbis Passage 2 Low MOI Supernatant Virion RNA"
name_dictionary["pass3"] = "Sindbis Passage 3 Low MOI Supernatant Virion RNA"
name_dictionary["pass4"] = "Sindbis Passage 4 Low MOI Supernatant Virion RNA"
name_dictionary["pass5"] = "Sindbis Passage 5 Low MOI Supernatant Virion RNA"
name_dictionary["pass2345"] = "Sindbis Passages 2 -> 5 Low MOI Supe RNA's Combined"
```

B)

```python
class Sindbis:

    # Lists of Proteins and Whole Genome Ranges.
    protein_list = ["nsp1", "nsp2", "nsp3", "nsp4", "capsid", "e3", "e2", "6k", "el1"]
    genome_list = ["5p", "nsp1", "nsp2", "nsp3", "nsp4", "26s", "capsid", "e3", "e2", "6k", "el1", "3p"]

    # Dictionary of Coding Sequences of Genome to Analyze.
    nt_ranges = {}

    # Full Polyproteine Coding Forms
    nt_ranges["nsp123_minus_stop"] = [60, 5747] # up to excluding stop codon
    nt_ranges["structural"] = [7647, 11381]
    nt_ranges["nonstructural"] = [60, 7598]

    # Full Protein Coding Region
    full_coding_region = nt_ranges["nonstructural"] + nt_ranges["structural"]

    # Individual Regions and Features
    nt_ranges["nsp1"] = [60, 1679]
    nt_ranges["nsp2"] = [1680, 4100]
```
PushpaScripts was developed with two primary philosophies throughout, after several attempts which did not ascribe to these philosophies, which nonetheless served as learning experiences. Philosophy #1: raw counts are used for calculations whenever possible, and pre-calculated ratios are not used for downstream calculations whenever possible. Examples of raw nucleotide counts are shown in Table 4.2 A. The reasoning behind this is simple, as espoused by Jason Banfelder, currently at Rockefeller University, during his lectures in my Quantitative Biology class: 1 heads in 2 coin flips and 1000 heads in 2000 coin flips produce the same ratio, but the latter is obviously more trustworthy as a fair coin. Using the Fisher exact test on 2x2 contingency tables of raw counts, as recommended by Jason Banfelder and Zak Singer, generates the odds ratio (~enrichment) between two frequencies, but also harnesses the full power in those raw counts for statistics. This power is lost if the frequencies are pre-calculated and then used for enrichment values, as is oft done, including by me in previous iterations of the program. Whenever possible, enrichment values in comparison of datasets and statistics (p-values, confidence intervals etc.) is calculated using this method, and deviations are made consciously and after much discussion that other approximations are appropriate.

Philosophy #2: all data in a database that is operated on remains in the
database, unless it produces confusion with new data added; in other words, new columns of computed data are added to existing columns of data instead of replacing them. For example, computing a frequency does not replace the raw counts that it was calculated from in that table or any downstream tables, such that raw counts are still accessible downstream as is required for the first philosophy. Earlier iterations of the program functioned akin to a handheld calculator - calculations are only performed as needed during analysis, previous values are cleared from memory as new values are computed, and information is 'built up' during analysis. PushpaScripts, in contrast, performs most calculations we can think of initially to build a giant database, and analysis is mostly a function of filtering down from previously computed data. This was easier in theory than in practice and data organization becomes an increasingly difficult battle. For example, Sindbis virus Toto1101 strain has 11,703 nucleotides, and information on the three possible nucleotide mutations in each position yields 35,109 rows of data. A row in each of the nucleotide, amino acid and codon databases yields ~100,000 rows, and an estimated 30 columns in each yields ~3 million individual pieces of data per sample in the computer's memory. Each of these ~3 million data points must be indexed and matched to the appropriate individual data point in all other samples such that calculations are only performed between appropriate data. Whereas there are multiple methods for organizing data in programming, this must also occur very quickly and efficiently, since slight differences over
millions of calculations can easily become unusably slow or fill up a computer's entire memory bank and cause a crash. After testing and implementing many time and memory optimization 'tricks', mostly discovered on StackOverflow message boards, the current version of PushpaScripts seems be able to handle most data wrangling, filtering and analyses pipelines we've run with ease.

4.4. PushpaScripts: A Brief Description of Major Program Modules.

Herein the directory structure and various programs that comprise PushpaScripts are listed and summarized (Figure 4.5). The 'app_info' folder contains various general descriptions of PushpaScripts, including many of the figures in this chapter. Perhaps the most useful file within is an excel file named Directory.xlsx that contains the necessary inputs and additional information for every function in every module (Table 4.3). The 'data' folder contains the required input files that are output from CirSeqScripts (Table 4.2) and nothing else. The 'data_output' folder contains stored database files generated by PushpaScripts and any analysis tables output by the user. Plots generated within python are also in this folder. The 'rscript_output' folder contains all plots that are output using R and ggplot2, including all the ones shown in Figure 4.3. This remainder of this section assumes basic knowledge of various python 2.7 data structures such as modules, classes, functions, lists and dictionaries.
Figure 4.5. PushpaScripts Program Files. The folder directory of the PushpaScripts program. TXT files indicate manuals, PYTHON files indicate python 2.7 code, and .r files indicate R and ggplot2 code. Sindbis_sequence.txt contains the consensus sequence of the Toto1101 strain as a continuous string with no spaces, and should be replaced if the user is analyzing a different strain or virus. Very brief summaries of the various programs and subfolders are described in the text below, and far more information can be found as comment lines within the code itself and in the accompanying manuals.
Table 4.3. Directory of Functions and Required Inputs. A snippet of the function directory within the 'app_info' subfolder, Directory.xlsx, is shown for the DataFrame Input/Output program module. The functions are numbered such that the function and relevant comments can be easily found within the python file itself with the search function. The number of columns after the function name represent the number of expected inputs, and the right parenthesis marks the end of expected inputs. () indicates no additional inputs, for example in the manual function #8, man(). The expected inputs are described, along with abbreviated format: str = string, obj = object, int = integer, flt = float, bool = boolean, dc = dictionary, df = dataframe, etc. An equal sign represents the default value used if no input is provided in that column. This directory provides a useful reference when applying any of PushpaScript's various functions.

<table>
<thead>
<tr>
<th>Module Functions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># Name</td>
<td>Inputs</td>
</tr>
<tr>
<td>D Combination Functions</td>
<td></td>
</tr>
<tr>
<td>6 output_dictionary(</td>
<td>input_dictionary  filename (str)</td>
</tr>
<tr>
<td>7 load_dictionary(</td>
<td>filename (str)</td>
</tr>
<tr>
<td>8 man()</td>
<td></td>
</tr>
<tr>
<td>E Sort and Crop Functions</td>
<td></td>
</tr>
<tr>
<td>9A sort_dataframe_by_column_highest_first(</td>
<td>input_dataframe  column (str)</td>
</tr>
<tr>
<td>9B sort_dataframe_by_column_lowest_first(</td>
<td>input_dataframe  column (str)</td>
</tr>
<tr>
<td>10 head_dataframe(</td>
<td>input_dataframe  number_of_rows = 20 (int)</td>
</tr>
<tr>
<td>11A sort_dictionary_by_column_highest_first(</td>
<td>input_dictionary  column (str)</td>
</tr>
<tr>
<td>11B sort_dictionary_by_column_lowest_first(</td>
<td>input_dictionary  column (str)</td>
</tr>
<tr>
<td>12 head_dictionary(</td>
<td>input_dictionary  number_of_rows = 20 (int)</td>
</tr>
</tbody>
</table>
The **Process** module contains the Process() class which comprises of a variety of variables and functions that primarily bind CirSeqScripts output files into processed dictionaries (**Figure 4.6**). It takes as input the q20threshold.txt file and mutationrates_MLE.txt file (both renamed) for a single passage and processes nucleotide-position (base) tables. It can compute codon and compile amino acid tables from the base table using genome information from the **Sindbis** module (**Figure 4.4 B**), which takes a few minutes for the entire coding region. It can use DataFrameIO module to write all tables or read codon and amino acid tables for faster processing and compatibility with R. It also has filter functions to filter any of the tables (in dictionary format) according to nucleotide position or any column value.
Figure 4.6. Process Program Module CirSeq Data Flowchart. A flowchart of the per-sample segment of the PushpaScripts database populated by the Process() class within the Process module. Initially various information on each nucleotide mutation is computed from raw counts and mutation rates files output by CirSeqScripts (Table 4.2) and added to the database. Then genome-specific information on coding regions is imported from the Sindbis module and codon-specific information is computed and added. Degenerate codons are then compiled and amino-acid specific information is computed and added. These tables are stored and cannot be further manipulated. For manipulation, an empty modifiable table can be loaded with any of these full genome or coding region tables, and a variety of functions can then be applied to filter these data to the user’s preference. Examples of filter functions are shown in green. Black boxes indicate the code used to generate or filter these tables, which is listed in the function directory (Table 4.3).
The **DataFrameIO** module comprises a variety of functions and classes that facilitate the transfer of information between Python, R, and other softwares by using Pandas. It has a variety of converters between dictionaries ⇨ Pandas dataframes ⇨ csv and feather files that work for dictionaries of lists of equal lengths. It also converts columns of strings between ascii ⇨ unicode for compatibility. It also has two transpose classes that rearrange certain values into columns (eg. mutation types) of unequal lengths and also process them into dataframes and output files. All information in PushpaScripts is output as both .csv files, which can be input and browsed in a variety of common programs, and as feather files. Feather files are binary table files that were developed specifically for interoperability between R and Python, and can only be read within these programs [5]. All tables loaded within PushpaScripts uses feather files for speed and consistency. The SQL programming language apparently provides additional database features, but the speed and convenience of feather files has made it seem unnecessary, but perhaps it will be implemented one day.

The **Compare** module, described above (**Figure 4.2 A**), composes of a variety of classes and functions that compile and compare (enrichment, difference) two datasets and binds related statistics. This uses several statistics modules, which strip down and organize statistical test functions from various scientific python packages (i.e., SciPy) for speed and convenience. The **MyFisher**
module contains functions that perform the Fisher Exact Test on contingency tables. The **MyStats** module is used for general statistics on a list of values, such as SD and SEM. The **MyDistro** module comprises of a variety of functions and classes that performs statistics calculations and basic plotting on a large distribution of values. It has functions to calculate the histogram, parametric density and non-parametric density modes and plots for a list of values. It also has classes that apply the above functions for general analysis of lists and coverage analysis of datasets, and is used for generating PushpaScript's default sequencing run quality plots. It can also interface with **R MLE** module, developed by Ashley, to calculate maximum likelihood estimates [1].

The **SiteCompare** module has a variety of classes and functions that can compare codon and amino acid mutants across the genome. It contains the Billboards() class that creates the enrichment-sorted tables described above. It also contains the PlotSites() class that interfaces with the **R amino acid barplot** and **R codon barplot** modules to generate all the plots and files shown in **figure 4.3**. The **Fitness** module calculates relative fitness values using the classical genetics equation shown in **figure 3.28** and populates whole genome tables. The **Chimera** module converts these values into a format that can then be input into the Chimera molecular structure visualization software to map fitness values onto 3D structures of viral proteins (eg. **Figure 3.30**) [4]. By
default, any set of values can be mapped onto all the structures shown in chapter 3. Mapping additional structures for Sindbis or other viruses requires manual input of select .pdb structure file parameters as described in comments within the module. Additionally, the MemoryTest module contains a variety of functions for timing computations and calculating computer memory usage, that was used for optimization. All of the above program modules are interdependent and are used by the Builder program module to populate the database, and care must be taken to not break other modules when changing a core function within. Finally, the various modules with the x figure prefix use the above programs to generate many of the plots and analyses shown in the prior chapter, and can serve as examples of PushpaScripts usage and further analyses. None of the above program modules are dependent on functions within figure modules, and they are independent and may be freely modified by the user without the risk of introducing bugs elsewhere.

4.5. CirSeq Analysis Software Installation Guidelines.

On a computer with a fast CPU processor, running CirSeqScripts on raw sequencing files can take anywhere from several hours for a MiSeq dataset to several days for large HiSeq datasets, during which the computer cannot be put to sleep or interrupted, and the CPU thread runs at 100% for the entire duration. Therefore, using a dedicated computer for the duration of these alignments is optimal. A desktop is preferred, since the extra space available
generally allows for better CPU cooling and less CPU throttling. In PushpaScripts, after the database is initially populated for several hours, subsequent analyses usually only take a few seconds to a few minutes, and as such a modern laptop would be adequate. There is no requirement that both CirSeqScripts and PushpaScripts reside on the same computer, as long as the appropriate files output by CirSeqScripts is moved to the computer with PushpaScripts.

Neither CirSeqScripts nor PushpaScripts utilize CPU multithreading effectively in their current versions. This is due to a problem within the python programming language called the Global Interpreter Lock, and efforts to overcome this issue were met with failure and led to duplicated or disordered data rows. Consequently, single threaded CPU performance is the most important parameter when choosing a computer to run these analyses. A computer with CPU that has high single-threaded performance benchmarks is optimal for these analyses (Table 4.4) [6]. For example, the same MiSeq sample processed with CirSeqScripts took over four days on my 2010 Apple Macbook Pro 15", but took less than half a day on my desktop with an Intel i7-4790K CPU. CPU multithreaded performance and GPU (video card) performance does not have a major effect on processing speed. The number of CPU threads determines the number of samples that can be optimally run in parallel, but most contemporary CPUs have more threads than the number of
CirSeq samples that are typically submitted to a single sequencing run.

Additionally, due to the high number of read and writes performed while running these programs, using a solid state drive instead of a spinning disk hard drive is also highly recommended.

**Table 4.4. CPU Processor Single Thread Benchmarks.** A computer's single threaded performance is the most critical parameter that determines the speed of CirSeqScripts and PushpaScripts, due to being built using python. Shown is the top segment of a list of CPUs with the fastest such performance as of the time of this writing. Price does not correlate with single threaded performance at all, and tracks more closely with multi-threaded performance. Consequently, these scripts can run very fast on rather affordable computers with the appropriate CPU. There are many CPUs within the top 20 that are under $200. This list frequently changes - for example, when I built a desktop specifically for CirSeqScripts in 2013, I chose the CPU that was at first place on this list, an Intel i7-4790K. As of today, 71 processors with better single-threaded performance have since been released. Adapted from [6].

<table>
<thead>
<tr>
<th>CPU</th>
<th>Single Thread Performance</th>
<th>Updated March 2020</th>
<th>Price (USD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMD Ryzen 7 PRO 3700</td>
<td></td>
<td></td>
<td>3,027</td>
</tr>
<tr>
<td>AMD Ryzen 9 PRO 3900</td>
<td></td>
<td></td>
<td>3,019</td>
</tr>
<tr>
<td>Intel Core i9-9900KS @ 4.00GHz</td>
<td></td>
<td></td>
<td>2,992</td>
</tr>
<tr>
<td>AMD Ryzen 9 3950X</td>
<td></td>
<td></td>
<td>2,982</td>
</tr>
<tr>
<td>AMD Ryzen 7 3800X</td>
<td></td>
<td></td>
<td>2,978</td>
</tr>
<tr>
<td>AMD Ryzen 9 3900</td>
<td></td>
<td></td>
<td>2,952</td>
</tr>
<tr>
<td>AMD Ryzen Threadripper 3960X</td>
<td></td>
<td></td>
<td>2,941</td>
</tr>
<tr>
<td>AMD Ryzen 9 3900X</td>
<td></td>
<td></td>
<td>2,939</td>
</tr>
<tr>
<td>AMD Ryzen Threadripper 3970X</td>
<td></td>
<td></td>
<td>2,936</td>
</tr>
<tr>
<td>Intel Xeon E-2278G @ 3.40GHz</td>
<td></td>
<td></td>
<td>2,935</td>
</tr>
<tr>
<td>AMD Ryzen 5 3600X</td>
<td></td>
<td></td>
<td>2,911</td>
</tr>
</tbody>
</table>

If one wants to install and run CirSeqScripts and PushpaScripts on their own computer and operating system, package managers have made installation
onto any computer far easier than when I started graduate school. Mac OSX or Linux is highly recommended, as all the code was written to work on these OSs, and running them on Windows would require all the slashes to be changed to the other direction, along with numerous other tedious edits. The most efficient way would be to install the python 2 version of Anaconda with the GUI option, that automatically installs most required python packages [7]. Then one should install R Studio via the Anaconda Navigator GUI, which will install most necessary R packages. These remaining packages should be installed using conda in the terminal, and the exact commands can be found on Anaconda's webpage: feather-format, memory_profiler, pympler, r-ggplot2, r-feather, and r-cowplot. Installing PyCharm via the Anaconda Navigator GUI is also recommended, as the settings will automatically interface with these Anaconda environment packages. PDFtk Server command line tools must then be downloaded and installed [8]. Finally, instructions within the CirSeqScripts folder should be followed to compile it specific to the user's operating system before running it for the first time.

However, the easiest method would be to use the CircSeqLinuxHD OS, that is an OS distribution built on Linux Ubuntu 16.04 LTS (Figure 4.7). As described above, all the scripts used for CirSeq analysis are already compiled and installed, along with every software dependency and additional useful programs. A 40 gigabyte (GB) bit-by-bit copy of the drive that was used for all
bioinformatics described in this thesis dissertation is provided as an .img image file. The download link to this file can be obtained from the Rice lab, or from me directly (pmambrose@gmail.com). One can simply acquire a solid state disk larger than 40 GB, and that includes every such disk currently available for purchase, and copy the image file onto it. The remaining space on the drive can be partitioned for data storage. Upon connecting the drive to the computer and booting from it, the user will be ready to go for bioinformatics on CirSeq data. Copying the OS image onto the drive can be easily done from within Linux or Mac OSX with the 'dd' program [9]. With the target drive connected, the following command can be run in terminal after navigating to the directory containing the CircSeqLinuxHD OS image file:

```
sudo dd if=circseqlinux.img of=/dev/disk#
```

The # must be replaced with the number of the disk that the image is to be copied onto, that is assigned to it by the OS it is connected to. Getting this number wrong will completely destroy whichever drive that number is assigned to, and can wreck total havoc. As such, dd must be used with care, and using an unimportant, extra computer is recommended. I have only used dd for this purpose, since dd was also used to create the OS image, but there are many safer .img file writing programs that may also work equally well.
**Figure 4.7. The CircSeqLinuxHD Operating System.** An image of the desktop of the CircSeqLinuxHD, which can be used by copying the available OS image file onto preferably a solid state drive and connecting it to a computer intended for CirSeq analysis. Perhaps the OS image can be deployed as a virtual system on an existing operating system, though we have not attempted this. CirSeqScripts and PushpaScripts can be accessed on the desktop, and CirSeqScripts can be run using the terminal command indicated in the text. Additional useful commands are listed in the 'useful_commands.txt' file on the desktop. PyCharm contains several accessory scripts within, and Ashley's relative fitness computation program (Figure 3.29) is shown in the PyCharm window [1]. R Studio is preinstalled for analysis and plotting using R and ggplot2. The Ubuntu system monitor program can be used to indicate memory leak bugs and terminal program crashes. A download link to an image of the OS is available via the Rice lab or by emailing me at pmambrose@gmail.com.
4.6. Chapter IV References.


Chapter V. Perspectives and Future Directions.

5.1. Perspectives.

When the RNA of Sindbis virus was first sequenced, the existence of minor variants was a problem to be overcome. When sequencing the subgenomic RNA, the authors were confident; "The method we have used [ensures] there is no chance of selecting a minor variant in the population" [1]. However, a few years later while sequencing the genomic strand, two cDNA's complementary to the same viral RNA region had a mismatch; "It is of interest that these two sequences agreed perfectly with the exception of a single nucleotide.. we believe the difference may have arisen from cloning a minor variant in the RNA population. In any event it is clear that [there is] at most a very low level of mistakes" [2]. Sampling random variants was an annoyance since the goal was to acquire only the consensus sequence; "Every effort has been made to ensure that this [consensus] sequence is accurate.. the [consensus] nucleotide found at any position is the majority nucleotide present in the population of cDNA copies produced by reverse transcriptase, and the sequencing is insensitive to random reverse transcriptase errors, cloning artifacts, or minor variants in the virus RNA population" [2]. This insensitivity to what lay beyond the consensus sequence was an important and often touted strength at the time. Several technological advances have occurred since, and this perceived weakness of being "sensitive to random reverse transcriptase errors, cloning artifacts, or minor variants in the virus RNA population" is now
a potential feature of modern sequencers. We were especially curious as to what these "minor variant[s] in the virus RNA population" looked like, and perhaps there was even genetically useful information in there. CirSeq allowed us to isolate these variants from random reverse transcriptase errors, cloning artifacts, and also from NGS sequencer-specific issues such as basecalling mistakes [3]. What lay beyond the consensus sequence certainly turned out to be both interesting and useful to us.

Using CirSeq, we discovered that viral RNA in vitro transcribed from infectious clone plasmids is a diverse population of such minor variants. Many virologists have had experiences that made them suspect that this was the case. For example, while working on infectious clones of the Asibi strain of yellow fever virus, Charlie noticed that one clone produced around one hundred-fold less virus than was expected from experiences with the 17D strain. Curious what the mutation could be that was causing this, he sequenced the clone and found a termination codon within the reading frame! There was a base missing in the infectious clone within the NS2A region of the Asibi template, in a stretch of adenines, resulting in a frameshift and a stop codon in the alternative frame. Therefore, the consensus viral genome could not undergo the initial translation events required to begin replication. It seemed most likely that at a frequency of somewhere around one in a hundred, the SP6 polymerase polymerase slipped and added an extra base to the in vitro
transcribed RNA, restoring the original reading frame and viral protein expression such that the genome could be replicated. We proved that *in vitro* transcribed RNA is indeed a diverse population of minor variants by using CirSeq to produce the first highly accurate mutation profile of such RNA, which was transcribed from the Sindbis Toto1101 infectious clone (*Figure 3.2*).

Additionally, the fact that *in vitro* RNA polymerase errors create a diverse population of viral RNA led to a wonderful collaboration on hepatitis B virus (HBV) with Drs. William (Bill) Schneider and Yingpu Yu in the Rice laboratory [4]. HBV chronically infects over a quarter of a billion people globally, which leads to over a million deaths annually, and there are no efficient cures available. The general premise of this study is that by using *in vitro* transcribed RNA to launch infection of hepatitis B (HBV), which is a DNA virus, in addition to other advantages, creates a large diversity of HBV mutants (*Figure 5.1*). Based on this principle, Bill and Yingpu created a cell culture system that can be used to measure the efficacy and drug resistance profiles of HBV antiviral treatment candidates. The synthetic RNA was transcribed using the error-prone T7 RNA polymerase, and we confirmed using CirSeq that most possible single-nucleotide mutations were already present in this population as minor variants. Launching infection using this RNA allows the sampling of a large diversity of mutants in response to selective pressures - for example, adding a drug would only allow drug-resistant mutants, which are already present as
minor variants, to rise in frequency and be identified (Figure 5.1). Indeed, the system was able to predict the two most common drug-resistant mutants that arise in patients with chronic hepatitis B given the drug lamivudine, which was in common use over the past two decades. This paper is in submission, and these data can be seen at the time of writing this document (March 2020) in the BioRXiv pre-print titled "An RNA-based system to study hepatitis B virus replication and select drug-resistance mutations" [4].

![Figure 5.1. Harnessing in vitro Transcribed RNA Viral Variants to Identify Drug Resistance.](image)

In our study, similarly, the presence of these Sindbis variants in the in vitro transcribed RNA allowed us to see how they respond to selection pressures faced when they are introduced into cells and tracked through multiple passages (Figure 3.8). Intuitively, mutations in critical sites of the virus should be filtered out since they would not be able to complete replication steps, and
adaptive mutants which can replicate better than the consensus genome should increase in frequency within the population. Comparison of variant frequencies in the post-selection and pre-selection samples, using bioinformatics developed for CirSeq and described in the previous chapter, recapitulated fitness expectations in historically well-studied sites, wherein many mutants were constructed as infectious clones and tested by Sindbis researchers (Figure 3.20 & 3.21). However, previously known sites in Sindbis are largely based on studies of (i) temperature-sensitive mutants, (ii) mutants in regions macroevolutionarily conserved among different alphaviruses which disrupt a conserved function, and (iii) mutants screened to produce a desired phenotype. Consequently, most known sites are biased towards ones in which mutations produce the aforementioned phenotypes, which are described in detail in the first chapter. Our method allows the analysis of the entire genome in an unbiased way. We focused on highly detrimental synonymous mutants, which should indicate important features in the viral RNA, since these were the most difficult to discover previously. Studies of such features have been mostly restricted to macroevolutionarily conserved nucleotide regions, and our data also recapitulates growth rate expectations of historically studied mutants within these conserved regions (Figure 3.25). Just a couple of synonymous mutants outside of these regions which arose during phenotypic screens have been previously identified. One such synonymous mutation was in the codon next to the opal stop codon of the virus, which made the virus ts in mosquito
cells, indicating that the nucleotide context adjacent to the stop codon was important to the stop codon's function. In our data, the two most detrimental synonymous mutants across the whole genome were adjacent to the nonstructural polyprotein start codon, and suggests and the nucleotide context surrounding the start codon is also important for viral function (Figure 3.24). Even just within this comparison of \textit{in vitro} transcribed RNA and passaged virion RNA, these findings haven't even begun to scratch the surface of predicted variants of interest that have been previously missed. The software developed described in the previous chapter generates and makes available plots and tables of variants within every codon and amino acid site, in order scrolling across the genome and also ranked by most or least enriched, and by synonymous or non-synonymous (Figure 4.3). An investigator interested in a particular region may use these tools to ascertain general fitness of mutants in the region and to inform construction of mutants likely to have a large differential growth phenotype. The list of highly adaptive and highly detrimental mutants also provide good fodder for further study to elucidate the basis of such fitness differences.

Some of the historically best studied sites on the Sindbis genome affect the cleavage of the viral nonstructural protein into individual proteins, and studies of mutants near these sites uncovered an elegant transcriptional regulatory mechanism that is described in detail in the first chapter. Despite these efforts,
only several substitutions of the conserved glycine penultimate to the cleavage sites have been studied. Our data recapitulates the previous viability observations of these mutants, but also provides predictions of novel mutations in the penultimate sites and also in all sites near where the cleavages occur (Figure 3.20). These cleavages affect the amounts and ratios of the various species of viral RNA, but only valine substitutions have been examined for effects on RNA. Additionally, most of these studies were conducted using radiolabeling which precludes analysis of early infection; "The first event in RNA replication is the synthesis of a full-length minus strand on the incoming genomic RNA template. This occurs during the 1st hour after infection, but it is not seen in labeling studies. [only at] 3 hr postinfection, RNA synthesis is easily detectable, and both plus and minus strands are being made" [5]. Dr. Zakary (Zak) Singer and I, based on a lunch discussion, sought to utilize his graduate school experience with single-cell RNA FISH microscopy of overlapping RNA species with some of these CirSeq-predicted cleavage mutants, to analyze a wider range of transcriptional regulatory outcomes than had been done previously [6]. We developed a method that combines RNA FISH and immunofluorescence to individually detect and quantify viral plus, minus and subgenomic RNA strands in ssRNA and dsRNA forms, in single cells, which allows acquisition of spatial and temporal information starting at very early infection (Figure 5.2). Numerous discussions with Drs. Brandon Razooky, Andrea Branch and Charlie helped to refine these
ideas, and this project has evolved into quite an elegant study of Sindbis early infection dynamics, which was an absolute pleasure to be a part of. Please keep an eye out for Zak's manuscript, which will be submitted shortly, if the reader is interested in these results. This was the first time I can recall that brainstormed ideas from a lunch discussion that I was a part of, which were not on anyone else's radar at the time, materialized into an exciting project that has generated very exciting results. Consequently, the next section - future directions - is presented with slightly more confidence than would be otherwise.

Figure 5.2. Sindbis Single Molecule RNA Immuno-FISH Allows Single Cell Imaging of Viral RNA Species. Our Immuno-smFISH method differentially detects single stranded RNA using RNA FISH probes and double stranded RNA using a fluorescent dsRNA antibody [7]. Genomic plus strands are exclusively detected by green probes complementary to the 5' end, and minus strands exclusively by yellow probes in the middle of the genome. The 3' end of the plus strands that is shared by the genomic and subgenomic RNA's are detected by the red probes. These strands are then sorted into genomic and subgenomic strands based on colocalization with the green probes which exclusively bind to the genomic strand, using tools developed by Zak during his graduate studies.
This method provides unprecedented resolution of early infection events and alphavirus transcriptional regulatory events. Adapted from figures provided by Zakary Singer.

We were also interested in comparing the above CirSeq datasets to intracellular viral RNA. Only a small fraction of Sindbis genomic RNA inside the cell is encapsidated and exits the cell in virions, representing a bottleneck with associated selective pressures. We were curious how the viral variant population adapts to these selective pressures. However, sequencing intracellular viral RNA posed a problem; extracellular virions largely only contain viral RNA and thus purified virions can be extracted into pure viral RNA. Sindbis infected cells only contain around 5% viral genomic RNA during a maximal infection, and this cannot be separated from surrounding cellular RNA's with the same ease. Therefore, total Sindbis infected RNA in an equivalent sequencing run would only produce about 1/20th of viral RNA reads as our extracellular RNA samples, precluding the quality of data we had come to expect. Our successful attempt to overcome this problem by developing an oligonucleotide hybridization-based RNA capture method is detailed in the second chapter. With CirSeq of what was now essentially pure intracellular viral RNA, in comparison with the extracellular viral RNA using bioinformatic tools described in the previous chapter, we identified dynamics that are likely beyond mere stochasticity, which indicated that not all viral variants inside the cell are equally likely to make it to the outside (Figure 3.35).
Right around the time the Sindbis capture method was developed, an epidemic of Zika virus hit the world, and Yingpu and Dr. Mohsan Saeed recruited me to also similarly purify intracellular Zika virus RNA. Zika virus does not produce enough virions to practically extract the amount of pure viral RNA necessary for many NGS library preparations. Moreover, the amount of RNA inside the cell is a lower fraction than Sindbis and is only around 1% of total RNA. The fun, excitement, and rollercoaster that was working with Mohsan and Yingpu to push this 1% to 99% purity, described in the second chapter, remains one of the highlights of my time in graduate school (Figure 2.1A). In addition to other assays, this allowed us to perform CirSeq on Zika virus Puerto Rico strain for the first time, which showed that it is a very highly diverse population (Figure 2.38). This additional diversity is likely because the virus strain was continually passaged from its isolation instead of being generated from an infectious clone and passaged only a few times. Adapting the method to Zika convinced us that this oligo capture method is broadly adaptable to any RNA species. Additionally, the final protocol was faster, easier and cheaper than initially expected. Consequently, it is quite a powerful tool for subsequent analyses that would benefit from target RNA purity, and one that several lab mates and additional laboratories have already started using, and have already adapted to other RNA viruses in unrelated virus families. It allows the comparison of intracellular and extracellular RNA’s for any type of comparative RNA analysis, for viruses that produce enough
extracellular virions for such analysis. For viruses that do not produce enough virions, enriched intracellular RNA is the only way to get fairly pure viral RNA for such analyses.

Figure 5.3. Study of Various RNA Modifications Benefit from RNA Enrichment. Shown are a variety of different RNA modifications that are being increasingly discovered in many RNA species, including in the genomes of many RNA viruses. These mechanisms are currently being unraveled by the field of epitranscriptomics. Most of this information is lost when RNA is reverse transcribed into DNA as a requirement for target PCR amplification. Consequently, these analysis benefit from increased purity of the target RNA within the total RNA pool, and our oligo capture method has been extensively used to boost signal for these collaborative studies described in the text. Adapted from [8].

The analysis of RNA modifications is one such application that benefits from having pure target RNA, since this information is usually lost during reverse transcription and subsequent amplification. A multitude of RNA modifications
are being increasingly discovered in various RNA species, including in RNA viral genomes (Figure 5.3) [8–12]. Consulting on purifying viral RNA's to better discover and profile these modifications in collaboration with Stephan Bluethgen and Dr. Inna Ricardo-Lax in the Rice laboratory and Noah Alexander, Daniel Butler and Alexa McIntyre in the Mason laboratory, among others, has been an absolute delight. For further details on such modifications on a variety of RNA viruses and their effects on viral replication, Inna's upcoming publication will be of interest to the reader.

5.2. Future Directions: Sindbis Virus.

Even within the best studied mechanisms in Sindbis, such as the nonstructural protein (nsP) cleavage mechanism and use of the opal codon for translational regulation, there are several long-standing gaps in our understanding that would be nice to address, which would require quite different tools than the ones we have used and developed thus far. When the opal stop codon a little upstream of the nsP3 and nsP4 cleavage site is read through to encode the nsP4 polymerase, which amino acid is inserted into that site, if one is inserted at all, is unknown. Various theories are presented in detail in the first chapter. This readthrough occurs at far higher frequencies than RNA mutations to sense codons, and thus is beyond the purview of nucleotide sequencing technologies. It would be interesting to subject this region to protein sequence analysis to determine what residue lies in its place, if one does. If the protein
sequencing method is sensitive enough, it would be of interest to see if the same amino acid is always used during translational readthrough, or is there a population of various residues at this position, when the RNA has the stop codon. These would again be likely different in mosquito cells compared to vertebrate cells. It is increasingly being discovered that cells produce and utilize a variety of stop codon readthrough protein products [13]. Thus, further study of this one example could reveal mechanisms relevant to translational biology as a whole. Additionally, as described previously, as a consequence of the stop codon there are two forms of nsP3 - one that stops at the stop codon and one that undergoes translational readthrough and is cleaved several residues downstream. The effect that this additional region has on the protein is unknown. In some cases, single amino acid changes can completely abrogate or add protein functionality in certain contexts, and in other cases large insertions are tolerated. It would be interesting to place these additional amino acids within this range of phenotypes, and see how much, if at all, this contributes to the viability differences in mutants with altered ratios of these two forms. Perhaps the potential discovery of a ts mutant or alternative phenotype within this short stretch by the screen described above could be of assistance.

Even Sindbis nonstructural protein cleavage events, which control translation regulation, are not very well understood during infection. As described in detail
in the first chapter, the sequence of these events, the ratios of cleaved and uncleaved proteins, and the cis and trans preferences of the Sindbis protease during infection were largely inferred from studies performed in cell-free in vitro translation or trans-expression systems. The rates of trans polyprotein cleavage, which play a central role in the current model of nsP cleavage wherein increasing concentration of the polyprotein leads to faster processing of it (Figure 1.22), could perhaps be quantified by using a trans-cleavage fluorescent reporter system [14]. Ideally, 3D structural information of the nsP's themselves while a part of active replication complexes would be attained. Perhaps single molecule fluorescence resonance energy transfer (smFRET) techniques, similar to ones used often in studies from Dr. Scott Blanchard's laboratory, can be harnessed [15]. Briefly, the smFRET method allows the "detection of biological interactions in terms of motion.. represented by changes in distance between individual particles or domains to which donor and acceptor fluorophores are site-specifically attached" [15]. It would be interesting to attach a fluorophore donor and acceptor on adjacent nsP's, which if tolerated, would indicate when they are still adjoined in a polyprotein and when they are cleaved and in individual nsP forms, under the highly likely assumption that cleavage leads to conformational differences in their interaction. Such studies performed on each pair of nsP’s can reveal spatial and temporal information on the ratios of cleaved and uncleaved proteins within replication complexes during active infection by the wild type virus and
viable cleavage mutants. Designing an smFRET system would benefit from 3D structures of the nonstructural proteins in cleaved and uncleaved forms such that molecular dynamics simulations can be performed to inform fluorophore placement sites [16, 17]. That in itself is quite an undertaking which would provide quite a few benefits to our understanding of the virus, but with recent advances in cryo-electron microscopy, such efforts are more in reach than ever before [18–22]. Combining these data with information on various RNA ratios using our immuno-smFISH method will create a far fuller portrait of the elegant transcriptional regulatory system employed by Sindbis virus and other alphaviruses.

I was given the idea to structure this section as a response to the question, "if I had to start my PhD again knowing what I know now, would I still do it similarly?" The short answer to that is yes - much of what I'd like to do builds on prerequisite findings and depends upon many of the tools we've developed over this time. Therefore, the remainder of this section will address a slightly modified version of that question; "If I had, at the beginning, all the tools in my arsenal as I do now, what are some additional studies I would have liked to have performed?" Our pipelines were built to rapidly analyze the behavior of all minor variants that result from single nucleotide mutations as the viral population responds to selective pressures. The following paragraphs describe additional comparisons I consider to be interesting.
The first chapter provides numerous examples of the utility of ts mutants in determining much of what is known about Sindbis genetics. However, most of these studies are based on about half of around 50 ts mutants originally isolated which had the least reversion rates, and consequently likely to be less detrimental than the other half that were thrown away (Table 1.5). Much of what is known about Sindbis, and alphaviruses in general, are biased towards the mechanisms disrupted by this singular collection of ts mutants. As such, the next comparison I would like to do, potentially beginning as soon as this document is submitted, is to perform an unbiased genome-wide screen of viral populations at both permissive and non-permissive temperatures. Given the results of a transposon-based random mutagenesis screen which induced a number of ts mutants while mutagenizing a small segment of the viral genome, we predict that such comparison would find many, many new ts mutants [23]. Mutants in proteins that are not well represented by existing ts mutants - nsP3, E3, 6k and TF - would be most interesting and useful. Additionally, 'reverse' ts mutants, which are defective at the normally permissive temperature but fine at the typically non-permissive temperature would be a cool discovery. Our bioinformatics pipeline can find these easily, if they exist, and if they present a phenotype beyond the random noise threshold. This experiment can be done by simply infecting chicken embryo fibroblast cells with passaged virus that we have already analyzed by CirSeq, at a permissive temperature of around 28°C.
and a nonpermissive temperature of around 40°C (Figure 5.4). Then performing CirSeq on the resulting viral RNA and comparing them using our pipeline should generate a list of variants which are affected by temperature, which could provide mutants in understudied regions and mechanisms which can be turned on and off by simply changing the incubation temperature. Additionally, most $ts$ mutants were generated using a variety of mutagens to increase the rate of $ts$ phenotype production (Table 1.6), and some of these are thought to be biased toward certain types of mutations (eg. nitrosoguanidine preferably inducing $G \rightarrow A$ mutations). It would be interesting to quantify and confirm the effect that these various mutagens have on the viral variant population.

Sindbis virus is an arbovirus with a natural vertebrate and mosquito host cycle, and must contend with highly diverse physiological environments and defense mechanisms while replicating well enough for transmission. Consequently, performing CirSeq on mosquito cells and comparing the viral population between vertebrate and mosquito cells, and also profiling the dynamics of this population as infection passages are alternated between these hosts, is a rather obvious idea that I’ve been thinking about for a while. Mosquito cell culture infections generally produce far less Sindbis virions than vertebrate cell culture. For our studies on extracellular virions, we needed to grow Sindbis virus in multiple 500 cm$^2$ plates of confluent vertebrate cell monolayers to get
microgram-quantities of viral RNA for CirSeq, and scaling that up for mosquito cells was not practical. However, our oligo capture method makes acquiring pure intracellular RNA from within mosquito cells rather trivial. We suspect that certain regions with higher diversity in our vertebrate cell datasets, such as the 3' UTR, will be subject to much higher selection in mosquito cells (Figure 3.14). Ideally, I envision this as being conducted in mosquito cells at 28°C in addition to the ts mutant study above, such that we would have temperature-matched controls for comparison of viral populations in both vertebrate and mosquito cells (Figure 5.4). If this mosquito cell infection is done in C7/10 mosquito cells, we can also perform a similar ts study by comparing the population grown at the typical 28°C versus 34.5°C, which is as high as mosquito cells can handle [24]. Then we can generate a list of previously undiscovered ts mutants which are ts in both cell types, and also those which are only ts in one host and not the other, which would provide interesting tools for further study. A natural and far less practical expansion of this study would be to compare Sindbis virus from live infections of its primary natural hosts - say, a transmission between Culex mosquitoes that feed on infected birds such as crows and so on (Table 1.3). This would give a better picture of how these viral populations look "in the wild," where they are less cell-culture adapted and have to contend with whole-organism physiology as well. No further comment will be made on how to go about this¹.

¹ This was an actual reviewer comment on a grant submission, wherein we did have to
Comparison of vertebrate and mosquito cells somewhat provide a *de facto* comparison of how the virus population handles different immune systems. However, to specifically look at the effect of the immune system in these hosts on the viral population, it would be preferable to use cells with more intact immune systems. The cells generally used and described above mostly have some immune system defects which allow the virus to replicate highly and produce enough virions to meet the minimum RNA threshold of our subsequent analyses. Using intracellular viral RNA purified via oligo capture allows far more leeway of cell types that CirSeq data can be acquired from. In mosquito cells, comparison of viral populations in cells competent and defective in the RNA interference pathway, such as using an isogenic cell line with and without knockout of AGO2, would be of interest (*Figure 5.4*) [25]. In innate immune-competent vertebrate cells, this experiment can be simply performed with and without the addition of interferon. Ideally, this experiment would be performed on a cell line with and without knockouts of portions of the immune system of interest (e.g. knockout and reconstituted STAT1 mouse embryo fibroast cells) (*Figure 5.4*) [26, 27]. Perhaps a comparison of these CirSeq datasets will reveal differences in codon bias and dinucleotide frequency tolerances, which have been shown to be influenced by various

---

comment on the impracticality of handling virus-infected mosquitoes and febrile birds within a laboratory situated in Manhattan, New York.
genes in the vertebrate innate immune system [28–31]. I, however, cannot speak more to this without doing additional reading on interferon-stimulated genes. Using the visualization methods developed by Zak described above would also be of interest to quantify the relative amounts of dsRNA, which is thought to trigger the innate immune system, to the various ssRNA species, and to also gain insight into spatial and temporal differences.

Figure 5.4. Potentially Interesting Comparisons of Viral Populations Using CirSeq. Illustration of various ideas presented in the text for comparing genome-wide minor variant populations that arise during infections in different conditions. Comparison of chicken embryo fibroblasts at 28°C and 40°C and alternatively mosquito cells at 28°C and 34.5°C will reveal ts phenotypes in each respective cell type. A comparison of the populations between the aforementioned chicken cells and mosquito cells at 28°C will provide a temperature-controlled mimic of the predominant natural viral cycle of transmission between birds and mosquitoes. Similarly, a human transmission cycle can be mimicked by comparing mosquito cells grown at their normal temperature, 28°C, to human cells grown at normal human body temperature,
37°C. Human innate immune system effects can be illuminated by comparing populations in immune-competent cells with added interferon, or alternatively in such cells with and without knockouts of interferon-stimulated genes such as ZAP or STAT1. Similarly, the effect of invertebrate RNA interference on viral populations can be probed by comparing mosquito cells with and without knockouts of components in the RNAi pathway such as AGO2. Y-axis represents temperature, and x-axis represents organism the cell line is derived from. Yellow cells represent invertebrate cells and red cells represent vertebrate cells. Protein structures adapted from [32–34].

5.3. Future Directions: Arboviruses.

We primarily chose Sindbis for our analyses because (i) it makes a lot of virion RNA in cell culture, and (ii) the large history of study which allowed us to compare, adjust and build our bioinformatic analyses pipelines. Additionally, as Charlie likes to say, "it is safe for graduate students to use", and not requiring a Biosafety Level 3 (BSL-3) facility saves quite a bit of time and effort [35]. Now, however, having both the bioinformatic pipelines and the oligo capture enrichment method available allows us to expand all of the above undertaken and proposed analyses to any number of arboviruses, which do not produce as much virion RNA nor have quite a long history as Sindbis, as shown by our Zika virus Puerto Rico strain data (Figure 2.38). During a visit to our campus, Dr. Anthony (Tony) Fauci, currently head of the National Institute of Allergy and Infectious Diseases (NIAID) at the National Institutes of Health (NIH), told me over dinner that there have been two constants for him throughout his 50 years of experience; (i) many of his friends, having grown up
in Brooklyn, are still annoyed that the Dodgers moved to Los Angeles, and (ii) there has been, and it's highly likely that there will always be, a new global outbreak every few years from a virus that we have no chance of predicting.

Tony's team at the NIAID in Bethesda, Maryland, state; "About 700,000 deaths due to vector-borne diseases occur globally each year.. In recent years, the Americas have witnessed a steady stream of.. emerging or reemerging arboviruses, such as dengue, West Nile, chikungunya, Zika, and Powassan, as well as increasing numbers of travel-related cases of various other arboviral infections" [36]. Many of these occurred just within my time in graduate school [37, 38]. There will surely be more to come [39].

In terms of clinically relevant alphaviruses, which are related to Sindbis virus and likely share many common mechanisms with Sindbis virus, chikungunya and eastern equine encephalitis viruses (EEEV) have caused recent epidemics as of the time of this writing [36, 37]. My friends who work in horse farms have been quite aware of EEEV for a while and the annual equine vaccination schedule for it, but a recent human outbreak that was larger than in previous years occurred just a few months ago and raised some alarms [40]. Tony's team at the NIAID describes it's impact; "In the summer and fall of 2019, nine U.S. states have reported 36 human cases (14 of them [~39%] fatal) of one of the deadliest of these diseases: eastern equine encephalitis.. This year’s EEE outbreaks may thus be a harbinger of a new era of arboviral
emergences.. a sobering thought, given the high case-fatality rate of diseases such as EEE” [36]. Comparison of chikungunya and EEEV populations in mosquito cells versus human cells, and even in equine cells for EEEV, would be of medical and veterinary interest. As far as I am aware, there were only a couple reported ts mutants of EEEV, which have not been mapped and don't seem to be still available, and none for chikungunya. Performing the ts mutant screen described above for these viruses can generate a list of mutants whose replication cycle defects can be switched on and off. Comparison of mechanistic ts defects with overlapping well-studied ones in Sindbis and Semliki Forest viruses will reveal how similarly they operate, which can illuminate the how these severely pathogenic viruses have differentially evolved to meet the requirements of their unique environments over large swaths of space and time. Similar CirSeq studies on clinically relevant flaviviruses, which we had initiated in Zika virus, can be performed for the related dengue, West Nile, Powassan, Kyasanur Forest and yellow fever viruses, which continue to be global threats [37, 41–46].

Once CirSeq data exists for multiple alphaviruses and flaviviruses, preferably grown in the same cell type, we believe it would be worth the effort to build a pipeline that incorporates and analyzes all these minor variant microevolutionary data from across viruses within a family. These will likely require statistical methods that are beyond what are used for CirSeq analysis
thus far, which only perform comparisons using a single consensus sequence. Much like the rapid increase in knowledge of viral genetic features that occurred once multiple alphaviruses were first sequenced and even just a few genomes were aligned, as described in the first chapter, adding this new dimension to these alignments should accelerate further discovery. I can only imagine the predictive power that could be wielded by not only having information on the conservation between the consensus sequences at each aligned site, but also the mutational tolerances of all single-nucleotide variants at all of these sites. Such a combination of macroevolutionary data and microevolutionary variant tolerance data is almost guaranteed to reveal many new and interesting findings, and I would be willing to bet nearly anything I own on it. Additionally, a pipeline to compare CirSeq data between different consensus sequences could also be applied to different strains within the same virus. It would be of interest to compare strains isolated in the wild to cell culture adapted strains of a virus to further illuminate the underpinnings of cell culture adaptation, as for example, between Sindbis patient-derived strains or the original mosquito-derived AR339\textsuperscript{2} strain to the cell culture adapted Toto1101 strain. As a corollary, perhaps most interesting would be to compare virulent strains with vaccine strains of the same virus to examine attenuation mechanisms, such as between the Asibi patient strain and 17D vaccine strain

\textsuperscript{2} Sindbis AR339 is not truly a wild isolate like more recent isolations, and was passaged in infant mouse brains for several generations before the viral stock as assigned as the strain (Table 1.2).
of yellow fever virus, or any of the other arboviruses listed in the following table (Table 5.1).

Table 5.1. Several Attenuated Vaccine Strains of Clinically Relevant Arboviruses. The left column lists a clinically relevant and highly infectious arbovirus that requires, at minimum, a BSL-3 facility due to its dangerous outbreak potential. The right column lists attenuated vaccine strains of each respective virus which can be handled in BSL-2 facilities. Some of these vaccine strains are used in humans and animals, and some of these vaccine strains are primarily used in laboratories due to the additional convenience of BSL-2 facilities (common) as opposed to BSL-3 (rare) or BSL-4 (very rare). CirSeq comparison of virulent and vaccine strains can potentially inform directed attenuated vaccine design. Adapted from [35].

<table>
<thead>
<tr>
<th>Virus</th>
<th>Vaccine Strain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chikungunya</td>
<td>181/25</td>
</tr>
<tr>
<td>Rift Valley fever</td>
<td>MP-12</td>
</tr>
<tr>
<td>Venezuelan equine encephalomyelitis</td>
<td>TC83 &amp; V3526</td>
</tr>
<tr>
<td>Yellow fever</td>
<td>17-D</td>
</tr>
<tr>
<td>Japanese encephalitis</td>
<td>14-14-2</td>
</tr>
</tbody>
</table>

Many of these vaccines and others were empirically designed by simply injecting into humans a virus stock that has been passaged in the laboratory for a long time and then, as in the case of prisoners and other human "volunteers" injected with yellow fever 17D strain, challenging them with injection of the virulent strain to see if they survive or not [47]. It took a while, but we as a society have now generally accepted that such direct empirical experimentation of vaccine efficacy on humans is rather unethical.
Comparison of CirSeq data between virulent and attenuated vaccine strains may reveal interesting evolutionary mechanisms behind such attenuation, which can potentially inform directed vaccine design. Additionally, whereas a select few arboviruses have vaccines for humans, none have any effective antiviral treatments. Development of an antiviral drug screening and resistant mutant profiling system for arboviruses, analogous to Hepatitis B one by Bill and Yingpu mentioned above, could potentially lead to tremendous benefits for human health [4]. Additionally, as exemplified by developments in Hepatitis C treatments in the past several years, any increase in our understanding of the life cycles of these viruses continues to increase our chances of finding effective therapies [48].

5.4. Future Directions: Additional Applications.

There have been various efforts toward using Sindbis virus as a vector for gene therapy and transgene expression, and even a webpage during the very early days of the internet was built by Dr. Henry Huang at Washington University in St. Louis to promote this idea [49, 50]. Peter Bredenbeek, now at Leiden University Medical Center in the Netherlands, as a postdoc in Charlie's lab developed such a vector system in the 1990's whereby the structural and nonstructural polyproteins of a Sindbis virus infectious clone was split into separate plasmids (Figure 5.5 A) [51]. These had the advantage that "The highest levels of expression of heterologous products are achieved
when the viral structural genes are replaced by the heterologous coding sequences. Such recombinant RNAs are self-replicating (replicons) and can be introduced into cells as naked RNA" [51]. Consequently, separate structural protein containing 'helper' can be used to package the replicon RNA, and "These helpers should be useful for applications in which expression of the viral structural proteins or virus spread is not desired" [51]. Based on these plasmids, an interesting recent application called viral evolution of genetically actuating sequences (VEGAS) was recently developed in 2019 by Bryan Roth’s laboratory at the University of North Carolina School of Medicine (Figure 5.5 A) [52]. Briefly, the above replicon system was adapted to subject mammalian transgenes to continuous directed evolution within a 'native' mammalian cell environment. Not unlike the basis of many experiments above, the transgenes themselves accumulate mutations based on the viral polymerase's error rate, and the mutation frequency continues to increase since the requirements of the viral life cycle shouldn't impose selective pressures on this transgene (Figure 5.5 B). This system was then used to successfully evolve a (i) transcription factor, (ii) a G-protein coupled receptor, and (iii) llama-derived nanobodies toward various functional goals [52]. Interestingly, after presenting an early part of my project at one of my annual departmental retreats a few years ago, I was approached by several membrane protein physiologists wondering if Sindbis virus and its high error rate could be used to perform directed evolution on mammalian membrane
channels and transporters. Now it looks like it can, and using such viral vector systems for directed evolution can impact many different projects in a broad variety of fields.

Figure 5.5. Transgene Directed Evolution using Sindbis and RNA Pol II Proofreading Fidelity Measurements. A) Schematic of the Sindbis VEGAS system, utilizing an RNA replicon containing a desired transgene to be subjected to continuous forward evolution, and a helper which aids in packaging the replicon into virions, both derived from the full length Sindbis genome. B) Mutation frequencies within this transgene increased over time due to continuous input of viral errors and the lack of negative selection. C) Kinetic measures of RNA Pol II, wherein longer pausing times (y-axis) represents higher incorporation of the wrong nucleotide, is shown by WT Pol II (blue) and a Pol II trigger loop mutant (red). X-axis represents 1 mM of additional rNTP.
added to a subsaturating pool of rNTP’s, to promote mismatching driven by excess nucleotide. Most studies of proofreading ability in the field is similarly conducted by using kinetics as a proxy for fidelity, and CirSeq allows direct measurement of mutation frequencies in synthesized RNA. Adapted from [52, 53].

Whereas RNA virus polymerases do have the highest mutation rates in biology, one limitation of only using the naturally occurring minor variants as the raw material for evolution is that single-nucleotide mutants are very highly represented. Unless there is very robust positive selection, double mutants within a codon or even nearby essentially occur at frequencies which can be estimated by multiplying the frequencies of the two single mutations together, which results in a frequency exponentially lower than either. This limits the nucleotide or amino acid contexts that can be sampled in most experiments described above. For example, within a codon, single nucleotide mutants result in a maximum of 9 possible amino acid mutants which could be sampled, and often less. These rates can somewhat be increased by using a variety of mutagens. Performing CirSeq on such libraries would be of interest as several of these mutagens, such as ribavirin, are also used in patients to treat viral infections and is often a component of combination antiviral therapies. Ribavirin is a nucleoside analog that can get incorporated by the viral polymerase and act as an RNA mutagen, among other activities (Figure 5.9 A) [54]. At least some of ribavirin’s antiviral efficacy is thought to derive from increasing the error rate and mutational load of viral replication. Further
probing the mechanism of such mutagens could perhaps assist in the design of the next generation of antiviral drugs.

These rates can also be increased by synthetically producing a library of viral plasmids (i.e., infectious clones or replicons) with mixed nucleotides at determined positions, and then transcribing the viral RNA from it. These libraries can be constructed by ordering 'selectively randomized' oligos and swapping them into the viral backbone plasmid. This provides the advantage that all possible codons and therefore all possible amino acids can be sampled at each protein position, or alternatively, a variety of nucleotide contexts can be sampled by randomizing multiple degenerate codon positions enabling identification of cis-acting RNA elements [55, 56]. Within the Rice laboratory, Bill and Yingpu have implemented such a method to sample more variants in the aforementioned hepatitis B drug screening platform [4]. I would be curious to apply the same method to understudied proteins and regions in the Sindbis virus genome. Given our finding that synonymous single nucleotide mutations do not greatly disrupt recognition and packaging of the viral RNA into the nucleocapsid and there is a fair level of redundancy (Figure 3.35), it would be interesting to perform a similar CirSeq analysis between intracellular and extracellular viral RNA while randomizing large regions of the packaged genome with degenerate codons. Alternatively, this analysis could also be performed in a trans-packaging system similar to those used for replicon
packaging (Figure 5.5 A). Such a system could divorce the effects that these mutations would have on general replication from just the packaging function, and would allow greater leeway in the amount and types of mutations that can be sampled. Analyses using this randomization method, however, is suboptimal for whole genome analyses unlike those that depend on naturally occurring minor variants or those using slightly increased frequencies via addition of mutagens. Use of such synthetic randomized libraries is best done within relatively short regions, as the complexity of the library rapidly increases beyond the number of molecules produced thereby reducing sampling, and also has practical oligo synthesis and cost limitations. Consequently, there is an art to achieving this balance of mutation frequencies.

Another limitation of CirSeq analysis is that due to the 80-100 bp fragment size limit, linkage analysis beyond that size can only be inferred using statistical predictions. For CirSeq datasets with suspected linkage events, long read sequencers available from Oxford Nanopore or PacBio can be used to supplement these data [57–59]. All mutations which have linkage will need to be at frequencies above each machine’s respective noise threshold to be detected, which are at present quite higher than Illumina machines due to higher inherent sequencing error rates (Table 5.2). These error rates will likely continue to decrease as these technologies are further developed. Additionally, Oxford Nanopore sequencers are now capable of directly
sequencing RNA to produce long reads. Even though the direct RNA sequencing error rate is quite high at the moment, this presents tremendously exciting opportunities for acquiring information on structural variants in viral RNA populations, such as recombinant genomes and truncated defective interfering particles, and information described above that is lost when RNA is reverse transcribed into DNA (Figure 5.3) [60].

Table 5.2. Brief Comparison of Common Current Next-Generation Sequencing Platforms (2020). Comparison of an Illumina MiSeq, Oxford Nanopore MinIon, and Pacbio RSII sequencing platforms on sequencing performed primarily on the same extract of Methicillin-resistant Staphylococcus aureus DNA genomes [61]. Illumina platforms are used for CirSeq analysis, and the short read lengths preclude linkage analysis. We use the HiSeq platform when much higher coverage than output by a MiSeq is desired. Platforms with higher read lengths currently have much higher sequencing error rates. These technologies are rapidly evolving and many NGS companies and technologies have came and went during my time in graduate school; these values are highly likely to be quite different in the future. Data compiled from [61, 62].

<table>
<thead>
<tr>
<th>2020 NGS Technologies</th>
<th>MiSeq</th>
<th>Pacbio RSII</th>
<th>MinIon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Read Length</td>
<td>300</td>
<td>10,000</td>
<td>10,000</td>
</tr>
<tr>
<td>Maximum Read Length</td>
<td>300</td>
<td>40,000</td>
<td>2,272,580</td>
</tr>
<tr>
<td>Relative Genome Coverage</td>
<td>72</td>
<td>130</td>
<td>12</td>
</tr>
<tr>
<td>Basecalling Error Rate</td>
<td>0.002</td>
<td>0.008</td>
<td>0.03</td>
</tr>
<tr>
<td>Normalized Error Rate</td>
<td>1</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>Basecalling Insertions Rate</td>
<td>2.00E-06</td>
<td>0.008</td>
<td>0.107</td>
</tr>
<tr>
<td>Normalized Insertions Rate</td>
<td>1</td>
<td>4,000</td>
<td>53,500</td>
</tr>
<tr>
<td>Basecalling Deletions Rate</td>
<td>1.00E-05</td>
<td>0.007</td>
<td>0.115</td>
</tr>
<tr>
<td>Normalized Deletions Rate</td>
<td>1</td>
<td>700</td>
<td>11,500</td>
</tr>
</tbody>
</table>
By profiling the mutational profile of its output RNA using CirSeq, we have analyzed the mutation rates of SP6 and T7 bacteriophage polymerases in vitro and Sindbis viral polymerases during infection, but what about cellular RNA polymerases? How do they compare? Eukaryotic cells utilize three RNA polymerases; Pol I which transcribes the 18S, 28S and 5.8S rRNA species, Pol II which is responsible for mRNA transcription, and Pol III which transcribes the 5S rRNA, tRNA's and a few other RNA species, such as Alu elements in primates. Pol II is taught to utilize proofreading to reduce the mutation rate relative to the other RNA polymerases, but the extent to which it does has not been quantified [63]. Additionally, mutations in Pol II's trigger loop decreases this fidelity [53]. However, the majority of these studies have performed using polymerase speed as a proxy for fidelity, measured using optical methods or by measuring synthesized RNA, since pausing has been shown to be indicative of RNA polymerase proofreading activity (Figure 5.5 C) [64–66].

Instead of using a proxy, CirSeq allows us to directly measure the mutation frequencies in synthesized RNA, and the comparison of RNA Pol fidelities can be performed by CirSeq by differentially profiling RNA species synthesized by each. Oligo capture can be used to easily enrich a specific target if that would be preferred. For example, a specific mRNA that is known to be transcribed
from one copy in the DNA genome can be enriched and sequenced from cells containing WT Pol II versus a Pol II fidelity mutant to determine the effects of this proofreading capability on specific mutation rates. Similarly, virioids transcribed by Pol II in plant cells would be quite intriguing. Whereas the mutational frequency of target rRNA molecules can acquired easily, determining the mutation rates would be more difficult for Pol I and III due to the high number of ribosomal DNA copies in eukaryotic genomes, and not much is known about biases in transcription from these copies. Therefore, it will be more difficult to assign what percentage of the mutation frequency is due to the error rate of these polymerases and what percentage is due to the rRNA being transcribed from multiple rDNA copies with different sequences, but this is perhaps doable with statistical methods.

I stumbled upon this idea more or less by accident. Before oligo capture worked, intracellular viral RNA was a non-starter, but even virion RNA from cells infected with Sindbis at high MOI's gave a much lower percentage of virus-aligned reads than usual. The most noticeable difference between a high and low MOI infection was in cytopathogenicity; more cells died sooner in high MOI infections. I aligned this CirSeq dataset to various RNA species simply to find what was 'contaminating' my sample, and the majority aligned to rRNA, indicating that ribosomes are being released into the media by dead cells and its rRNA component is diluting out the virion RNA. However, since various
plots and analyses are automatically generated by CirSeqScripts as described in the prior chapter, a gander at these showed more interesting results than I'd expected. I have since been hoping to design and perform this rRNA CirSeq experiment on better samples than partially degraded RNA from ribosomes released from dead cells that have been floating in the media for hours. The experiments proposed above would be of significant interest to those interested in eukaryotic RNA transcription.

5.5. Future Directions: Coronaviruses.

As I'm writing this chapter, we are in the middle of a period of virus discovery that has some differences, but also some similarities, to that period from around 70 years ago described in the first chapter. Active arbovirus surveillance for West Nile virus, using similar tools as previously described, is still occurring across the United States using a combination of finding and checking dead birds, testing mosquitoes, horses and humans, and using sentinel chickens (97% of detected West Nile virus mammalian cases in the USA are from horses) [67–69]. However, at the moment, the world is gripped by panic over a febrile respiratory disease descriptively named coronavirus disease 2019 (COVID-19). The causative agent was discovered just a few months ago and was initially creatively named 'novel coronavirus' (Figure 5.6), and is today referred to as SARS coronavirus 2 (the sequel; stay tuned for the trilogy). Acquiring the genome sequence of the virus upon discovery has
gotten a bit faster, from a gap of three decades for Sindbis virus to just a few days for SARS-CoV-2.

Figure 5.6. SARS Coronavirus 2 Virions. Transmission electron microscope image of SARS-CoV-2 originally isolated from a U.S. patient with COVID-19. Virus particles were propagated in Vero cells and visualized with post-processed with false color to highlight envelope (blue), nucleocapsid (yellow) and genomic RNA (red) density regions. This family of viruses are named corona (i.e., crown in Latin) viruses due to the crown-like appearance imparted by the blue 'spikes'; their spike envelope glycoproteins (S-proteins). Adapted from [70].
COVID-19 has just the other day been announced by the World Health Organization as a pandemic, and graphs of the number of people infected in many countries mimic the familiar exponential curve portion of plaque assay-titered cell culture infection growth curves (Figure 5.7) [71]. Every single headline on the front page of New York Times involves the virus, and all major sports leagues and concerts have been suspended and much air travel is restricted. Many schools, universities and companies have shuttered their doors and moved everything they can to remote access, the trains and subways in New York City are emptier than I've ever seen before, and the public portion of my upcoming defense must be given virtually via teleconference [72]. These measures are taken for good reason - for example, 76% of all current SARS-CoV-2 cases in the state of Massachusetts is linked to a single meeting at the biotechnology company Biogen, which likely consisted of many people with a biology background who would have been assumed have known better [73].
**Figure 5.7. Exponential Growth of People Testing Positive for SARS-CoV-2.** The number of confirmed cases in the past few weeks, as of the time of writing this document, outside of mainland China is shown using linear and log y-axes. The virus has now spread globally, after it spent a few months mostly inside China since it was discovered in November 2019. The outbreak originated surrounding the Huanan Seafood Wholesale Market, a live animal market in the city of Wuhan in the Hubei province of China. The number of cases outside of China is currently increasing by one log every 16 days, and may or may not hit carrying capacity soon (it seems to have passed the inflection point of a logistic growth curve within China). The curve is reminiscent of the exponential portion of characteristic one-step growth curves of viral infections in cell culture wherein viral titers at different timepoints are measured via plaque assay. Adapted from an excellent YouTube video on the mathematics of exponential and logistic growth by Grant Sanderson [71, 74].

Pseudo-information and misinformation abound, people have stopped drinking Corona-branded beer, and local supermarkets have run out of toilet paper for reasons I do not fully understand. A journalist from a major medical news publication, who is certainly more informed than most journalists who are discussing coronaviruses on the news all day at the moment, stated the following in response to a question on an internet forum thread titled 'We are a team of medical experts following COVID-19's progression closely. Ask Us
Anything.'; "We don’t know about mutation of this particular virus yet, but in general, coronaviruses don’t mutate very easily. They contain a unique mechanism called a proofreader that corrects errors in their genetic sequence. That ensures that they will continue to copy themselves correctly each time. It also means they are slow to mutate" [75]. This is wrong³. First, a direct causative link between microevolutionary rates, *i.e.*, polymerase error rates, and macroevolutionary rates, *i.e.*, "slow to mutate" within the above context, for viruses, is suspect. And most importantly, basic awareness of the principles of evolution as espoused by Charles Darwin would indicate that likely nothing on this planet ever has or ever will "make a perfect copy each time" [76]. But just how wrong is it in the context of coronaviruses?

Coronavirus family viruses contain the largest known RNA virus genomes and can reach 32 kb in size, about three-fold the size of Sindbis virus and most other known RNA viruses. Without some proofreading, and with the same mutation rates as other RNA polymerases (*Figures 3.3 and 3.27*) coronaviruses will generate about 6 mutations in each genome copied, which seems to be a high mutational load. Thus, intuition dictates that coronaviruses would need some proofreading to bring mutations per genome to within a more common range, and they were discovered to encode a viral RNA

³ Which is okay; I may have made a few mistakes while sounding authoritative within this document, and I am surely often wrong on many topics outside of RNA virology.
proofreading exoribonuclease, the first of its kind known in RNA viruses (Figure 5.8) [77]. However, these estimates were determined for various coronavirus family members using biochemical methods, kinetic methods as described above for Pol II, Sanger sequencing of multiple clones, and NGS of short fragments without accuracy optimizations which are subject to each platform's inherent noise threshold (Table 5.2, additional description in first chapter) [77]. The accurate mutational profile of a transcribed coronavirus genomic RNA population has not been performed, and can be easily generated using CirSeq. I perhaps would have done this already if I didn't have this document to write. Various proofreading mutants with defects in this exoribonuclease have been shown to have various effects on viral fitness and growth rates, but CirSeq can precisely identify effects on mutation rates using experiments similar to those described above for Pol II fidelity mutants [78, 79]. It would be of interest to profile the viral populations generated by these mutants and compare them to those from the WT virus using the CirSeq pipelines described in the previous chapter. These experiments would answer the following outstanding questions in coronavirus biology; (i) What is the true mutation rate in a WT coronavirus, i.e., how many mutations per genome copied occurs normally in the viral population? How does this compare with those of RNA viruses of typical ~10 kb length genomes, such as Sindbis virus? Does the proofreading ability get the coronavirus mutation rates part of the way there and the population just tolerates a higher mutational load, does
it equalize it with other RNA viruses, or does it push it even lower such that each whole coronavirus genome copy on average has less mutations in it than other RNA viruses? (ii) What is the true effect of the exoribonuclease on this proofreading ability? Does it suppress all specific mutations rates (e.g., G→A substitution rates) equally, or does it specifically recognize and proofread certain types? Do coronaviruses with defective exoribonucleases show the same mutation rates as other RNA viruses as theorized in Figure 5.8, or does the coronavirus polymerase core without a functional exoribonuclease already possess a higher fidelity than other RNA viruses?

Figure 5.8. Estimated Coronavirus Mutation Rates in Comparison with RNA Viruses and DNA Replication. Unlike most other RNA viruses, coronaviruses utilize proofreading to increase their replication fidelity, likely necessitated because coronaviruses have around three-fold the genome length as other RNA viruses. How these estimates were primarily attained is described in the text. Estimated range of mutation rates are shown using dashed arrows for coronaviruses, RNA viruses, and cellular DNA replication as occurs within our cells. For each class of organisms, shown are the inferred contribution to total replication fidelity by nucleotide complementarity during incorporation facilitated by the polymerase core (blue-green), by proofreading activity (green), and by mismatch and excision repair. Nsp14 ExoN indicates the
exoribonuclease proofreading component of coronaviruses. Estimates of mutation rates of additional virus families is shown in figure 1.3 A. Adapted from [77].

If the reader thought that alphaviruses with just one subgenomic RNA species was complicated upon reading the first chapter (as I still do), coronaviruses produce a large multitude of them, which is of interest as they still contain many mysteries [80]. Practically, our oligo capture method can make short work of isolating the genomic RNA from these various subgenomic RNA’s and surrounding cellular RNA for higher sequencing depth, and consequently, more powerful genomic RNA analysis. If certain coronavirus subgenomic RNA’s are of interest, these can be specifically isolated too, with adapted capture methods that Dr. Inna Ricardo-Lax has developed in our lab to specifically isolate Sindbis subgenomic RNA. Our RNA immuno-FISH method (Figure 5.2) can perhaps be used to quantify the ratios of these with single-cell resolution over time and space. Additionally, temperature sensitive mutants have been isolated for various coronaviruses, but not for SARS-CoV-2 as of yet, since the generation of these is low-throughput and time intensive. As with Sindbis virus, as described in detail in the first chapter, these have been tremendously useful in the study of various coronavirus mechanisms such as protease activities [81, 82]. A ts mutant CirSeq screen as described above (Figure 5.4) can rapidly identify ts mutants for SARS-CoV-2, providing additional tools to accelerate the study of this specific global pandemic agent.
Finally, many existing antiviral drugs that shown varying levels of efficacy against other RNA viruses are currently under trials for COVID-19, primarily due to the convenience of already having undergone development and early-stage human clinical trials [83–85]. Some of these, such as remdesivir that was originally developed for treating Ebola, are thought to get incorporated into viral RNA, avoid recognition by the exoribonuclease, and cause elongation termination downstream (Figure 5.9C) [86]. It would perhaps be of interest to see the distribution of coronavirus RNA's in the population with such termination events, and whether they contain additional mutations when the incorporated drug in the template is copied without termination, using a combination of long-read direct RNA nanopore sequencing and CirSeq.
Figure 5.9. Several Drug Candidates in Trials for Coronavirus Disease 2019. A) Ribavirin is a guanosine analog which is used to treat pediatric respiratory syncytial virus infections and viral hemorrhagic fevers including Lassa fever, Crimean–Congo hemorrhagic fever and Hantavirus infection. Prior to sofosbuvir, was used in combination with interferon to treat chronic hepatitis C virus infections. It is often used in cell culture to increase viral RNA mutations. B) Sofosbuvir is a nucleoside analog prodrug that is currently used as a monotherapy or in combination therapies to effectively treat Hepatitis C infection, and has a cure rate of ~98% across all viral genotypes with minimal side effects. C) Remdesivir is an adenosine analog prodrug that was originally developed during the 2014 Ebola virus outbreak. It has broad-spectrum antiviral activity in cell culture infections.

Our old viral RNA mutagenizing friend, ribavirin, is also currently being tested in combination with sofosbuvir, a hepatitis C virus antiviral, and with interferon (Figure 5.9 A & B) [87]. The coronavirus exoribonuclease has been experimentally shown to provide some ribavirin resistance by removing incorporated ribavirin during proofreading, but perhaps it can still work in trials at higher concentrations or in combination with these other drugs [88]. Profiling
the genomic RNA populations of WT and proofreading-defective coronavirus in the presence of varying concentrations of ribavirin using CirSeq would be interesting, and would reveal the extent of ribavirin excision activity on viral mutation rates. Additionally, ribavirin has been used for decades in combination treatments for patients with all sorts of viral infections since its addition generally improves cure rates in clinical studies, but why it has this effect is unknown [89, 90]. The aforementioned experiments on ribavirin in combination with these other drugs that are in current trials as combined therapies for COVID-19 can illuminate whether the combinatorial clinical effect is due to additive effects on RNA mutation rates, or if they act via completely separate pathways. Whereas these CirSeq studies would generally indicate any adaptive mutants above a set statistical significance threshold, I would imagine that the most powerful application would be an implementation of a drug resistant mutant positive selection platform similar to the one Bill and Yingpu developed for hepatitis B virus (Figure 5.1). If, as with hepatitis B virus, we can predict drug resistant mutants to various drug candidates - as monotherapies and in combinations - before they arise in patients given these treatments, it would provide tremendous benefits towards the development of effective antiviral cures for what at this moment looks to be a prominent human health problem for years to come.

In summary of these future directions sections; there is much to do.
5.6. Chapter V References.


549


47. Calisher CH (2013) Lifting the Impenetrable Veil: From Yellow Fever to Ebola Hemorrhagic Fever & SARS.


69. The California Department of Public Health West Nile Virus Website. 
http://westnile.ca.gov/


71. Sanderson G Exponential growth and epidemics. 
https://www.youtube.com/watch?v=Kas0tIxDvrg


74. Sanderson G Who/what is 3blue1brown? 3Blue1Brown, https://www.3blue1brown.com/about

75. r/Coronavirus - We are a team of medical experts following COVID-19’s progression closely. Ask Us Anything. reddit, https://www.reddit.com/r/Coronavirus/comments/fdf5fq/we_are_a_team_of_medical_experts_following/


This appendix serves as a Sindbis virus genome feature index of the comprehensive 1994 review by Drs. Jim and Ellen Strauss:


This appendix is organized by viral protein or untranslated region, in the same order as in the text of the review. The left column lists the genome features, the "S&S Page" column indicates the page in the review where the feature is described, and the "Reference" column lists the associated references from the review. The same reference numbers are used as in the text, and therefore the references section of the review can be used to look them up.

For convenience, the genome features table is followed here by a list of only the references mentioned in the table, while retaining the numbering system from the original text. Features which have been found to have differential phenotypes in mosquito cells compared to vertebrate cells are highlighted in blue, along with the associated page numbers and references.
The review is a terrific source of information on various aspects of alphavirus genomes. However, it was written over a quarter of a century ago now, and is obviously limited to findings from before. Its primary limitation is in the discussion of host factors, which have been the topic of most studies since then. The first chapter of this thesis dissertation can help put chorological events in Sindbis virology in further context.

Reading this review constituted my first 'homework' assignment given to me by Charlie during my rotation. I read the review over that Christmas break at my Grandparents' house, and being quite overwhelmed with the amount of information before I even knew anything about viruses, made this map as an Excel spreadsheet while reading it to help keep track of the various features and which gene or untranslated region it belonged to. Since then, it has been a tremendously useful resource for me as a reference point to quickly get a refresher or read up more on features that I came across during the course of graduate school, and several labmates have found it useful also. I have included it here as a resource that a reader interested in alphaviruses may potentially find useful also.
<table>
<thead>
<tr>
<th>Gene (Structural)</th>
<th>Different in Mosquito (mq) Cells</th>
<th>S&amp;S Page</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PE2 (E2 + E3)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ER Signal (N-terminal, 19-40 aa's?)</td>
<td>500</td>
<td>139</td>
<td></td>
</tr>
<tr>
<td>• Carbohydrate site (conserved asparagine, aa 11-14)</td>
<td>500-501</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Membrane anchor + stop transfer signal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• C-terminal, hydrophobically conserved, length: 26 aa's)</td>
<td>501, fig 8</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>6K insertion signal (33 aa's past anchor)</td>
<td>502</td>
<td>301</td>
<td></td>
</tr>
<tr>
<td>PE2/6K signalase cleavage site (C-terminal alanine)</td>
<td>502</td>
<td>301</td>
<td></td>
</tr>
<tr>
<td>Glycosylation sites: Asn-196 &amp; Asn-318</td>
<td>502</td>
<td>214</td>
<td></td>
</tr>
<tr>
<td>• 4x E2 glycosylation sites mutant: temp-sensitive</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Asp-275 site in ectodomain, 34.5°C prevents budding</td>
<td>502, 508</td>
<td>105, 107</td>
<td></td>
</tr>
<tr>
<td>• Grows well (not ts) in mosquito cells</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• less bulky &amp; uncharged (no sialic acid in mosquitos)</td>
<td>502, 508</td>
<td>105, 107</td>
<td></td>
</tr>
<tr>
<td>• Antibody resistance with new glycosylation site</td>
<td>502</td>
<td>520</td>
<td></td>
</tr>
<tr>
<td>Ectodomain lateral glycoprotein interaction sites</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Mutants: H291→L, A344→V</td>
<td>508</td>
<td>304, 178</td>
<td></td>
</tr>
</tbody>
</table>
**Gene (Structural) | Different in Mosquito (mq) Cells**

<table>
<thead>
<tr>
<th>PE2 (E2 + E3) (<em>Continued</em>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Furin cleavage site (PE2 → E2 + E3): R_(K/R)R sequence</td>
</tr>
<tr>
<td>• Early event in mosquitoes, late event in vertebtrates</td>
</tr>
<tr>
<td>• PE2 cleavage mutants: AR89 &amp; insertion strains</td>
</tr>
<tr>
<td>•• Produced PE2 virions in vertebtrates, not in mosquitos</td>
</tr>
<tr>
<td>•• S1→N cleavage mutant attenuates mice neurovirulence</td>
</tr>
<tr>
<td>••• Cleavage prevented by new glycosylation site (AR89)</td>
</tr>
<tr>
<td>Palmitoylation sites: 4 to 6 total</td>
</tr>
<tr>
<td>• E2 cytoplasmic tail: conserved C’s: C396, C416, C417</td>
</tr>
<tr>
<td>• Palmitoylation occurs post-cleavage in mosquito cells</td>
</tr>
<tr>
<td>•• primarily on E2 rather than PE2</td>
</tr>
<tr>
<td>Phosphorylation of PE2 cytoplasmic tail: conserved TPY</td>
</tr>
<tr>
<td>• required for cytoplasmic reorientation in ER membrane</td>
</tr>
<tr>
<td>Gene (Structural)</td>
</tr>
<tr>
<td>------------------</td>
</tr>
<tr>
<td><strong>PE2 (E2 + E3) (Continued)</strong></td>
</tr>
<tr>
<td>E2 tail capsid binding domain (for budding/enveloping)</td>
</tr>
<tr>
<td>• Some E2 tail mutants were only inhibited in one host</td>
</tr>
<tr>
<td>• aa 403 → 420 (from SINV/RR chimera experiments)</td>
</tr>
<tr>
<td>•• E2 tail-only synthetic peptide binding experiments</td>
</tr>
<tr>
<td>•• Conserved binding structure: conserved Pro's and Cys's</td>
</tr>
<tr>
<td>Cell Receptor binding domain (aa 170-220)</td>
</tr>
<tr>
<td>• G172R mutant enables neurovirulence</td>
</tr>
<tr>
<td>Neutralizing antibody A/B binding domain (aa 172-220)</td>
</tr>
<tr>
<td>Neutralizing antibody C binding domain (aa 62,96,114,159)</td>
</tr>
<tr>
<td>Mice neurovirulence: Q55H E2 mutation important (AR339)</td>
</tr>
<tr>
<td>• Host immunopathology determines encephalitis severity</td>
</tr>
<tr>
<td>• His 55 overcomes bcl-2 inhibition of apoptosis</td>
</tr>
<tr>
<td>• Deletion of aa 55-61 in RR attenuates RR in mice</td>
</tr>
<tr>
<td>Gene (Structural)</td>
</tr>
<tr>
<td>------------------</td>
</tr>
<tr>
<td><strong>E1</strong></td>
</tr>
<tr>
<td>Glycosylation sites: Asn-139 &amp; Asn-245</td>
</tr>
<tr>
<td>• C-terminal chain is simple or complex depending on host</td>
</tr>
<tr>
<td>Membrane anchor + stop transfer signal</td>
</tr>
<tr>
<td>• Upstream of C-terminal stop codon</td>
</tr>
<tr>
<td>• Cytoplasmic domain deletion or change has no effect</td>
</tr>
<tr>
<td>Stop codon (C-terminal)</td>
</tr>
<tr>
<td>Palmitoylation site: C430</td>
</tr>
<tr>
<td>Envelope fusion domain (aa 80-98)</td>
</tr>
<tr>
<td>• Low pH mutants (aa72 &amp; 313) have altered neurovirulence</td>
</tr>
<tr>
<td>• Fusion pH in mosquito cells dependent on growth medium</td>
</tr>
<tr>
<td>Conformation changes require cholesterol</td>
</tr>
<tr>
<td>• Mosquito cells don't produce cholesterol, must intake</td>
</tr>
<tr>
<td>Neutralizing antibody MAb33 binding domain (~ aa G132)</td>
</tr>
<tr>
<td>Gene (Structural)</td>
</tr>
<tr>
<td>------------------</td>
</tr>
<tr>
<td><strong>Capsid</strong></td>
</tr>
<tr>
<td>• Structure</td>
</tr>
<tr>
<td>• Packing morphology</td>
</tr>
<tr>
<td>• smaller nucleocapsids, with DI's</td>
</tr>
<tr>
<td>Proteinase catalytic triad</td>
</tr>
<tr>
<td>Proteinase cleavage site (serine)</td>
</tr>
<tr>
<td>N-terminal domain (disordered, doesn't crystallize)</td>
</tr>
<tr>
<td>• unconserved, high positive charges for RNA binding</td>
</tr>
<tr>
<td>C-terminal domain: cap-cap lateral interaction</td>
</tr>
<tr>
<td>• conserved, determines T=4 icosahedral shell</td>
</tr>
<tr>
<td>E2 binding domain (for budding/enveloping)</td>
</tr>
<tr>
<td>• WEEE = EEE capsid + SINV E1 E2</td>
</tr>
<tr>
<td>Ribosome binding domain (aa 94-205)</td>
</tr>
<tr>
<td>**Gene (Structural)</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td><strong>6K</strong></td>
</tr>
<tr>
<td>• entire protein can be deleted with no effect</td>
</tr>
<tr>
<td>• but point mutations reduce fitness! (obscure)</td>
</tr>
<tr>
<td>Signal sequence for E1 translocation (C-terminal)</td>
</tr>
<tr>
<td>6K/E1 signalase cleavage site (C-terminal alanine)</td>
</tr>
<tr>
<td>Palmitoylation sites: C35, C36, C39</td>
</tr>
<tr>
<td>Gene (Non-Structural)</td>
</tr>
<tr>
<td>-----------------------</td>
</tr>
<tr>
<td><strong>nsP1</strong></td>
</tr>
<tr>
<td>Function: Synthesis of - strand RNA</td>
</tr>
<tr>
<td>• A348→T Mutant: ts, no (-)RNA</td>
</tr>
<tr>
<td>• 4000-7000 (-)RNA strands produced before shutoff</td>
</tr>
<tr>
<td>Capping function: methyltransferase activity</td>
</tr>
<tr>
<td>• S-adenosylmethionine binding site: R87, S88</td>
</tr>
<tr>
<td>• S-am low Km mutant replicates in low methionine mq cells</td>
</tr>
<tr>
<td>Capping function: guanyltransferase activity</td>
</tr>
<tr>
<td>• GTP binding site: Q21, S23, V302</td>
</tr>
<tr>
<td>• Ribavirin &amp; mycophenolic acid resistance: K21, N23, M203</td>
</tr>
<tr>
<td>• Ribavirin &amp; mycophenolic acid thought to reduce cell GTP levels</td>
</tr>
<tr>
<td>nsP1/nsP2 cleavage site (GA*A) by nsP2</td>
</tr>
<tr>
<td>(+)RNA Copromoter: complement on (-), 51 nt stem-loop</td>
</tr>
<tr>
<td>Packaging signal: nt's ~946-1226 (+encapsulation efficiency)</td>
</tr>
<tr>
<td>• RNA size matters (small RNA = low packaging efficiency)</td>
</tr>
<tr>
<td>Gene (Non-Structural)</td>
</tr>
<tr>
<td>-----------------------</td>
</tr>
</tbody>
</table>

**nSP2**

RNA Helicase Domain (N-terminal, aa's 1-459)

- 2 NTP-binding domain motifs: GSGKS @ nt189, DEAF @ nt252
  
Nucleus (N & C-terminal) and nucleolus localization signals

Proteinase Domain (papain-like, C-terminal aa's 474-728)

- Catalytic dyad: C481 & H558 (N609 helps also)
- Cleavage is regulated, polyprotein P123 is functional
  
  - N614→D increased cleavage but was lethal

Many mutants (5) reduce subgenomic RNA relative to genomic

nsP2/nsP3 cleavage site (GA*A) by nsP2

**nsP3**

Function not well understood, opal codon makes 2 forms

- Required for RNA synthesis, F312→S mutant: ts, RNA(-)

Presence in polyproteins P123/P1234 enhances P3/4 cleavage

N-terminal domain: highly conserved (~325 aa's)
<table>
<thead>
<tr>
<th>Gene (Non-Structural)</th>
<th>Different in Mosquito (mq) Cells</th>
<th>S&amp;S Page</th>
<th>Reference</th>
</tr>
</thead>
</table>

**nsP3 (Continued)**

C-terminal domain: highly variable, heavily phosphorylated

- Large deletions are well tolerated 511 297

Opal Codon & downstream C (UGAC tetra-nucleotide)

- Conserved in all alpha’s with max 1 substituted nt
- Adjacent C is leaky (A/G/U terminates efficiently)
  - Readthrough (for P1234) occurs 10-20% of the time 509 300
  - SINV AR86 Mutant: lacks opal codon, only produces P1234 509 484
  - Sense codon mutant more restricted at low MOI than high 518 299

nSP3/nsP4 cleavage site (GA*Y) by nsP2 513, fig 15 486

**nsP4**

Function: RNA dependent RNA polymerase (has GDD motif) 511 241

- ts mutations (3 known) render it RNA(-) 511 181
- ts mutants (4) more restricted in mq than chicken cells 511 283
<table>
<thead>
<tr>
<th>**Gene (Non-Structural)</th>
<th>Different in Mosquito (mq) Cells</th>
<th>S&amp;S Page</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>nsP4 (Continued)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Underproduced than other nsP’s (needs opal readthrough)</td>
<td>511</td>
<td>516</td>
<td></td>
</tr>
<tr>
<td>• decreased readthrough (other codons) impairs replication</td>
<td>511</td>
<td>299, 283</td>
<td></td>
</tr>
<tr>
<td>• UGAC→UGAU mutant more restricted in mq than chicken cells</td>
<td>511</td>
<td>283</td>
<td></td>
</tr>
<tr>
<td>Excess nsP4 rapidly degraded (N-end rule pathway)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• stable nsp4 is within replicase complex</td>
<td>511-512</td>
<td>93</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>**Untranslated Regions</th>
<th>Different in Mosquito (mq) Cells</th>
<th>S&amp;S Page</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>5' UTR</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(+)RNA promoter: complement on (-) (nt's 1→44)</td>
<td>519</td>
<td>370, 386</td>
<td></td>
</tr>
<tr>
<td>5' UTR SINV/RR Chimeras are attenuated</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Suggests viral protein interactions (on (-)RNA also?)</td>
<td>525</td>
<td>273</td>
<td></td>
</tr>
<tr>
<td>Cap hyper-methylation: affects translation</td>
<td>535</td>
<td>216</td>
<td></td>
</tr>
</tbody>
</table>
Untranslated Regions | Different in Mosquito (mq) Cells

| 3' UTR |
|------------------|--------|--------|
| (-)RNA promoter, on (+) (last 19 nt's) | 519 | 272 |
| • Upstream region large deletions tolerated in chicken | 519 | 272 |
| • Deletions are more detrimental in mq than chicken cells | 519 | 271 |
| • 16-19 nt deletion host specific, detrimental in chicken | 519 | 271 |
| • 16-19 nt deletion grew better than wt SINV in mq cells | 519 | 271 |
| High affinity mq host protein binding sites (first 250 nt) | 520 | 389, 390 |
| • Different host proteins in chicken & vero cells | 520 | 367, 390 |
| 3x 40-nt repeated elements | 521 | 387, 485 |
| • Deletion far more deleterious in mq than chicken cells | 522 | 272, 273 |
| • Regulates translation: deleted DIs +replicate -translate | 522 | 360, 362 |
| 3' UTR SINV/RR Chimeras are indistinguishable from WT | 525 | 273 |
| • Suggests host protein interactions | 525 | 273 |
### Intergenic Junction (& Start of Capsid)

Subgenomic RNA promoter: 24 nt’s (SGP) (complement on (-)?)

- Changes within 24 nt can increase/decrease activity
- Some mutants deleterious in mq but not in chicken cells
- Upstream sequence presence increases activity
- Other alpha SGP’s can decrease or increase activity (eg. RRV)

26S: Cap required but insensitive to inhibiting analogs
Recombination can occur at detectable frequencies

<table>
<thead>
<tr>
<th>Untranslated Regions</th>
<th>Different in Mosquito (mq) Cells</th>
<th>S&amp;S Page</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intergenic Junction (&amp; Start of Capsid)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Subgenomic RNA promoter: 24 nt’s (SGP) (complement on (-)?)</td>
<td>519</td>
<td>384, 294</td>
</tr>
<tr>
<td></td>
<td>Changes within 24 nt can increase/decrease activity</td>
<td>520</td>
<td>162</td>
</tr>
<tr>
<td></td>
<td>Some mutants deleterious in mq but not in chicken cells</td>
<td>521</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td>Upstream sequence presence increases activity</td>
<td>520</td>
<td>411</td>
</tr>
<tr>
<td></td>
<td>Other alpha SGP’s can decrease or increase activity (eg. RRV)</td>
<td>520</td>
<td>201</td>
</tr>
<tr>
<td></td>
<td>26S: Cap required but insensitive to inhibiting analogs</td>
<td>535</td>
<td>26, 539</td>
</tr>
<tr>
<td></td>
<td>Recombination can occur at detectable frequencies</td>
<td>544</td>
<td>601</td>
</tr>
</tbody>
</table>

Provided herein for the reader’s convenience are only the references from the aforementioned review that are listed within the “references” column of the preceding table. Numbering and formatting are the same as in the review, such that the reader can look up the reference in either the original text or below:


Appendix B: Extended Acknowledgements.

Charlie Calisher's book described in this dissertation's first chapter, "Lifting the Impenetrable Veil: From Yellow Fever to Ebola Hemorrhagic Fever and SARS", also contains perhaps my favorite preface and acknowledgements sections. Therefore, I figured I ought to similarly better express my gratitude than possible within the front matter, and include a few more details on the people I've described therein who have made my educational journey, and this thesis dissertation, possible. Additionally, I've told my friends that I will show my appreciation of their support by permanently attaching their names to my thesis dissertation, which many of them have been excited about. Charlie Calisher describes the serendipitous event that started his career in virology; "It probably is worth relating here how I found myself as an employee at CDC.. As I had just been liberated from graduate school and needed some time to ponder the rest of my life, I chose to drive south from Washington, D.C., to visit Ocracoke Island, North Carolina.. I erected and arranged my tent on the beach.. Two fellows approached my tent and asked me whether I had previously camped and whether I had actually set up a tent in the past.. They told me I had the triangular tent upside down and that the vent was on the ground.. Taking pity on me, they invited me to join them and their families for dinner.. They were a marvelous, fun-loving and generous bunch.. One of the men said he worked for the Federal Aviation Administration, the other, James L. Goddard, said he worked at CDC.. Jim Goddard turned out not only to work at CDC but was its director.. I wonder what my life would have been like had I erected that tent properly". Interspersed amongst my expressions of gratitude are a few stories of how I ended up in graduate school and on this thesis project and consequently a virologist, via many strokes of serendipity, largely presented in reverse chronological order.
I thank Charlie Rice and Peggy MacDonald and all members of Charlie's lab. My first email to Charlie, before I knew who he was, was to inquire about auditing his Tri-Institutional Virology course during my first year. I only later realized how busy he usually was, but he replied to me immediately, and invited me over for a long chat on viruses before I even took the course, where I told him that I knew nothing about viruses but thought they were cool. I'm certain he had better things to do with his time than to entertain a kid for hours who read a very short layman's book on viruses one summer and thought it was a fascinating read. As I describe below, the course was fantastic, and I asked to rotate, and even though I still didn't know anything about viruses he took me on anyway. Coming in as essentially a blank slate without a relevant background and not knowing anything about viruses, the amount of money I've cost Charlie's lab for my seven years of training and experiments may match or even exceed all that I've cost my parents since birth, and that is not a small amount. There were likely better uses for it, on both Charlie's and my parents' behalves. Also, throughout grad school Charlie would leave the door open to his office all day. I knew that as long as there was no one else in there having an 'official' meeting, which always included after 5 pm or on the weekends, I could just walk in, throw a piece of paper on his desk, with data or a paper I was reading which was confusing or I just thought was cool, and ask for his opinion. I've probably interrupted quite a few trains of thought, and delayed quite a few of his emails and responsibilities. Many of my friends describe meeting their PIs as waiting by the office door akin to a leopard ready to pounce in the small chance that their PIs show up and before they lock their doors, and I never once had that experience. Both my friends’ stories of finding their PIs and the pandemic lockdown has shown what a special resource that was. Peggy was always a tremendously nice and helpful resource for pretty much everything. She has spent so much of her time holding weekly subgroup meetings for many of us students so that we can
build presentation and literature reading experience on a weekly basis, and
discuss our data and experimental plans far more often than otherwise would
be possible. I am so grateful that Charlie and Peggy have decided to hang up
some of my wildlife photographs in their weekend home in Connecticut, and
they continue to share photos of bears and other wildlife captured by the trap
camera in their backyard. Their Aussie Shepherds during my time here, Hattie,
Bruno and Lily, were also so nice and always provided a fun time hanging out.
Santa told me that I was Hattie's second favorite lab member and Bruno's
favorite, which I agree with. Lily's preferences are a bit more mysterious.
Hattie and Lily prefer to roll over and insist on me giving a tummy rub until they
made it known that they were content, and Bruno preferred to chase and be
chased around, sometimes with a tennis ball and sometimes without
depending on his mood. It was nice to know that no matter what is going on
with my science or my day, my first priority was providing tummy rubs and
playing catch, getting my face licked and occasionally providing a banana.

Troels Scheel and Melody Li, who both now have their own labs at the
University of Copenhagen and UCLA respectively, were essentially my initial
lab 'parents', and taught me how to work on viruses on the bench and
 computationally, with tremendous patience and grace. Most of the protocols
and methods listed in this thesis are descriptions of what they taught me in
person, often multiple times until I 'got it'. None of what is in this dissertation
would have been possible without their efforts, and it's been a tremendous
privilege to be their 'apprentice' and continue to stay in touch when we're each
3000+ miles away. Ashley Acevedo, who originally developed CirSeq, has
been so nice and helpful throughout - via email before she joined the Rice lab,
in-person after she joined the lab, and via email and phone after she joined
Myriad Genetics in San Francisco. I still talk to her frequently, mostly to make
fun of her new Tesla vehicle and the speeding tickets she deserves. Hachung
Chung, who now has her own lab at Columbia University, designed excellent Sindbis virus qPCR probes which solved one of my initial roadblocks during my rotation and was really nice and helpful throughout, and gave me my first experience of Korean barbecue. Eating short ribs at all-you-can-eat Korean BBQ places until standing is a painful experience has since become a routine activity. Leonia Bozzacco, now at Regeneron, was a great resource troubleshooting any cell-culture virus-growing issues. Meg Scull, who now has her own lab at the University of Maryland, was really nice and gave me an expensive bottle of streptavidin beads which initiated my oligo capture enrichment optimization process. I likely would've put it off for a while if she hadn't, and after I got that bottle I didn't have that excuse anymore. The timing ended up being perfect for when Zika hit, as Sindbis was just about optimized, and it led to a tremendously fun time collaborating with Yingpu Yu and Mohsan Saeed on optimizing Zika RNA enrichment. Yingpu was always there to chat and help with anything I ever needed, and showed me ways of doing certain things which were absolutely critical for optimization. Many of the methods listed in this dissertation are also descriptions of what he taught me in person. Mohsan, who now has his own lab at Boston University, was also always there to chat and help, and we also had a lot of sports discussions. When I was a child my Grandpa and uncle took me to an India vs Pakistan cricket game live, back when Pakistan was a superior team. More recently, however, India has been a stronger team and those matchups have been more akin to the Yankees playing a college team, so those discussions were a bit one-sided. Mohsan offered to help edit my paper draft, and I thought he'd give me a few suggestions and ask me to send it to the next person. However, he sent it back to me saying, "it's not good enough to send out yet, make these changes and send it back to me", and I did, the paper in slightly better shape than before. He then did that 15 more times. Mohsan has probably spent more time editing my manuscript than I originally spent writing it, when he didn't have to. I
initiated the 'repayment plan' by using my Subaru Outback and helping him move to his new apartment and job at Boston, which was a fun trip. That period of excitement and discovery surrounding Zika virus when the epidemic first hit along with Mohsan and Yingpu was one of the most exciting parts of graduate school.

It goes without saying that Bill Schneider has been tremendously helpful to me, as he is to everyone in the lab as indicated by the fact that he is on every lab member’s acknowledgements slide in their respective talks. This started my very first day in the lab, when he sourced a reagent for me from someone he knew from his graduate school institution. He’s been a tremendous resource for book recommendations and a variety of discussions ranging from the science and history of virology and discovery, to how cool the extraction and forging of metals are. Joe Luna also shares an appreciation of the variety of topics and random facts ranging from the history of virology to metals, and can always be counted on for exciting conversations. Avery Peace was always a source of great conversations on wild animals and the outdoors, and Stephanie Sarbanes attempted (unsuccessfully) to make me slightly musically cultured when we'd stumbled upon some Met Opera tickets. Ype de Jong and Mohammad Kabbani, two of the clinicians in the lab, helped throughout with various medical and grant-related topics and always showed concern on some of my health habits, especially those relating to Coca-Cola intake. Mariel Bartley helped edit parts of this thesis that were too boring to entice others to bite when she was on her post-defense vacation, as she did with many presentations and documents prior, and helped cover my lab meeting food responsibilities multiple times. I thank the lab’s support staff for helping keep the lab running, including Santa Pecoraro, Ellen Castillo, Anesta Norris, Aileen O’Connell, Glen Santiago, Joe Palarca, Sonia Shirley, and Michael Pearce, who can be counted on for almost anything. Ellen puts in a tremendous effort
into keeping the cell culture room running and providing a constant source of cells for experiments, and also into sending thoughtful cards and messages for every big event and occasion. I also thank the personnel at Rockefeller University - Melvin White, Katrina Gray and Isaiah Curry always had a smile and a nice conversation, and Alzatta Fogg made frequent visits to the Abby dining hall a must and shared many interesting stories from her experience of New York City in various eras since WWII. I ran into Anesta and Isaiah at a Korean BBQ place in Queens by happenstance, and I’ve known since that they also share an appreciation of Kalbi short ribs, followed by vanilla ice cream. The campus security team has always been so nice, it's been a great pleasure to get to know so many of them, mostly from the earlier years when I used to loiter around the front gate for smoke breaks, and we've had many great conversations. Inna Ricardo-Lax has been a tremendous collaborator on a variety of fun projects and has always been a source of help and discussions. The viral epitranscriptome and direct RNA sequencing projects, in collaboration with Chris Mason's lab, has been filled with exciting discoveries. Stephan Bluethgen, her student, was a great 'apprentice' of the oligo capture method and it filled me with pride when he designed and purified a viral RNA species from scratch. And the RNA FISH microscopy project with Zak Singer and Brandon Razooky has been a tremendous pleasure to be involved in. Zak and I came up with the project after we saw each other's lab meeting presentations and thought something along the lines of the current project would be cool, and brainstormed some initial ideas over lunch. We had no idea if we were totally bonkers, but Charlie said to go for it anyway. Through many, many refinements during the weekly post-lab meeting lunches with Zak and Brandon and discussions with Charlie and Andrea Branch, the project evolved into something special, and is now a paper and a solid foundation for grants and future papers, including on coronaviruses. Most of my projects in graduate school were refinements or natural outgrowths of ideas given to me,
and in contrast this one was completely our own, at least initially. It was based on gaps in knowledge that no one had attempted to tackle in about 20 years, and was completely outside of anyone else's radar, and I'm pretty confident the project timeline would've been different without that lunch brainstorming session, if it would've happened at all. Involvement in this project has, for better or worse, given me confidence that I can come up with original ideas that may be worth something. I am not a microscopy expert and so I feel my contributions have been a very small portion of the project, but involvement in this project has also been one of the most fun parts of graduate school.

I thank Chris Mason and all the members of Chris's lab. I first met Chris during my interview week dinner and our first conversation was on a range of topics from science to the elegance of our Grandfathers' shotguns. Initially for my thesis project I had no idea what I was doing and continually ran into roadblocks during the rather arduous next-generation sequencing sample preparation phase. He was always willing me to meet and help when he had better things to do, and was able to quickly diagnose and tell me what was wrong and suggest fixes. The sample prep had so many sequential steps and I would have no idea which step was faulty and how to even begin fixing it. His suggestion of cleaning up primer dimers using his BluePippin machine enabled my first successful CirSeq sample prep, the foundation of my thesis project, and without his input I'm not sure I would've ever gotten it working. It was pretty clear I knew nothing about next-generation sequencing, the focal point of his lab, but he took me on as a student and became my co-mentor anyway. It has been an absolute pleasure to be involved and be able to help where I can in various projects in his lab ranging from direct RNA sequencing of viruses to viral epitranscriptomics. Chris also made time to meet weekly and multiple times a week at times, and I never had to gather intel on potential office door openings in order to do so. Noah Alexander, now a graduate
student at UCLA, was a great collaborator and friend, and the two of us along with Yogesh Saletore had numerous lunch meetings on the Upper East Side. I'd like to say we explored the area's food scene, but we almost always ate at the same place. Alexa McIntyre, now a postdoc at the University of Zurich, also shared a deep interest in virology, sent me many photos of her cat, and was always willing to help with programming issues. Dan Butler and Ben Young were helpful throughout with various experiments, reagents and discussions. I cannot even begin to imagine having better co-mentors than Charlie and Chris, and it's nice that I was but a small part of the ongoing collaboration between both labs.

I thank the various members of the graduate school for being so helpful and gracious throughout. Dean Randi Silver has gone out of her way, while being so busy, to be so nice and helpful on all aspects of life - from my first day ever on campus during my interviews, to my last day and beyond. Dean Augustine Choi and his wife Mary hosted me for dinner at their house and has made time for school-related and sports discussions while having a medical college to run. Dean David Christini spent much of his time checking up on me and helping me write and edit grant applications. Dean Carl Nathan provided me with discussions and insights into the history of immunology which I would not have been aware of otherwise. Dean Barbara Hempstead gave me great advice on school and the value of skills I would gain from the AMBA program. In addition to their interactions with me, I especially thank them all, and all the graduate school administrators involved, in running all the day-to-day operations in the background that keeps the school running so efficiently. Running a graduate program or a medical school is not easy nor is it cheap, and I've been able to be in a sort of a cocoon where I can just happily think about my science all day and I didn't once have to think about my stipend or where the money was coming from for various activities or services. For that, I
am eternally grateful. Living in such a bubble without having to consider finances is not a tenable long-term proposition, and I hope to gain some of the skillsets they all contribute to the school during my AMBA program.

I thank all the members of my PhD program, the Department of Physiology, Biophysics and Systems Biology (PBSB). I cannot imagine a better department or a better fit, and I am so lucky to have been a part of it. It has truly felt like a family, and my thanks to Basil Hanss for recommending it to me when I was considering grad school programs. Harel Weinstein, the chair of PBSB, in addition to running such a great program always took time to catch up and give me great scientific advice. Emre Aksay routinely helped with progress within the program and administrative requirements, and made an effort to introduce me to virologists when I'd just mentioned to him that I thought viruses were interesting. I've had numerous nice conversations with Larry Palmer, Olaf Andersen, Crina Nimigean and Olga Boudker on a variety of topics within science and outside of science, including National Park visits and road trips in Subarus. Among others, Dan Gardner, Bernice Grafstein, Carl Blobel and Luce Skrabanek ran wonderful and helpful courses. Luce told me that they changed the first year course schedule so that the students can attend my dissertation defense, which was so, so nice! I hope they found it at least a little entertaining. Audrey Rivera was the point-person for anything and everything relating to the program, and I cannot thank her enough for continually being so nice and so helpful. She represents essentially the 'core' of PBSB, and neither I nor most other PBSB students can imagine our grad school experience without her. She won Weill Cornell Medicine's inaugural Excellence in Student Service Award in 2016, and I think she deserves it every year. I do apologize to Audrey and others, one more time, for all the paperwork and program milestone deadlines I have missed which caused them unnecessary stress! I've been honored to be invited to help with PBSB
recruitment every year, and with orientation and retreat pre-ACE student mentorship events. Presenting my poster every year and helping with student-coordinated recruitment events managed by Cindy Liang, Shana Bergman and others has been a lot of fun. I'm grateful to the various recruits who ended up joining PBSB who reached out and assumed, for better or worse, that I was a good person to chat with about the program both prior to joining and after, and who also gave me really nice comments about my efforts, which among others includes Justine Lottermoser, Rafael Colon, Zita Aretz, Arielle Messer and Jiwoon Park. If they are to be believed, I guess I was useful to some degree. My PBSB batchmates, which includes Sahil Chopra, Kofi Deh, Patrick Grinaway and others, made the first couple years of courses a great shared experience. And speaking of first year courses, I must share a story of Jason Banfelder, director of the required first year course quantitative biology (qBio), who is a tremendous mentor and has since become a good friend with whom I've had many, many helpful discussions and lunches with. The course involved learning biological statistics and applying it using the programming languages R and MATLAB. I had never programmed before, I had never taken linear algebra which was used for the MATLAB component, and my research up to that point had not involved computational biology and I thought I'd never use it. It was really hard for me, I thought I didn't have a good enough background and felt behind, and it caused quite a bit of stress. So I went to him and asked him if I can drop the course and take it later, and he said "No". He said that even if I had to go to his office for 3 hours a day every day, he'll help get me caught up and up to speed. So I went to his office for several hours a couple times a week, and he drew stuff like vectors on napkins which I'd never seen before until I got it, when he certainly had many better things to do. What's funny is that I thought I'd never use it in my research, and shortly after I was given the option of choosing the rotation project which became my awesome computational-heavy thesis project, in which I do what I learned in
the class every single day. Without getting through the course, I wouldn't have had the background or confidence to start the project and would've had to choose something different, which almost certainly would not have been as fun. Jason's exceedingly generous efforts changed the course of my career and my life.

And outside of PBSB, I thank everyone involved in the amazing courses and seminars that I've been so lucky to be a part of. The Microbial Pathogenesis course hosted by Luciano Marraffini was an excellent source of knowledge of the various non-viral infectious disease pathogens and the discovery of genetic tools, such as restriction endonucleases and CRISPR, that have been derived from them. The Translational Science course run by Henry Murray for each session had a lecture on recent research and the development of a therapeutic based on that research, followed by an interview of a patient taking that treatment, which provided fascinating perspectives on the impact of modern science research. The fascinating Drug Development course run by Ignacio Rodriguez had scientists from Roche come in and teach us about every step involved in designing a drug based on current scientific understanding and taking it to the 'finish line' of treating patients. Georgia Frost organized a Biochemistry course taught by Gregory Petsko which remains one of the most unique and interesting courses I have ever taken. Dr. Petsko spent the entire course solely on glycolysis, spending each 2 hour class on a single reaction within it and a deep dive into "what problem is the cell trying to solve here", and taught us a way to look at biochemical reactions I'd never seen elsewhere. Even if I hadn't even thought of it in years, I can still think through and draw out the glycolysis cycle without any prior memorization, whilst I'd immediately forgotten it the day after all my undergrad exams on it. He also conducted a session where we read Greek epic poems and used them as a framework to think deeply about our career paths. The Programming in Life
Sciences course taught by Seth Syberg provided me with the python programming skills which made large swaths of my thesis project analyses possible, which I would not have picked up within a reasonable timeframe otherwise, if ever. The Infectious Diseases Area of Concentration run by Harjot Singh and Laura Kirkman organized innumerable exciting seminars that I got to enjoy, and recently being able participate in New York Presbyterian Hospital's COVID-19 response and clinical update meetings hosted by Trip Gulick has been extremely exciting and informative. The Clinical and Research Genomics course run by Chris, my mentor, helped put what I was doing within the context of all the exciting current research going on driven by next-generation sequencing advances. And being a part of the Virology course, run by Charlie, my mentor, and Paul Bieniasz, every time it has been offered while I was in grad school, has been one of the great pleasures of my life. I got to learn from and have exciting lunch discussions with virus experts from around the world, including Eckard Wimmer, Stephen Harrison, Terry Dermody, Vincent Racaniello, Sean Whelan, Doug Lowy, Bernie Moss, Grant McFadden, Julie Pfeiffer, Elodie Ghedin, Steve Goff, Lynn Enquist, Marty Markowitz, Adolfo Gracia-Sastre, Robert Lamb, Richard Condit, Colin Parrish, Theodora Hatzioannou, Ian Lipkin and many others. I am so grateful to have been given the opportunity to develop and run the fun discovery-themed discussion session of Charlie's flaviviridae class for the students after I had officially taken the course. And it was an honor to work with Vincent Racaniello and Glenn Rall as a student editor on the next edition of their Principles of Virology textbook. I got to do deep dives into various exciting aspects of virology and assist in refining the primary textbook for the next generation of students interested in viruses. I am grateful to be recognized as a student editor in the upcoming edition, and the copy I received due to being listed as such saves me the money I would have otherwise spent acquiring the amazing textbook upon release, as I’d done with the previous edition. The Tri-
Institutional campus, which includes Weill Cornell, Rockefeller and Memorial Sloan-Kettering, has continually invited amazing seminar speakers, most of whom are really nice people, and always reserves a meal for the speaker and only graduate students, so that we'd feel more comfortable given our varying degrees of inexperience and naivete. Highly memorable lunch and dinner conversations include those I've had with Tom Monath, Sondra Schlesinger, Anna Pyle, Jim Strauss, Ellen Strauss, Jack Szostak, Laura Kramer, Anthony Fauci, Bruce Buetler, Ronald Breaker, Stan Lemon, Eric Betzig, Olaf Schneewind, Andrew Fire, Volker Lohmann, Francois Penon, John Taylor, Chuck Samuel, Karla Kirkegaard, Pedro Vasconcelos, Andrew Goodman, Joseph Heitman, Sydney Brenner, and Carolyn Coyne, among many others. Tom Monath gave me essentially a crash course on the history of arbovirology and added multiple dimensions onto what I had read in Charlie Calisher's book, and graciously mailed me the DVD of Telford Work's virus discovery film which is described throughout this thesis and allowed both the aforementioned flaviviridae session for students and a fun lab movie night. After a wonderful discussion, Anna Pyle took the effort to send me immune-activating RNA constructs and an ultra-processive reverse transcriptase enzyme before they were scalable and commercially available. Sondra Schlesinger spent an entire day with Zak and I chatting about our work on Sindbis and the history of Sindbis virology, in which she played a critical role as described in the first chapter, and it was one of the best days of my time in grad school (which was filled with great days!). Chatting with Jim and Ellen Strauss about birdwatching, forests, and the lives of many of the virologists mentioned in the first chapter, many of whom started their careers in their lab, was a great experience. I have a feeling, which is probably true, that these classes and seminar series were the best such courses provided anywhere in the world, and I am so privileged to have been a part of them.
I thank my committee members and my lab rotation laboratories. My committee members past and present, Scott Blanchard, Jeremy Dittman, Xin-Yun Huang, Randy Longman, Jason Mezey and Joao Xavier, volunteered their time for the sole benefit of my science and graduate school progress. Scott was passionate and engaged with my project throughout grad school even before joining my committee, and provided many great discussions and suggestions, and I'm grateful that he continued to be a part of my project even after moving to St. Judes in Tennessee. I look forward to continuing to discuss and further probe cellular RNA polymerases and ribosomal RNAs. Xin-Yun, who also happens to be Charlie's neighbor and shares an enthusiasm for pet dogs, assisted with many administrative tasks and hassles that were beyond my abilities. Randy, who was Charlie's student a long time ago, agreed to handle the administrative and probably pretty annoying chairperson responsibilities of my defense when I asked for a favor, after he had to spend the morning of my defense date in the clinic. Jason Mezey invited me to present my project to his group for great feedback, and gave terrific scientific advice on the focus of the project that I hadn't considered before, which helped chart its course. And Jeremy hosted me for my second rotation, on presynaptic transmission using the model system Caenorhabditis elegans. I'm grateful for his continued involvement throughout my time in grad school, and all the multi-hour conversations we've had that ranges every interesting topic imaginable each time. Jeremy has such enthusiasm for the history of science and the principles of discovery that I learn new cool facts and perspectives about the world each time we chat. In addition to the discussions of my project, he provided thoughtful feedback on this entire document. Ishani Basu helped mentor me both in his lab and in navigating early grad school, and building and troubleshooting the lab's microscope computer was a fun experience. I was involved in a totally wild idea to ascertain complexin's effective distance on the SNARE complex from the presynaptic vesicle by
using a linker and random walk principles. Later on, Justine approached the question from a different angle and got it working and was able to effectively investigate the question, which was really nice to see! Alessio Accardi hosted me for my first rotation on ion channels, which was a natural progression from my previous work and formed a nice transition into graduate school, and also taught a great course on membrane proteins. There, Daniel Basilio mentored me for my first graduate school project and helped get me used to the additional research and creative responsibilities, and Malvin Vien, a previous ‘Project Runway’ participant and currently a medical student at the University of Colorado, was a good friend. My project was a fun and totally crazy idea to try to convert a chloride channel to a fluoride channel, and I got to meet Randy Stockbridge and Chris Miller at the Biophysical Society Meeting, who had first identified the fluoride channel I was using as a guide and provided interesting input. I’m so grateful to have been given the option in both those rotations, among more conservative projects, of attempting a crazy and highly risky project where I was told from the get-go that we’re essentially driving into the dark with no headlights and there’s a good chance nothing may come of it. It was tremendously exciting and gave me the confidence to choose and attempt what ended up being my thesis project in my last rotation. I imagine taking on a safer and more routine project would’ve also ended up fine and some thesis dissertation would have come out of it, but it certainly would’ve been a whole lot less fun.

And I thank my friends from the program, without whom grad school would have been far less enjoyable. Especially, among those not described above, Mark Spurgeon, Mike Levine, Chad Kurylo and Will Shipman. I met Mike during interview week through a mutual friend, and now that he is also working on viruses, I look forward to continuing discussions on virology and non-science topics. Mark provided great company at many lunches, events and
concerts, and Will knew that I had mentioned an interest in viruses and that I was bad at keeping track of mass emails and told me to specifically check for the message announcing Charlie's virology course. If I'd missed that email, which was quite likely without Will's efforts, the trajectory of my life would've been drastically different. Chad accompanied Noah and I for numerous lunches and discussions in the Belfer building and nearby, and Chad hosted me for my AMBA program visit and was very helpful throughout the process. I am grateful to be able to stay in touch and see how their careers have evolved and will continue to evolve since the "early days" of grad school.

I thank the many people who have put in an effort to contribute to my success during my time in graduate school. I am so grateful to professor Jamie Rosenwald for replying to my 'cold email' immediately and allowing me to audit his great course on investing. I've had the privilege of attending his fun and informative course every year since then, and also the wonderful dinner that his wife Laura hosts every year. I hope to build on the foundations that he's taught me in my next steps. Laura introduced me to Uriel at one of the dinners, an undergraduate student in the course, who expressed an interest in medical research and it was really nice to be able to inform him about the field and then have him as a highly collaborative member of Chris's lab. Uriel also introduced me to a rather inspirational TV show which begins each episode with the phrase, "Welcome to the kingdom of science. Get excited!". I am very grateful to Scott Aylor, Lisa Su, John Taylor and Matt Hurwitz from Advanced Micro Devices (AMD) for taking an active interest in computational virology and in learning more about my project. They graciously provided me with recommendations, resources and technical assistance that tremendously accelerates my analyses described within this dissertation, and the work of many of my labmates. I am very thankful to Judi Byers and her team and Cornell SC Johnson for all their advice and assistance (and Judi's instagram
post!), and I'm incredibly grateful to Charles Lee and his family for their support of me with an extremely generous scholarship. I hope to continue to stay in touch and build upon their generous efforts and the efforts of many others during my time at Weill Cornell.

I thank Arielle Klepper, now a physician at UCSF, for introducing me to science, and without whom likely none of what is described so far in this dissertation would have occurred. Arielle, the older sister of my friend Julian who is described below, initially took an active interest in my development right after I graduated college. We would hang out mostly at Julian's DJ gigs at the now defunct Bar On A in the East Village (story goes the new owner didn't know how to bribe effectively enough to make noise complaints go away). She was an MD/PhD student at Mount Sinai at the time and convinced me to try biological research due to my background of being a TA in anatomy and physiology, taught me how to read scientific papers, and introduced me to her physiology course professor Basil Hanss which initiated my research career. We had a pseudo-argument over the phone recently where she said, "I didn't do much you would've gotten here eventually" and I disagreed and said, "I'm pretty sure I would've been unemployed and homeless without you". Maybe in an alternate universe I'm somewhere between the two extremes, but I'm so grateful that I don't even have to imagine where I'd be without her efforts. When I asked her many years ago how I can pay her back she replied, "pay it forward", which I intend to do. Arielle's mentor, Andrea Branch, has joint lab meetings with Charlie's lab and it was a funny moment when I walked into my first Rice lab meeting and unexpectedly saw Arielle and went "what are you doing here?!". Andrea Branch has become a mentor to me as well, and an incredible and passionate collaborator and advisor on the aforementioned microscopy project along with Zak and Brandon and also on many parts of my other projects. I am incredibly indebted to Arielle and to Andrea as to how
awesome my research experience has been and my life is currently, without whom none of what I describe here would have probably panned out (I'm guessing if Arielle had had a lesser experience in grad school she wouldn't have recommended it to me). I will describe the serendipitous way I first met Arielle's brother Julian below. A nice aspect of my defense becoming a teleconferenced one due to the pandemic was that Arielle and many other friends who were geographically distant were able to join and have the same experience of the presentation as everyone else.

I thank everyone involved during my time at Mount Sinai. Basil let me join his lab even though I had no prior research experience, where I broke the first pipette I'd ever held on my first day, turns out the volume adjusters have limits to how far they can turn. I had such a tremendously fun time learning the basics of molecular biology and investigating various nucleic acid membrane channels and pursuing many, many crazy and fun ideas. This experience developed my preference for such projects which led to my aforementioned grad school rotation activities. Dac Anh Nguyen, a MD/PhD student in Basil's lab and my experiment mentor, essentially taught me everything I know about biological research. He told me how cool viruses were and had me read parts of Field's Virology, which probably played a central role in where I ended up. He pushed me to think of better and more combinatorial ways of approaching problems by endearingly calling me a clown and saying, "what are you in clown school?", whenever I suggested a suboptimal solution. I still utter that phrase to myself when generating ideas to motivate me to 'think outside the box' and do better. Dac also hosted me at his apartment for a while when housing was in flux, where I stayed on a couch next to a full studio music console and multiple 3D printers and all sorts of engineered parts, which was quite an experience. Justin Costa, also an MD/PhD student in Basil's lab at the time, also shared a passion for fun ideas and experiments and was the other
partner-in-crime of mentoring me and teaching me everything I knew about research. Justin would grow all sorts of interesting plants with his background in botany, had an adorable and gigantic pet that was mostly wolf and can speak for hours about specific games of Go played by Buddhist monks centuries ago while we’d watch his extensive nature documentary collection. After my first successful mutagenesis and ligation construct, Justin poured some water on my head and performed a little ritual to officially declare me a molecular biologist. Basil, Dac and Justin were critical to fostering my excitement in continuing with biological research in grad school, whilst I had many friends who also tried the field after college and began to hate it and went to do something else. I am so grateful to continue to stay in touch with them and discuss science and innovative approaches, receive advice and occasionally (frequently) be called a clown.

I am so grateful that I got to know Eddie Bottone. Dr. Bottone grew up interested in cycling and joined the military, and then started as a bacteriology technician at Mount Sinai in the 1950's and eventually worked his way up and became the director of the entire microbiology department in the 1970's, winning all of the most prestigious awards given by Mount Sinai along the way. Even after retirement, he used to take a break from driving his Audi A8 and come into Mount Sinai and Elmhurst Hospital a couple days a week to mentor a variety of physicians and students, as purely an unpaid hobby. While walking around the floor in the Annenberg Building that Basil's lab was in, which also had portions of the Infectious Diseases department, he'd stop by and chat with Dac every once in a while. When I joined Basil's lab, he had recently slipped on snow and shattered his arm, and had a cast on for a while. He had learned boxing and martial arts while in the military, and several months later after the cast was removed, during one his visits to the lab he saw me and said, "hey you look fit, come hold these punching bags for me while I box to gain my arm
strength back". So I did, in the ID department's conference room, and as his arm got stronger he would also hold the punching bag and teach me how to punch and kick effectively, incorporating muscles from around the body I didn't know could be involved in each hit. Eventually he began to teach me about various infectious agents, mostly through fascinating stories, and advise me on graduate school and encourage me to take the next step of my career. He then invited me to his weekly rounds at Elmhurst Hospital, seeing patients and the impacts of infectious disease research and treatments. At Elmhurst, they'd reserve the most perplexing cases for his weekly visit and Dr. Bottone would quickly do a diagnosis and teach the physicians about the infectious agent, often with similar phrases to "I'd seen this once in the 1960's", kinda like the protagonist of the TV show 'House MD'. One sentence from him cured my longstanding struggle with dandruff - turns out the practice in the town in India my family is from of rubbing coconut oil on our heads for "healthy hair" provided required food for the organism that causes most dandruff, *Malassezia furfur*. No dermatologist I'd seen in India or in the USA had been able to diagnose the issue, and Dr. Bottone quickly recalled that the organism is frequently misdiagnosed in clinical laboratories because technicians would often forget to rub a layer of olive oil on the culture plates. He was on the government's anthrax committee during the scare, and often told a funny story about when during a random search the cops found a bunch of anthrax-related items and documents on him, and he had to have an advisor to George W. Bush call the police station and clear things up. He insisted on writing a wonderful recommendation on my potential, which I'm sure played a key role in my admission to Weill Cornell, and encouraged me to grab the opportunity. When I was about to leave Mount Sinai, Dr. Bottone gave me a cherished copy of a 1987 book that he wrote along with Gary Wormser and Rosalyn Stahl, "AIDS and Other Manifestations of HIV Infection", one of the earliest textbooks on the disease, with the handwritten inscription; "To
Pradeep: Interlocking paths in life heralds friendship and well-being and liberates the human spirit. May this volume usher in memories of a friendship borne out of mutual caring and endeavors. With much affection, Dr. B". Dr. Bottone, now 86 years old, has been experiencing deteriorating health, and I'm so lucky that he put in extraordinary amounts of time and effort when he was able to advise and befriend me (and teach me how to effectively punch and kick stuff). Additionally, Vladimir Brezina allowed me to audit his terrific course on ion channels at Mount Sinai after Dac and Justin recommended it to me, which gave me my first taste of graduate school courses and gave me the confidence that I can handle such coursework. I am so grateful Dr. Brezina, who passed away in December 2016, allowed me that opportunity, gave me great advice on grad school, and also wrote me a nice recommendation describing my abilities which I am sure impacted my admission into Weill Cornell.

I thank all my friends from my time at Mount Sinai and Weill Cornell, and my post college era in general. I'm grateful to my friends in New York City, in particular Brian "Reaper" Hageman, Emma Lee Hanson, Nic Grigorian and Noelle Mihalinec. Brian, who grew up on a farm in South Dakota, has come to India with me twice and is one of my Grandparents' and extended family's favorite people, and must have some intrinsic luck as we spotted a tiger both times we visited this one particular forest, Kabini in Nagarahole National Park. Without his interest, I would've never made it to the Taj Mahal and other historic sites in North India which ended up being cooler than I was expecting. He was basically my roommate in the East Village for a while, and my summer crashing on his couch in Williamsburg Brooklyn right before starting grad school was quite a summer. Brian's advice, "relax, just because you start doesn't mean you have to finish, just do it while it's fun and stop when it's not", was the deciding factor in me accepting my grad school offer. Emma has been
an extremely sweet and supportive friend, and made a time period after everyone else in our close group of friends moved out of the city quite enjoyable. Nic, my roommate for a little while, is a celebrity hair stylist and typically charges upward of a thousand dollars for his clients, but gave me numerous haircuts in our bathroom for free, and provided great company. My friends from my East Village neighborhood of 1st Avenue and 10th Street - Guido, Keri, Markus, Justin, John, Albin, Mike, Satcharn, Natalie, Stephen, Kevin and many others - many of whom worked at the Thirsty Scholar and the 13th Step, made those couple of streets feel like a little village and way too fun. I'm grateful to my friends from Central New Jersey, who have made that area feel more like home than any other place I've been, especially Jack Klemens, Pete Czajkowski, Jeremy Olshefski, Phil Carrillo, Mike Barcz, Jackie Genco, Lie Griffith, Alexandra DeGennaro, Eileen Coppolla, Tommy Eichhorn, Bianca Rodriguez, Stephanie Magalhaes, Andra Nicole, and Chris Tychinski. I've probably spent the majority of the weekends over the past decade overall hanging out on Jack's couch watching various TV shows and playing video games, and discussing and arguing over the most unimportant and irrelevant issues. As a professional chef with a sous vide and a bottle of duck fat at home, he's made me better food for me than any restaurant. It was tremendously fun to have Jack and his wife Mina at my cousin Navin's destination wedding in Greece, and they gave me enough grief that I accompanied them to the Parthenon instead of laying around in my room streaming Yankees games, which in retrospect was a good decision. The texting group chat with Jack, his cousin Brandon and his brother Trey, and Phil is a constant source of random entertainment. The current argument that started just now is over the soundtrack to the original Tony Hawk's Pro Skater. Everyone needs a friend who would wait weeks, without even being asked, to catch up on The Mandalorian while there's a deadline to catch. Pete and I have spent innumerable hours also on a couch and at restaurants, music
festivals, the gym, eating thousands of calories immediately after the gym, coming up with unique band names and occasionally searching for lost items in parking lots. Pete, who is now in Indiana working on trucking logistics, was a hairstylist when he was in Jersey and also gave me numerous free haircuts, which were apparently difficult as he would say, "your hair is as messed up as your brain". Jackie has taken over the duty of giving me haircuts since Pete moved out of town, always accompanied by a fun conversation. Jeremy, an engineer at NJTransit who I like to call and blame whenever there's a train delay which has nothing whatsoever to do with him, saw things in me before I was able to see them in myself, and was always a great friend. Chris and his sister Tina always made sure that my stomach was full and there was something fun to watch and do whenever I was over. Tommy, in addition to all the fun discussions, recently hosted me at an event where we got to chat with and eat dishes made by multiple master chefs from 'The Food Network' which constituted my first official 'foodie' experience. Barczcz has hosted us for fun summer barbecues at his lakehouse and it's been tremendously fun designing and fabricating various things using copper and aluminum 6061 alloy at his Dad's machine shop. These projects were interspersed at the time as he'd routinely either get fired by his Dad or quit, and I look forward to doing these projects again as, in his words, "my Dad's calmed down a lot since having grandkids". I'm grateful to my friends from the local Starbucks in Hillsborough and Princeton where I did much computational work and wrote a good portion of this thesis - Vanessa, Alex, Joey, Mor, Keith, Kyla, MK, Rad, Amy, Yak and many others - who have made my many hours there quite enjoyable. And I'm grateful to my friends from North Jersey, who I'd initially met through Jon Jurow who was from the area, who I describe below. They have been a tremendously fun and supportive group of friends I'm glad to have in my life, which includes Allan Stein, Brett Sanders, Wes Appel, Jamie Levi, Jamie's Mom Reesa, Ian Lagowitz, Johnny Silverberg, Steve Tee and Jamie Blaustein,
among many others. Allan, who is training to be a dentist, and I keep in touch nearly every day, and it's been a really nice routine to discuss everything from daily minutiae to life-changing events and share great advice whenever needed throughout the entire course of grad school. Him and Jamie are usually some of the first ones to sense whenever something's not alright and reach out. Brett, currently in LA managing an investment company he co-owns, took me to numerous events around town when he was around, is always so full of energy and motivation, and it's always a pleasure catching up. Reesa, Ian and Johnny have been really nice and caring mentors for various aspects of life and grad school. I am so incredible lucky that the various circumstances after undergrad led to me meeting all of my friends described above - I cannot imagine my life without them, and best of all I don't have to.

I thank my teachers and mentors at the University of Rochester, my undergraduate institution. Due to Rochester's rather unique system of not having any required courses, I was able to take a multitude of interesting courses in biology, chemistry, history, and comparative religion. Douglas Brooks was one of the most amazing professors and orators I have ever had, and taught me more about human perspectives and behavior than I thought was possible to know. Mike Kirsch initially convinced me to take a course of his, and by graduation I had taken 6 courses with him. In my final year, Dr. Brooks mentored me on an independent study project where I primarily studied the works of Jonathan Z. Smith, who remains my favorite author, and taught me the subtleties in the art of comparisons which continues to influence many aspects of my life. Paul Muller-Ortega's Theories of Religion course, where we had to read the entirety of books such as William James' 'The Varieties of Religious Experience' and Emile Durkheim's 'Elementary Forms of Religious Life' every week, was perhaps the most difficult and one of the most
rewarding classes I have ever taken. Sean Hanna coached me and helped facilitate my 'Take Five' program in comparative religion, without whom most of the aforementioned experiences would not have occurred. John Jaenike, who became my biology major advisor, taught me evolution and ecology which formed the early foundations of my work on RNA virus evolution described in this dissertation. Robert Kreilick taught a fascinating course, Chemical Instrumentation, which was one of my first serious exposures to how science is practically conducted and gave me my first experiences of keeping an organized lab notebook and writing thorough lab reports, and he also graciously recommended me to his colleagues as a TA. Richard Eisenberg's course, Inorganic Chemistry, furthered my interest in chemistry and initiated my friend Kenny Lotito's career in the field. As sophomores Kenny and I went to a nearby Jamaican restaurant that did not check ID's to buy alcohol (as 40's) and ox tail, and on the way back to campus needed a light for our cigarettes. So we stopped by the chemistry building (Hutchinson Hall) since we figured they should have sources of fire, found a matchbox on the floor which was a fateful sign, and ran into Thomas Krugh who spent several hours discussing the magic of chemistry and convinced us to sign up as chemistry majors. Though I unfortunately don't remember much chemistry, their efforts developed analytical skills critical to my graduate school studies. William Hauser taught a wonderful course on Japanese history, and convinced me and encouraged me to do a history major when it wasn't even a thought that I had had previously, and became my major advisor. The major required a senior seminar in a focus area, for which I chose World War II, and as a result I had the good fortune of taking Celia Applegate's Fascism in Europe course, along with my friend AC Karchem. She was an amazing lecturer, and through all the readings and discussions I learned a whole lot about the complexities of how the world works. I remember one class where we were all tasked with coming up with a consensus on the definition of fascism and writing it on the
blackboard by the end of the 3 hour class, but as we had all developed an increased awareness of subtleties from the course, after much debate the only definition we could agree on was a frowny face. The book for one of my reports in the class, Primo Levi's 'The Periodic Table', remains one of my favorite books today, and I was recently able to recommend and share the experience with Bill Schneider and others in the Rice lab. Their efforts in teaching me to read and write history formed the foundations for most of what is in the first chapter.

And I am eternally grateful for having Dr. Alan Dietsche as my undergraduate mentor. Dr. Dietsche taught Mammalian Anatomy and Mammalian Physiology, and made them so interesting that he won more teaching awards than anyone else at the university. We heard through the grapevine that these courses were a must, and though it sounded weird to us that what people usually claim is the most boring subject could be the most interesting, Kenny and I tried it out as sophomores and fell in love with the class. Dr. Dietsche was a horse farmer and a Vietnam War veteran with the US Navy, who then pursued his PhD and became a professor and taught the most fascinating course on anatomy and physiology humanly possible. He interweaved the class with fun and relevant stories of these various experiences and highly inspirational messages, and taught us to watch out for BS from people with much higher degrees than we had at the time, and to not take anything at face value until we analyzed it ourselves. He would end such life lessons with the phrase "don't be the darkness, be the light", which always gave us a euphoric realization that we were our own arbiters of truth. He would culture buckets of his own style of blue cheese and deliver it weekly to the National Hotel in Leicester, NY, which was certainly the best blue cheese I have ever had with Buffalo wings. Upon our visit to the National Hotel they poured us a drink named the Dietsche Iced Tea, which was a tall Long Island Ice Tea glass filled
to the brim with ice and Johnnie Walker Black Label. He never wore a jacket in
the frigid Rochester winters, and said that before he could afford a car he was
so annoyed at constantly layering and fearing the cold that he decided to train
himself not to fear it anymore. I found him and his class so interesting that I
asked to be his TA my junior year, which was a hot commodity as most of my
classmates felt the same way. Because he got a lot of requests, he said he's
usually only taken on seniors as TA's, but took me on anyway. In retrospect,
this changed the course of my entire life. Being his TA during both my junior
and senior years was a tremendously fun experience. The lab itself at
Hutchinson Hall was a museum - it was decorated with the skulls of hundreds
of animals including a hippo skull at the entrance, real human full skeletons,
numerous filled formaldehyde jars, and we'd pick up organs at the local
butcher shops for dissection each week. Real human skeletons are apparently
no longer available, and he told me that ever since some medical students
broke off the temporal styloid process of one of the skulls, he'd reply to
subsequent requests from the medical school to borrow the skeletons by
informing them that he owned a shotgun - they were solely reserved for his
undergrads to learn from. He coached me to develop and take complete
ownership of my lab lectures to students, speak with authority and present
with confidence, which was far more responsibility than usually afforded to
undergrad TA's. As I mentioned above, him fostering my interest in anatomy
and physiology led to me beginning biological research in Basil's lab.
Furthermore, shortly after he was diagnosed with Alzheimer's, when I asked
him for a recommendation he made a personal phonecall to Basil and told him
about me and asked him to take me on in his lab and continue to foster my
love of physiology, which no doubt led to my time at Mount Sinai and forged
my life's trajectory. Dr. Dietsche passed away on December 2018, and his
obituary reads: "Requests from his family, if you're reading this, have a drink in
his honor; and 'Don't be the darkness, be the light!'". I was always impressed
with Dr. Dietsche's wall of binders full of notes, and now I've also been able to accumulate such a collection thanks to graduate school, which I hope he has noticed. One common thread among all the varied labs I've been in - the laboratories headed by Charlie Rice, Chris Mason, Jeremy Dittman, Alessio Accardi, Basil Hanss and Alan Dietsche - was that prior to considering joining, I noticed that they were all excellent and amazing teachers in their respective courses, and were repeatedly recognized as such by many. It has worked out every single time for me so far to consider a lab based on the mentor instead of the topic, and perhaps that is a good benchmark to consider for future students choosing rotations or labs or mentors.

I thank my friends from my time at Rochester, all of whom I am lucky to continue to be in touch with and have known for more than a decade now. My freshman hall was filled with an assorted bunch of us who never submitted the roommate and hall preferences paperwork, and I am so grateful that we are all still very close today, which according to Dean Matthew Burns is very rare. He would remind us of how much he was impressed by our continuing friendships every time we got in trouble for a variety of activities and had to go meet him, the last of which happened the week before graduation. Jacob Zoske and Jonny Stulberg lived in the room right across my door, and we'd met when our parents moved us in and looked at each other as soon as they left going "finally.. let's go find some beer". I've probably spent the majority of days over that vast period of time on Jacob's couch. Jacob's visited my Grandparents' house in India, and hosted me at his extended family's home in Rehovot, Israel, and we've also spent a summer in Italy where he discovered and introduced me to the best steak I've ever had in my life. He also introduced me to the gym where I couldn't even bench the bar at first, and greatly influenced my current fashion sense. Jacob's Mom hosted me in Jacksonville over a winter break where she made probably the best breakfast I've ever had and I
got to see wild alligators for the first time at Okefenokee Swamp, including the famous elder Oscar the gator when he was still alive. Oscar was a 14 ft and 1000 lb gator who, "surviving a shotgun blast to the face, at least three bullet wounds, broken bones and arthritis", made it to a 100 years old according to the news report and is now displayed in their museum. Jacob and his wife Sarita continue to be great friends who I try to see as much as possible.

Jonny, who was in the Navy and is now a lawyer in LA, was such a tremendous actor in college that our school's play director essentially had a nervous breakdown when Jonny wanted to take a break for a semester, and I got an A for a stage management class where I had no clue what was going on just due to my association with him. His and Grace's original wedding date got postponed due to a medical emergency, and Jacob, Sarita and I flew out to LA using our original tickets anyway to all hang out. After an incredible weekend, the last day was spent on a couch watching the entirety of the first season of ‘Stranger Things', and a pact was made that I must fly out to LA and watch each subsequent season there upon release. Jonny's brother Mike and his wife Elizabeth hosted me at Yale to tell me about the grad school experience and show me around and help with grad school applications, and also get me some *Agrobacterium tumefaciens* for some pretty wild transgenic experiments. Mike continually gave me good advice from a couple years before I started to even after he watched my thesis defense. Of note, his advice of not doing a summer rotation, "this is the last time you'll have some extended time off for a while without having to think about anything so chill out and start in the fall", led to proper timing of my last rotation being in Charlie and Chris's labs. Without it, I would have completed my required rotations before that one, and the entire trajectory of my life since starting grad school would have been different. Jon Jurow lived on another floor in the same building, and we continued to live nearby and hang out all the time throughout college, New Jersey, and the East Village, and he saw things in me and
inspired changes which made me the person I am today. He was there for me when I wasn't even there for myself, and him and his wife Kelsey continue to be great friends, and we continue to have many meetups and numerous long phone conversations. Jon's Mom and Dad also hosted me for a summer which led to me meeting many of my aforementioned friends. Kenny Lotito's and Arlo Chapple-Berletic's room was down the hall and became a focal point of our group due its rather unfair larger size, Arlo's XBox, and Kenny's desktop where we rotated playing the very first 'Call of Duty'. We spent many hours strategizing optimal rifleman and sniper placements. Kenny also visited my Grandparents' home in India, and all the locals he'd met still ask me about him. Kenny starting his PhD program inspired my own grad school journey. Arlo's paintings still hang on my walls, and he's recently been giving me advice on life in Ithaca. Jon "JT" Terraciano is always full of energy and highly motivational and can make any situation hilarious, and Jon Henderson, "Hende", continues to be a good friend with many shared experiences. You may have noticed a continuing theme - we had four Jonathans in our freshman group of friends, and as such had to create differential nicknames for each. I am so grateful that I never got around to filling my housing preferences form and consequently met and continued to remain good friends with Jacob, Jonny, Jon, Kenny, Arlo, JT and Hende since my first day of college.

Random roommate placements in college continued to introduce really good friends into my life. Laura Teicher hosted many nice weekends in The Berkshires and has been giving me advice on an MBA experience. Arthur Kotchetkov, who I got to know through Laura, became my roommate a couple years later and taught me how to cook everything in bacon fat. Tim Haas cultivated a love of traveling and seeing the American landscape, and was a critical driver (no pun intended) of our post-college cross-country road trip. Tim, Kenny, Arlo and I started at San Francisco and made it through the
Farallon Islands, Lake Tahoe, Tulelake Lava Caves, Crater Lake, Redwoods NP, Mount Rainier, Olympic NP, Glacier NP, Yellowstone NP, Devil's Tower, The Badlands and an abandoned steel factory in Pittsburgh to end back up at Rochester, putting a total of 6,356 miles on my 2001 Subaru Outback named "Cybele". We stayed in a tent without floor padding (but right-side-up), ate primarily hot dogs, peanut butter, all kinds of mammal and fish jerky and berries from roadside bushes, and kept forgetting my car's license plate number for registering our campground site so named ourselves with the backronym 'Viper Kill Patrol #73 Squad' in an effort to remember. The road trip remains one the best experiences of my life, one I hope to do again while my spine can still handle long periods of time seated in a car. My Rush Rhees library group, which included Elliot Dolby-Shields, AC Karchem, Alissa Silverstein, Rebekah Mott, Vicky Banchevsky, Meegan Conrad and others, made long periods of studying enjoyable with camaraderie and our hourly smoke breaks, and it's been nice to routinely catch up after college and see how our career paths have evolved. Alissa has been there for me during moments when I needed it, and Elliot giving me his jeans and his critique has greatly contributed to my current fashion style. Occasionally I needed a change of scenery, and my Carson science library buddy, Emmanuel Menga who was in medical school at the time, inspired me to 'keep my eye on the prize', if you will. Quite a coincidence, much later on Emmanuel took excellent care of my Grandpa after his spinal stenosis surgery while an orthopedic surgery resident at NYU, for which I'm very grateful. Many others made undergrad an enjoyable experience and kept in touch afterwards, including Chris Kieliszak, Sam Selonick, Matthew Smith, Gabe Malseptic, Glenn Goldman, Dan Steinberg, Jazz McAuliffe, Brad Grattan and Chris Freeman. Sam started hanging out with Jacob and I during our summer in Italy, and remained a good friend, who as a statistician has been coming to lab and spending much of his time helping me with the statistical components of the
analyses described within this dissertation. And my friends during my semester study abroad in Melbourne, Australia - including Noah Stahl, Noah's Dad Rick, Ellise Akazawa and John Committi and many regulars from Percy's Bar - have become lifelong friends after that amazing shared experience, when Melbourne quickly became my second favorite place in the world after the New York City area. Julian Klepper, well before introducing me to his sister which led to my career as described above, had done a study abroad in New Zealand the semester before I did, and convinced me to overcome any fears of change and do one as well. After college, Julian frequently hosted me in New York City where I met Arielle, and Julian and his Mom and their dog Jasper also hosted me at their home in Pittsburgh for Jon and Kelsey's wedding, and introduced me to critical Pittsburgh experiences such as fries from the ‘Original Hot Dog’. I'd serendipitously met Julian my first week of college, even though him and his freshman group lived in a different dorm building across campus. Julian and Jonny were taking a course on the history of nonviolence movements such as Gandhi's and Martin Luther King Jr's movements. Whenever our groups would run into each other on the fraternity quad, Julian and Jonny would both shout "let's get nonviolent" and we'd all start wrestling each other, initiating friendships between our groups which then remained throughout college. I was reminiscing about that on the phone with Julian the other day, and he said, "man that was so dumb we definitely missed some expected maturity milestones as teenagers", which is true. I can't imagine how my life would have looked like now without Julian's efforts and the chain of events subsequent to our meeting and becoming good friends. Julian's current project is a bar and concert venue named ‘Wild Birds’ in Crown Heights, Brooklyn, and I encourage the reader to visit the spot for an entertaining time. I'd originally applied to a list of schools given to me by my high school guidance counselor without any campus visits and chose Rochester without a visit based on discussions with Christina Lee, who Emily
Sheu who went to my neighboring high school introduced me to. I'm so grateful how a seemingly random selection process in high school led to such a wealth of tremendous lifelong friends and experiences which have charted the trajectory of my life since.

I thank my teachers from my grade and high schools. In my second high school, Fox Lane HS, Dr. Dennis Maika had a rather unique approach in his AP History class and insisted that we always read the primary source first, followed by summaries in textbooks and articles. Though it was a lot more work at the time compared to others' history classes, it has had a tremendous impact on how I've approached topics since, and this impact is evident throughout this dissertation. Dr. Maika and Mr. Terlizzi also took an active interest in helping me get acclimated upon moving to the school. Dr. Eisenkraft, Mrs. Shea and Mrs. Kaczorowski gave me exposure to college-level science in their physics, chemistry and math classes. In my first high school, Holy Trinity HS, Mr. John Brady was a fantastic homeroom teacher and my chemistry teacher who inspired my interest in the subject. In my middle school, Mr. Wall also took an interest in getting me acclimated after moving, was a great homeroom teacher, and initiated my interest in history. And during my time in India prior to middle school, Mr. Martin, Mr. Charles and Ms. Rani among others built the foundations of my educational journey. I'm grateful to all my school teachers who did not allow me to develop a distaste for learning, which seems quite an easy thing to pick up as a child.

I thank my friends from my grade and high schools, who I am tremendously lucky and grateful to still be in touch with. From my second high school, Cody d'Ambrosio made many classes entertaining, took me to many fun concerts and events and continues to be a good friend, and I try to meet up with him and his fiance Molly, who also shares an interest in viruses, whenever
possible. Chris Acevedo made the schoolbus rides interesting and we spent many hours after class playing Starcraft. Matt Lazarus and Gus Ahrens also made an effort when I'd just moved to the school to get me integrated and out to many events. From my first high school, Adam Lyn, Andrew Dewar, Mike Arnes, and others in our group continue to meet up around the world, even though we all live in different places now, and most recently at Adam and Anne's fantastic wedding in Miami last year. Adam convinced me to do my study abroad in Melbourne having done the same a year prior, gave me terrific advice towards it, and then moved there permanently, but still makes it a point to stay in touch routinely and come visit America whenever possible. Lennox David and I had many fun arguments over who was better at certain Playstation games, Greg Emeribe debated TV Shows, John Allen, Joe Brown, Allen Berry, Sam Dellile and John Davis sat nearby due to alphabetical order seating and made many of my classes entertaining, and Darius Anderson helped with many teen events as someone who was a couple years older and had his locker next to mine. Hanging out with my friends from my junior high now - Sam Wallace, James Gancio, Chris Gordon, Amar Menton, Matt Wojciechowski, Jon Campo, Chad Barnes and others - even after fairly long gaps, still feels the same as in 8th grade and like nothing has skipped a beat. Sam, now 6'11", was the tallest guy in the grade and I was the smallest, and he still reminds me, "Pradeep, remember when kids wanted to mess with you but wouldn't because they'd be scared I'd crush 'em like a pancake", which I can't say I remember such situations well but I trust him. Sam made every situation hilarious and taught me to be funny, and we watched comedies including 'Mystery Men' and 'Friday' numerous times in middle school and most recently a couple months ago. Sam's Mom was a second Mom to me, and I still stop over to hang out even if Sam's not around. And his favorite food continues to be my Mom's curried meatballs, and he still eats 50 at a time when I secure some for him. Chris Gordon and James Gancio made the
schoolbus rides entertaining. And I'm grateful to still be in touch with my friends from the apartment complex I lived in during middle school, adjacent to Nassau County hospital and the county jail, from back when playing outside all the time unsupervised was something children did - Paley Yin and Paley's Mom and Dad, Mustansir Abbas and Mustansir's Mom and Dad, Hasnain Sial and his parents, Jeffrey Yu and his parents, and Imran Mohammed and his parents. I've seen Paley, and more recently his wife Annie and daughter Kelsey, at least each month since 7th grade and most recently a few hours ago. Our apartment's window looked onto the playground and my Mom initially didn't like Paley since he'd come push his younger brother Paige off his bike and drag him kicking and screaming, but then found out that since both their parents were at work Paley would cook for and feed his brother even at that age. Paley stopped by during my study abroad in Melbourne and has also visited my Grandparents house in India, when he fell asleep five minutes before we stumbled upon a wild leopard and unfortunately missed his chance. Mustansir was my direct neighbor and I'd spend a few hours after school each day at his house. One summer in college while taking a summer class at Stony Brook, Paley and I lived in a makeshift garage with just a George Foreman grill and I'd spend the weekends at Hasnain's house playing Xbox, which remains a memorable summer. And Ashley, Vijin and Evance made my elementary school time in India memorable. Moving so often was initially fairly annoying, but I'm grateful that it led to so many lifelong friends I would not have met otherwise.

And I thank my family. Not that there was ever a choice, but I'm so lucky and grateful to have been born into the one I'm in! My Mom was in the first generation in my small town in India where women got an education, and she had me in her mid-20's. I come from a long line of teenage mothers before that, which was the societal expectation in those times and in that part of India.
As a result, I got to know all of my Grandparents (to whom this thesis is dedicated) and two of my Great Grandmas well into adulthood, which is an absolutely tremendous privilege. Apparently I was a rather picky and annoying eater when I was a child, and so my Mom would get annoyed and give up trying to get me to eat, and my Grandma would go to her daughter, "you don't know what the hell you're doing" and try to get me to eat for a while. Then she'd give up and my Great Grandma would go to her daughter, "wow I can't believe I raised you, you clearly don't know what you're doing" and take over and keep trying. So it has "taken a village", if you will, along the whole way. My Grandparents have been tremendously supportive all throughout my life, even raising me at key points throughout. To whatever degree I'm aware now, they taught me right from wrong, and knew when to tell me to calm down and apologize when I was angry at my parents, and when to parent my parents. I saw my paternal Grandparents in Central New Jersey often, and my Grandpa continues to support me in many ways, and I've been quite privileged to visit my maternal Grandparents in Nagercoil, India, once or twice a year, including every Christmas, for as long as I can remember. When in India, my Grandma guilts me into eating painful amounts of freshly caught curried fish, and my Grandpa takes me out to his local farm in Anjugrammam every day to see livestock and photograph wild birds and small mammals, and has handled the logistics of my hobbies and my forest trips throughout my whole life. Each visit is a really nice experience. In fact, when Brian was over in India, we knew something was wrong and that my Grandma was having a medical emergency when she didn't seem to care how much we were eating, which was the first time that has ever happened, and it led to quick notice and successful intervention. My next trip will happen as soon as this pandemic is over. I am so incredibly blessed and grateful to have the extremely loving Grandparents that I do, and be able to see them as often as I do.
As I've gotten older and my friends have started having children, I've realized that having a child is quite an expensive and time-consuming endeavor. My Mom and Dad put in quite a bit of effort towards this endeavor, which I still haven't realized the full extent of, and towards supporting my education and my rather expensive hobbies. It seems that having your child be ungrateful and entitled to most of these efforts for a long period of time is a fairly standard parenting experience, and I did my part to give them that experience in full. I'm also grateful that they've been so nice to my friends, and as many of them say, a 'second set of parents' to them. As was the norm before cell phones, many of my friends would just show up and hang out at the house with my parents even if I wasn't there. However, this has continued even today, when my friends would now call and go hang out with my parents when I'm nowhere near, and my Mom makes each of my friends' favorite dish whenever possible. My little brother who is 6 years younger, Sandeep, allowed me to have the whole range of relatable experiences that come with being an older sibling. Amidst making up lies to get me in trouble and get my parents to take the TV remote away from me and give it to him, he introduced me to many entertaining TV shows. I drove up to Rochester to pick him up after his first semester in college, and when I first saw him outside his dorm he looked completely different and had on new fashionable clothes and a different haircut. I immediately tackled him to the ground and got his new clothes muddy and grassy and he said, "you know, my life has had a lot of changes recently and it's really nice to know that no matter what changes I go through, you'll still be the same ***hole. That's a constant I can rely on". I'm grateful to continue to be entrusted with such older sibling responsibilities.

And I should take a moment to recognize the efforts of my ancestors from before my time. I come from an extremely highly educated, financially stable, privileged and comfortable family, which has been that way for many
generations. I've always had a house to live in, traveled by car, have been able to travel extensively, and pursue whichever educational path and whichever hobby I desired at the time. And I come from a very large and highly networked family, my Mom has 27 first cousins on her Dad's side, all of whom have advanced professional degrees (and to be fair, quite a bit of alcoholism and anger issues as well). As such, to whichever degree one defines whatever "success" that I've had, I was "born on third base and made it to home plate", if you will. But it wasn't always that way, and it very, very easily could have not been that way today. I come from a low caste in India, called Nadar, who are officially titled as palm tree climbers. For a very long time, that was the only profession someone born to my caste was allowed to even think about doing. And they were subject to apartheid - for a long time they were not allowed to walk on the same street, live in the same region, or even wear clothes in the presence of higher castes, which was considered disrespectful and offensive and often lead to death. Part of the palm and coconut harvest belonged to the higher castes, and several days per week had to be reserved as slaves in their service. Then the British arrived, and at various points in the 1700's and early 1800's, all of my various ancestral lineages converted to Christianity which gave them access to clothes, education, networks via churches to those in charge and a mental framework which did not define the profession you had to go into. In the mid-1800's the British Queen issued a religious freedom decree which allowed members of my caste who were Christians to wear clothes in public for the first time, and the British military were dispatched to my town to quell riots and murders by members of higher castes who took offense and felt their wealth threatened. Over the course of a few generations, each lineage made it out of poverty at varying times. Most of this information is lost now, or I just have never heard it, but one Great-Great-Great Grandfather born in the 1860's was a surgeon who traveled the world and provided medical assistance during the Boer War, and another from that era was an engineer and helped
build dams that we can still see near that town. I've heard firsthand stories of them, the Grandfathers of my two Great Grandmothers that I've known. My Mom's paternal Grandpa, Mathias, who died just a few years before I was born, on the other hand was born to a pastor and grew up in a mud hut having to go to the bathroom outdoors and take baths in rather dirty lakes and rivers. He persisted and became the first member of his family to go to medical school and developed a very successful surgical practice. Apparently there were sometimes conflicts of old money vs new money somewhat akin to ‘The Great Gatsby’. As such, all of these people I have never met, and may never even hear about, had a critical role to play in allowing me to do the stuff I've been able to do, which includes this thesis dissertation. I was never forced to take my clothes off in public, only intentionally. The only time I've ever even heard about my caste, outside of some old people stories, is when my cousins in India were able to get into better colleges due to affirmative action for "backward castes". The only caste-specific thing I have ever done is buy 'official' coconut-cutting blades only because I thought they were cool. Without a long line of decisions, hard work and luck going the way it did, I would most likely be climbing palm trees and harvesting palm fruits and coconuts today, as many members of my caste still do. Which is fine, but I like my life a lot better this way, and it was through absolutely no effort of my own. I don't think I have what it takes, but maybe I would've made it out of poverty like my aforementioned Great Grandfather did if I was born that way, but I'm so grateful that I don't even have to think about it. I am very lucky that the combined efforts of the previous generations in my family got them all the way to third base so that I could come into the world on it.

Coming from such a large family, I have many uncles and granduncles and aunts and grandaunts and cousins to thank. My Mom and Dad's siblings and their families, Jeya(pa) and Reena(ma), Lulu Chitti and Ruben Uncle, and my
cousins Jonathan and Rowena, and my Gradma's brother Mohan have been incredibly nice and supportive throughout my whole life. On my Dad's side, I see Nalini Aunty, Sudha Uncle, Shobi Aunty, Sundar Uncle and my cousin Prita among other members of the 100+ person Ambrose clan often. My wildlife photography hobby did not happen in a vacuum, my family on my Mom's side have been going to forests for hunting since my Great Grandpa's young adult days in the 1930's, and then for wildlife watching after India banned hunting in the 1970's. As a young teenager, my Dad gave me his old camera with a broken battery compartment that I held together with duct tape, which would rip apart releasing the batteries onto the floor at opportune moments when I was just about to photograph a rare wild animal. When this got too annoying, I convinced them to buy me a digital SLR, and the photography component of my hobby grew from there and it's been amazing to be able to share the experience with much of my family. My cousins Rahul and Rithy and their parents Madhu Uncle and Asha Aunty have taken me to many forests since I was a child, including numerous trips to Balamore and many National Parks throughout India, and still do so whenever I visit. I've been accompanied by Rahul or Rithy almost every time I've sighted the elusive wild tiger, including once when we all saw a live tiger hunt of a deer, which remains one of the coolest things I've seen in my life. My cousin Navin has also accompanied me on various trips to Balamore, before and after moving to America, most recently with his wife Heather, daughter Evara, Mom Lalitha Aunty, and his Grandma (and my Grandpa's sister) Sushila Patti. Many more uncles, including Sudha Uncle, Sabu Uncle, Ranjit Uncle, Prasad Uncle, and others took me out to forests often. And many others from the area, including Sashi, Sundaram, Prabhu, Muthu, Devadas, Suresh, Chappandi, Saravanan, Chodalai, Mani and Chalgundu accompanied us and made those experiences tremendously fun. My Dad's medical school friends Dr. Moji Jini and Dr. Gnanaraj hosted me and my friend Annie at their homes in the
Himalayan Foothills and took me to see rhinos at Kaziranga National Park. I'm grateful to know so many people who share a passion for seeing and discussing wild birds and animals. My Great Grandpa on my Mom's side bought a large plot of land when he was young and most of his children and many of his grandchildren out of his 100+ descendants built their houses on it, so when I was a kid in India I lived on the same street as many of my cousins, including Navin, Rahul and Rithy who were in my age group. We played together almost every day, and played cricket frequently, which often ended in fights and angry shouting and threats with a bat over whether or not someone was out, since we lacked an umpire. Despite that, we all remain very close today. Navin lived with me in Central New Jersey for a while when he first moved to America, and now lives in Virginia Beach, and his and Heather's wedding in Greece was an incredible experience alongside many of our Jersey friends. Rahul also moved to America and now lives in San Francisco with his wife Aditi and children Vivaan and Viya, and Rithy lives in Bangalore, India with his wife Sujitha, who had a wonderful wedding this past year. We've known each other since our births, and it's nice to stay in touch often due to the magic of the internet and meet up whenever possible. I am so grateful to have all of the aforementioned circumstances surrounding my birth, which has made my life largely secure and enjoyable, and enabled the option of having all the experiences and the educational journeys that I've been privileged to have.

And finally, I thank the reader, for taking an interest in the work presented herein, and I hope that you found some portion of this dissertation a rather enjoyable read.