Expanding, Characterizing, and Repurposing the Invertebrate Virosphere

By Bryce Demopoulos, Ethan Lin
Department of Biomedical Engineering, College of Engineering

Abstract

Modern medicine demands the capacity to deliver genetic or biological cargo to specific cell types. Past efforts to achieve this goal have relied on the retooling and re-engineering of a small subset of vertebrate viruses with limited success. Remaining challenges with regards to in vivo delivery include finding novel viral vectors that can achieve different target specificities in addition to those that are more amenable to synthesize de novo. In an attempt to address these remaining limitations, we collected and sampled diverse invertebrate species to isolate and identify RNA viruses associated with them. As the invertebrate virosphere remains largely unknown, we hypothesized that we would identify novel viruses whose components could be characterized and repurposed to build a new suite of viral-based tools. To this end, we isolated and sequenced RNA from a diverse library of invertebrates (including 42 insects) by next-generation sequencing and subsequently performed de novo genome assembly on the reads obtained. Captured reads were analyzed for signatures of RNA dependent RNA polymerases (RdRps) – a necessary component of all RNA viruses. The two putative novel virus genome assemblies discovered were named Castor and Pollux, and were characterized and independently confirmed by quantitative PCR. These small RNA viruses or their RdRps (less than 5kB) will, in the future, be synthesized and artificially launched in mammalian cells to ascertain whether they can be selected via guided evolution to function and deliver a desired genetic or biological cargo.

Introduction

In the advent of recombinant DNA technology, science has witnessed the development of tools to generate antibodies from a plasmid, silence messenger RNA, deliver genes, and even edit DNA with single base pair resolution (Khan et al., 2016). However, to capitalize on these discoveries, the scientific community needs the ability to deliver the necessary cargo to the cells of interest (Dobson, 2006). Thus far, the issue of delivering genetic material to cells has come in the form of either repurposed viral vectors or the direct delivery of genetic material (Thomas et al., 2003).

With regards to viral vectors, these have largely focused on the use of lentiviruses, adenoviruses, or the adeno-associated virus (Robbins & Ghivizzani, 1998). While each of these viral vectors has demonstrated promise in some context, each has inherent issues preventing their widespread use. For example, early clinical trials with lentiviruses resulted in multiple integration events that culminated in the development of cancer (Condiotti et al., 2014). Conversely, use of an adenovirus, while it does not integrate, many in the human population have pre-established immunity to the vector rendering it ineffective. While both seroprevalence and integration are issues that can be addressed with additional testing or methodology, neither vector represents a tool that can be used in a more generalized platform (Vemula & Mittal, 2010). In place of these two popular vectors, a third expression system has gained popularity called the Adeno-associated virus (AAV). This vector, which in nature impacts a wide variety of animal species, can be repurposed for gene delivery and has shown significant promise in clinical trials (Naso et al., 2017). While AAV neither integrates nor

shows high seroprevalence, its limitations derive from the fact that it tends to deliver to the liver and has a very limited coding capacity (Robbins & Ghivizzani, 1998).

While we still use these vectors, the limitations are well understood and the scientific community is simultaneously looking for other solutions. Over the past few years, many researchers have focused on various lipids or synthetic nanoparticles to deliver recombinant DNA to cells (Zhao & Huang, 2014). However, this has proven difficult, largely owing to the inability to breach the barriers required to reach the nucleus. More recently, a promising technology in this area is the direct use of RNA. The use of RNA as a therapeutic is promising in that it can be easily manufactured and does not integrate. However, while promising, a remaining limitation of RNA is its inherent instability. In this regard, the identification of novel RdRps may also enable the engineering of self-replicating RNAs, thereby overcoming this limitation (Lundstrom, 2021). In an effort to find a small RdRp that will not show any prevalence in the human population, we sought to sample invertebrates for novel RNA viruses from which we could build, in two significant steps:

1. The initial aim of this project was to gather RNA samples from variegated sources. Multiple samples from a wide range of invertebrate species provided the necessary heterogeneity from which RNA was isolated. Subsequent construction of a diverse invertebrate RNA library allowed for the identification and classification of viruses present within each sample (regardless of genome type). The RNA library was then used to sequence, assemble and identify putative viruses.

2. Identification of a novel virus was immedately followed by a thorough characterization and analysis of open reading frames (ORFs). Compatibility with cloning and evolutionary relationships to other known viruses can then be assessed. As previously stated, small RNA viruses and/or viral RdRps that neither integrate nor have a high seroprevalence are ideally suited to work with and advance. Subsequent cloning via synthetic biology and launching in permissive cell lines serve as the next steps in the progression and development towards a self-replicating RNA.

Methods and Materials

RNA Isolation from Collected Samples

To address the first aim, insects were collected and stored in RNALater® from predetermined environments in a set area. We recorded the sample ID and suspected species using www.amentsoc.org/insects/what-bug-is-this/. The collected insect was then pulverized with small quantities of TRIzol reagent, with the exact amounts dependent on total sample size. Subsequent incubation allowed for the phenol in TRIzol to break down cellular components while maintaining RNA integrity. Chloroform was added to the solubilized RNA to induce phase separation, which occurred over a fifteen minute period of centrifugation. The generated supernatant contained RNA in the colorless upper aqueous phase and was transferred out of solution via pipetting. The red organic proteinaceous layer and DNA interphase layer were discarded. A quantity of isopropanol, equal to half the added amount of TRIzol Reagent, was mixed into the aqueous solution and allowed to incubate, and the insolubility of RNA in isopropanol yielded a white RNA pellet, albeit impure. Subsequent resuspension in 80% ethanol allowed for purification of the RNA due to ethanol's low dielectric constant and propensity of the salt to dissolve in water and force it out from the RNA. The remaining pellet of RNA was then characterized using a Nanodrop instrument®. This RNA was cataloged and stored at -80°C.

Next Generation Sequencing

High-quality RNA isolated from the previous step (as determined by the Nanodrop) was fragmented and used to generate an Illumina-compatible library for massively-parallel sequencing (see Figure 1). The process followed in this procedure was informed by Michael Quail's literature on the topic (Quail et al., 2009). In brief, each captured

RNA fragment was used to amplify an isolated pool of identical cDNA fragments that could be sequenced alongside each other. Using a high-resolution camera and real-time primer-mediated extension, the NextSeq Illumina instrument can generate 500 million reads from a single run. RNA samples were, therefore, cloned and processed in this way and sequenced and assembled de novo to identify contiguous RNAs that were greater than 6000 nts in length (as this exceeds the size of most mRNAs whereas viruses are commonly larger than this). "Contigs" were then translated in all 6 possible frames and putative proteins (larger than 600aa) were used in a BLASTx search to determine whether there was any homology to known RNA dependent RNA polymerases (RdRps), which are generally larger than 600 residues. Sequenced contigs showing homology to known RdRp were then characterized to identify additional open reading frames (ORFs). Each putative ORF was aligned to known viruses and fitted into a phylogenetic tree to ascertain which family of viruses it was contained within.

Viral Selection

Based on our final list of putative viruses, we prioritized which ones we move forward with using a number of criteria. First, it was essential to have high genetic coverage of the genome to be certain of the viral sequence. For this reason,

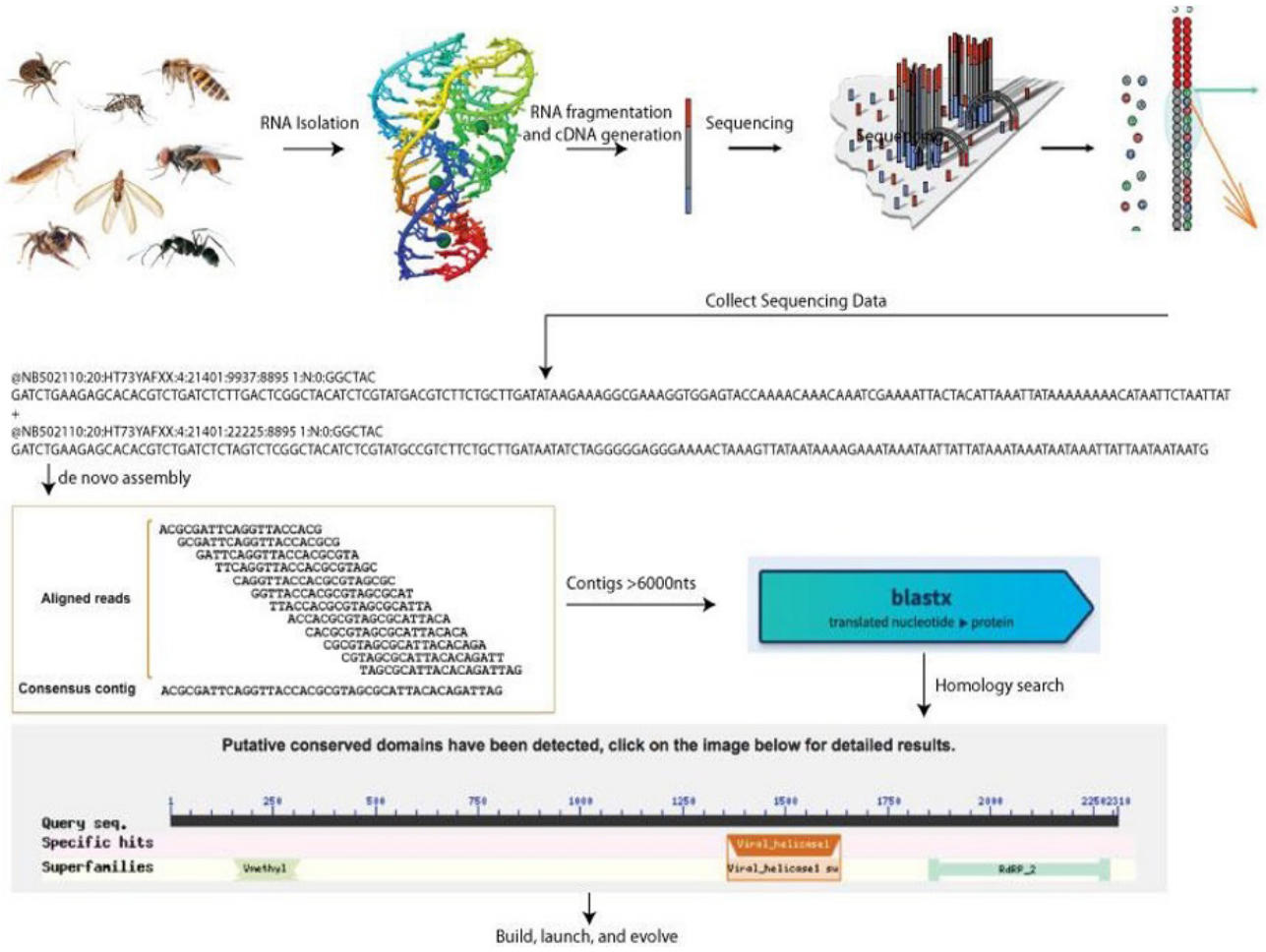

Figure 1: Animation Depicting the Overall Strategy of Data-Collection. The process begins with the collection of invertebrates. The samples then undergo RNA isolation in order for us to sequence them. In the sequencing step, the RNA is fragmented into many pieces before being amplified many times. The fragments will contain overlapping segments of genetic code, making them amenable to reconstruction via de novo assembly. The subsequent consensus contigs vary in length, but the longer ones (>6000 nucleotides (nts)) will be analyzed via Basic Local Arrangement Search Tool (BLAST) to contrast them with known viruses and determine if they are novel viruses. Putative small viruses and/or RdRps will then be synthesized and introduced into mammalian cells to determine if we can select them to function (this step is denoted as Build, launch, and evolve).

we only built viruses that have greater than 10x coverage across the genome at every position. Second, we prioritized viruses that are novel. And third, we chose the smallest viruses that fulfilled the above criteria as they can be synthesized relatively easily. Should we focus on a virus of positive polarity (which can be determined by RdRp homology) we would transcribe RNA and introduce it into cells for further study. Should we discover a virus of negative polarity, we would clone the polymerase into a plasmid to enable host production prior to introducing the genomic RNA for further study.

Results

The first step in interpreting the data involved construction of an RNA library generated from our diverse collection of arthropod and arachnid species. The RNA of 42 individual insects were sampled and analyzed throughout the duration of the project. One insect, the cricket, was split into two sections, for a total of 43 samples. Figure 2a tabulates information about each of these samples, including the insect of origin and the label associated with it. The RNA concentrations and purity are also shown. Of the 43 samples, five were rendered impotent by RNA purities that were too low to sequence. These samples are highlighted red in figure 2a. These results could have been due to human error in the isolation process, or the lack of a quantifiable amount of RNA in the insect genome. The 38 remaining samples were split into twelve different pools for sequencing, labeled A through L. The methodology behind splitting the samples consisted of organizing groups of samples with no overlap between insect types and pairing samples with lower RNA yields to those with higher yields. Loose approximations were made to mix ~1µg of each sample into each pool, for a total of ~3µg in a 100µl solution. Next Generation Sequencing was then performed following the outline described previously. The subsequent sequencing results were organized in an excel spreadsheet by length, and nine out of twelve pools contained "contigs" of greater than 6000 nucleotides. Contigs of this length or greater were considered potential viral candidates, while shorter contigs were disregarded. These remaining contigs were compared against existing libraries of RNA samples to determine if they were viral, and if so, whether they were novel. BLASTx revealed that the vast majority of the contigs had high homology to known viral or other RNA-containing species. In fact, in eleven out of twelve pools, none of the contigs were novel viruses. In Pool D however, two novel viruses were identified and were named Castor and Pollux. Figure 2b reveals a sample output of BLASTx for the longest contigs in Pool D. The longest of these contigs, with a length of 15614 base pairs, was Castor. The second longest, with a length of 12101 base pairs, was Pollux.

Characterization of Castor and Pollux:

Castor and Pollux were identified as viral RNAs by BLASTx because they contained segments with homology to known RNA dependent RNA polymerases (RdRps). These RdRps are essential proteins encoded by RNA viruses that have no DNA stage, and are thus a good but fallible indicator of viral identity. The viral RNAs were thus further characterized to identify additional ORFs. Figure 3a shows the five ORFs identified for Castor. The RdRp segment codes for the RNA-dependent RNA polymerase, and is the longest ORF at 7089bp. It codes for a protein with a Mw upwards of 270da. The nucleoprotein (NP) was identified due to homology with other viral NPs. With a length of 1479bp, it codes for a protein that encapsidates the viral genome and is a necessary element of all negative-sense RNA genomes. Also notable is the spike protein, which is almost certainly involved in penetration and infection of host cells. Figure 3b similarly shows the ORFs of Pollux. Like Castor, there are five identifiable ORFs, and with the exception of the ORF2 (the ORF coding for the spike protein in Castor), the ORFs in Castor and Pollux seem to be well aligned. Figure 3c, which shows the molecular weights of the ORFs, hints at a potential relationship between the viruses since the molecular weights of the RdRp and NP sections are similar. In order to confirm the existence of each of these ORFs, primers were designed and they were amplified and run through a gel. Figure 3d shows the result of one gel run for the Castor

ORFs as an example output; however, every ORF was individually separated and confirmed successfully.

Both RNAs were determined to be single-stranded, negative-sense viruses, and both originated from the family of Rhabdoviridae. Each putative ORF was aligned to known viruses to establish these viral relationships within a phylogenetic tree. Sample outputs of these trees, based on the RdRps of the two viruses, are shown in Figure 4. The nearest relation for both RdRps is an unclassified Coleopteran rhabdovirus. The similar results for each pair of ORFs, in addition to a 43% global homology rating between the two viruses, suggest a relationship and likely a common ancestor. Although confirmation of their origins was not determined, homology searches suggest that these viruses came from the same insect. Based on the loosely conserved RNA sequences present in the sample, we hypothesize that these viruses came from the only spider in pool D (sample #36). It is

Sample #	Sample Origin	Label	Final Concentration (ng/μL)	RNA Purity (260/280)	# μL to obtain ~1μg	Pool
Sample 1a	Cricket	B1	78.9	0.71	NA	NA
Sample 1b	Cricket	B2	105.8	1.07	NA	NA
Sample 2	Black spider	B3	594.4	2.13	2	A
Sample 3	Lady Bug	B4	1278.9	1.58	1.5	B
Sample 4	Pill Bug	B5	144.2	1.94	7	A
Sample 5	Lady Bug	B6	393.6	1.6	3	D
Sample 6	Centipede	B7	1720.6	2.2	1	C
Sample 7	Micro Spider	B8	89.4	1.46	11	I
Sample 8	Yellow Spider	B9	1126.2	1.59	1.5	H
Sample 9	Fly	A1	329.5	1.51	3	E
Sample 10	Mosquito	A2	138.2	1.39	7	B
Sample 11	Mosquito	A3	1560.8	2.16	1	D
Sample 12	Mosquito	A4	1631.3	2.12	1	E
Sample 13	Mosquito	A5	276.2	1.92	4	C
Sample 14	Mosquito	A6	1402.3	2.16	1.5	G
Sample 15	Macro-spider	A7	461.4	1.7	2	G
Sample 16	Mosquito	A8	226.6	1.44	5	H
Sample 17	Spider	A9	42.1	1.5	20	B
Sample 18	Unknown	C1	1504.8	2.05	1	F
Sample 19	Spider	C2	92.5	1.76	10	J
Sample 20	Green Lacewing	C3	163.2	0.71	NA	NA
Sample 21	Cricket	C4	129.9	0.57	NA	NA
Sample 22	Spider	C5	110.2	1.34	10	F
Sample 23	Mosquito	C6	169.7	1.4	8	K
Sample 24	Spider	C7	65.2	1.6	15	C
Sample 25	Mosquito	C8	159.8	1.3	NA	NA
Sample 26	Spider	C9	8.4	1.41	40	K
Sample 27	Spider	D1	86.1	1.34	12	E
Sample 28	Hornet	D2	18.9	1.58	20	A
Sample 29	Mosquito	D3	13.5	1.91	20	A
Sample 30	Mosquito	D4	22.8	1.79	40	L
Sample 31	Mosquito	D5	214.7	1.97	5	I
Sample 32	Mosquito	D6	898.3	2.09	1.5	J
Sample 33	Mosquito	D7	72	1.54	15	G
Sample 34	Mosquito	D8	34.9	1.89	15	F
Sample 35	Spider	D9	13.2	1.92	20	F
Sample 36	Micro Spider	E1	55	1.81	20	D
Sample 37	Fly	E2	876.4	1.89	1.5	H
Sample 38	Fly	E3	1184.9	2.09	1.5	I
Sample 39	Fly	E4	1114.7	2.08	1.5	K
Sample 40	Daddy Long Legs	E5	1441.2	2.05	1.5	L
Sample 41	Wasp	E6	468.1	1.58	2	L
Sample 42	Moth	E7	314.6	1.96	3	J

	Sequence	LENGTH
1_30613 flag=1 multi=181.0000 len=15614	AAAAAAATAAAAAACAAAATTATAATTCATTCTCCTCAGAGGAAATAGGCTCGTTTTTTTAATTTTAAAAGCCTCCATTCTAAC	15614
1_1112 flag=1 multi=220.0000 len=12101	CGGCTAACCCTCTAACCTGATCTCATTACCACTCATGTTCGAATATTATACTGAACGACTATTATACCTACTTCAATAGAAACAC	12101
1_16137 flag=1 multi=43.0000 len=9441	TAGAACTTTTCCCAAACTATTGTCTATCATTCTGGTACAATATGTTTAATGTAGTGAATTCATCTTTGCTAATACAACAAGTCAT	9441
1_43184 flag=1 multi=45.3600 len=7857	AAAAAGACTTCTGATTCTAAATTTAATACTTTGAAAACAGCAAAAAATCATAACAACAGCTACTCTCTGAATCCAAATGTATCT	7857
1_41369 flag=1 multi=37.6024 len=7101	ATCGAAAATTAAAAAGTTGAAACCTTAAGAATTCAATATTCAGAGGAACCTTCATTGGTTACAATTATCACTGATTACCTCAGT	7101

Figure 2: An Overview of the Samples and Initial Data. a, Tabular data showing which insects the 43 samples originated from and the pools into which they were grouped. Note that one insect, the cricket, was large and therefore split into samples 1a and 1b. **b,** Sample data showing some results for Pool D. Each pool generated similar data, with hundreds of contigs of varying lengths. The highlighted segment denotes the automatically generated ID, nucleotide sequence, and length for the first contig, which was the novel virus dubbed "Castor."

It is possible, however, that they came from different insects within the same pool. It is due to their similarity to each other that the two viruses were named after the twins from Greek mythology, Castor and Pollux.

Discussion

Viruses are omnipresent, and yet many viral genomes are unrecognized and undocumented. This project has demonstrated, first and foremost, the potential of next-generation sequencing and de novo genome assembly to expand the invertebrate virosphere. However, despite the discovery and analysis performed in this study, much work remains to be done. After all, with the discovery of two novel viruses comes the introduction of a new suite of tools that could be repurposed for gene editing. To develop such tools, we will first need to verify expression of Castor and Pollux within their cognate RNA samples to ensure and verify their sequences. The negative stranded character of the two viruses suggests the promise of cloning the polymerases into a

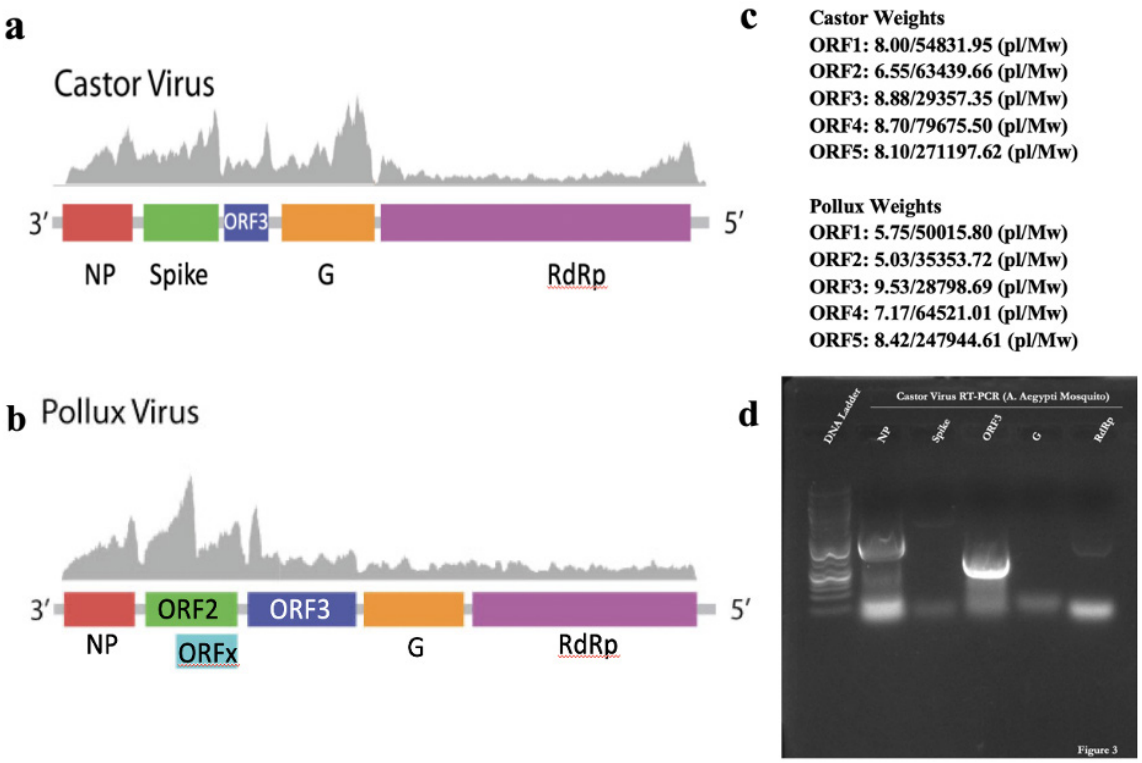

Figure 3: Open-reading Frames of Castor and Pollux. a, A breakdown of the ORFs identified in Castor. The RdRp is the RNA-dependent RNA polymerase, and NP is the nucleoprotein. Five ORFs were identified overall. **b,** A breakdown of the ORFs identified in Pollux. The RdRp and NP regions are similar to those in Castor. Five ORFs were identified overall, but there is no conventional spike protein like that found in Castor. **c,** Provides some information for each of the ORFs found in Castor and Pollux, notably the molecular weights of the corresponding proteins. **d,** An example gel demonstrating confirmation that these ORFs exist, are separable, and are well-defined.

plasmid, so PCR amplification and subcloning into plasmids suitable for in vitro transcription and/or eukaryotic expression is a direct next step. Enabling such host production will be followed by introducing the plasmid into insect or mammalian cells (C6/36 or BHK cells, respectively), and PCR can then be used to determine whether evidence of self-amplification can be observed. Should we see some levels of "replication," we will continue to passage the viruses to determine whether we can guide their activity and study their biology. Successful replication will, in the long term, be followed by additional analysis of the putative ORFs and isolation of the RdRps to qualitatively determine the potential of guided evolution to achieve a functional enzyme in mammalian cells.

Also notable about Castor and Pollux is their close relationship to each other. Initial analysis suggested only one viral discovery, but a closer look quickly demonstrated that two viruses with a high homology were in fact present. An interesting further study could test for interdependence between these two viruses. While it is well-known that viruses are fully dependent on host cell machinery in order to replicate, it would be a novel phenomenon for two viruses to also be dependent on each other.

Continuing to expand the virosphere should be a major scientific goal, and more effort should be put into identifying and characterizing new viral genomes. The fact that Castor and Pollux were discovered in such a small sample size suggests the large number of viruses yet to be discovered. Previous similar experiments have yielded many more viruses in even smaller populations. The invertebrate virosphere contains remarkable variety and flexibility as a result of the frequent rate of recombination and horizontal gene transfer. Continuing to take advantage of such rapid evolution and diversity has the potential to yield numerous novel therapeutic vectors. At the very least, continuing to find such viruses will continue to

expand our knowledge of the virosphere and the diversity and mysteries it contains.

Conclusion

RNA viruses represent one of the greatest sources of biodiversity in the world, and yet knowledge of the many species and families remains limited. Our historical emphasis on studying viruses in cultures or as disease-causing agents has caused us to neglect large and diverse groups of more unremarkable populations. This study sought to begin to analyze one such population—the invertebrate virosphere. By isolating and sequencing the RNA from 42 insects, and creating a diverse RNA library via next-generation sequencing and de novo genome assembly, we were able to identify two novel viruses. These putative novel viral genomes were named Castor and Pollux, and were subsequently characterized and independently confirmed by quantitative PCR. Aligning the ORFs of the newly discovered viruses to preexisting counterparts allowed for the determination that they are single-stranded, negative-sense viruses from the family Rhabdoviridae. While much work remains to be done to achieve real medical progress, Castor and Pollux exemplify the unrecognized and underappreciated diversity and potential of RNA viruses, whose rapid evolution and variable genomic size, structure and segmentation make them wildly promising prospective candidates for various therapeutic applications. The data recovered from these pursuits will not only allow for the development of viral vectors and novel therapies, but will also inform our knowledge of the world around us and provide perspective on the evolutionary intricacies, patterns, and developments within the viral world.

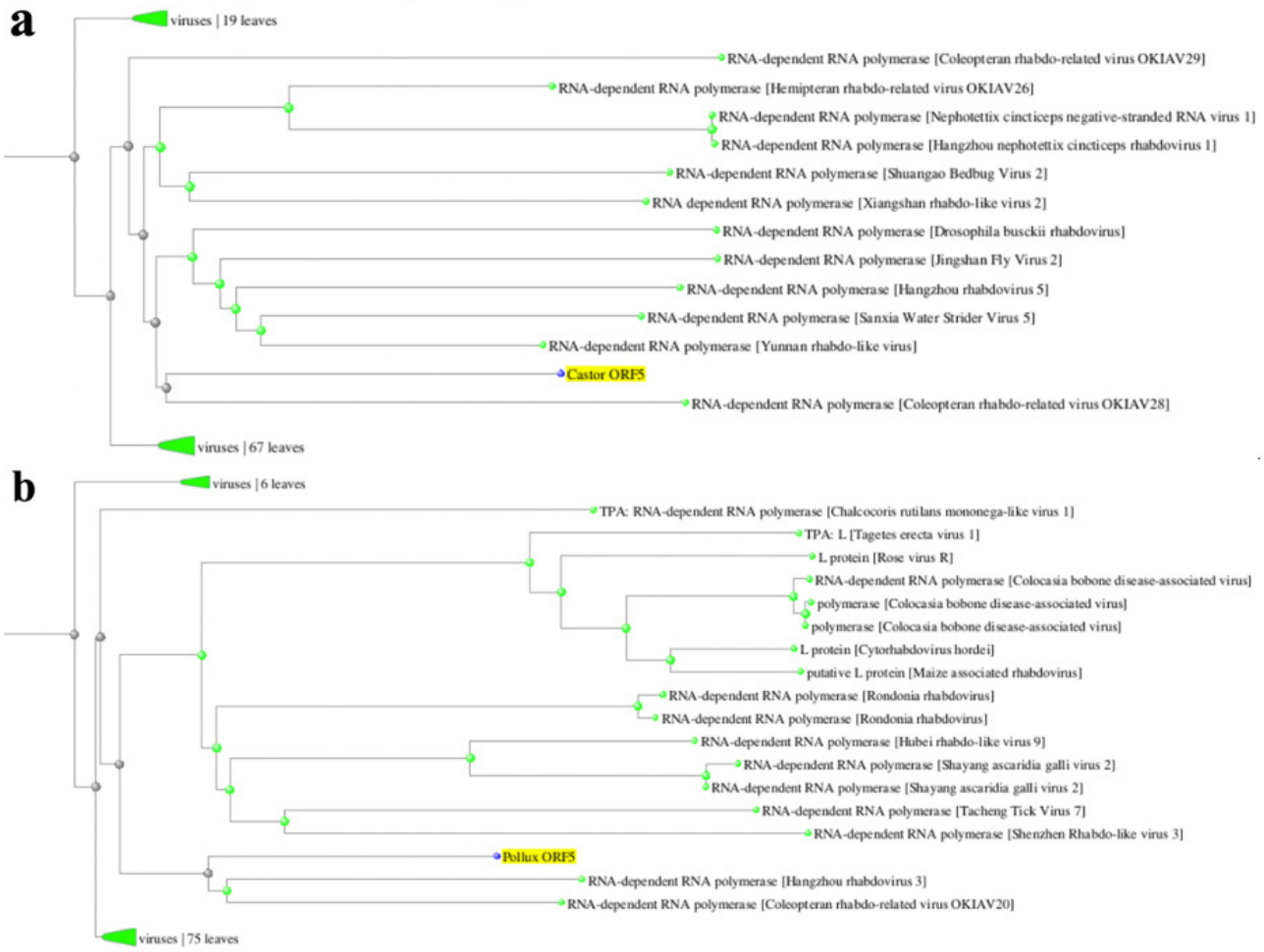

Figure 4: Sample Phylogenetic Trees for Novel Viruses. a, A neighbor-joining phylogenetic tree for Castor's RNA-dependent RNA polymerase (labelled Castor ORF5). **b,** A neighbor joining phylogenetic tree for Pollux's RNA-dependent RNA polymerase (labelled Pollux ORF5).

References

Condiotti, R., Goldenberg, D., Giladi, H., Schnitzer-Perlman, T., Waddington, S. N., Buckley, S. M. K., Heim, D., Cheung, W., Themis, M., Coutelle, C., Simerzin, A., Osejindu, E., Wege, H., Themis, M., & Galun, E. (2014). Transduction of fetal mice with a feline lentiviral vector induces liver tumors which exhibit an E2F activation signature. Molecular Therapy, 22(1), 59–68. https://doi.org/10.1038/mt.2013.193

Dobson, J. (2006). Gene therapy progress and prospects: Magnetic nanoparticle-based Gene Delivery. Gene Therapy, 13(4), 283–287. https://doi.org/10.1038/sj.gt.3302720

Khan, S., Ullah, M. W., Siddique, R., Nabi, G., Manan, S., Yousaf, M., & Hou, H. (2016). Role of recombinant DNA technology to improve life. International Journal of Genomics, 2016, 1–14. https://doi.org/10.1155/2016/2405954

Lundstrom, K. (2021). Self-replicating RNA viruses for vaccine development against infectious diseases and cancer. Vaccines, 9(10), 1187. https://doi.org/10.3390/vaccines9101187

Naso, M. F., Tomkowicz, B., Perry, W. L., & Strohl, W. R. (2017). Adeno-associated virus (AAV) as a vector for gene therapy. BioDrugs, 31(4), 317–334. https://doi.org/10.1007/s40259-017-0234-5

Quail, M. A., Swerdlow, H., & Turner, D. J. (2009). Improved protocols for the Illumina Genome Analyzer Sequencing System. Current Protocols in Human Genetics, 62(1). https://doi.org/10.1002/0471142905.hg1802s62

Robbins, P. D., & Ghivizzani, S. C. (1998). Viral vectors for gene therapy. Pharmacology & Therapeutics, 80(1), 35–47. https://doi.org/10.1016/s0163-7258(98)00020-5

Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., Qin, X.-C., Li, J., Cao, J.-P., Eden, J.-S., Buchmann, J., Wang, W., Xu, J., Holmes, E. C., & Zhang, Y.-Z. (2016). Redefining the invertebrate RNA virosphere. Nature, 540(7634), 539–543. https://doi.org/10.1038/nature20167

Thomas, C. E., Ehrhardt, A., & Kay, M. A. (2003). Progress and problems with the use of viral vectors for gene therapy. Nature Reviews Genetics, 4(5), 346–358. https://doi.org/10.1038/nrg1066

Vemula, S. V., & Mittal, S. K. (2010). Production of adenovirus vectors and their use as a delivery system for influenza vaccines. Expert Opinion on Biological Therapy, 10(10), 1469–1487.

Zhao, Y., & Huang, L. (2014). Lipid nanoparticles for gene delivery. Nonviral Vectors for Gene Therapy - Lipid- and Polymer-Based Gene Transfer, 13–36. https://doi.org/10.1016/b978-0-12-800148-6.00002-x

Acknowledgements

I would like to thank the Urban Barcode Research Program of Cold Spring Harbor Laboratory, Dr. Christine Marizzi of the Harlem DNA Lab, and the Pinkerton Foundation for providing me with the opportunity to perform this research. Thank you also to the Icahn School of Medicine at Mount Sinai for laboratory facilities, and specifically to Dr. tenOever for his unparalleled guidance throughout the research process.