Diversity and Evolution of Viroids and Viroid-like RNA Agents

Benjamin D. Lee

Christ Church College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2024

Abstract

Viroids are the smallest and simplest known pathogens. Comprised of circular noncoding RNA, these agents can cause fatal diseases in plants. Although other viroid-like agents such as satellite RNAs, ribozyviruses, and retrozymes have been discovered, the evolution and true extent of viroids and viroid-like RNA agents remains poorly understood.

The primary aim of this project was to examine the diversity and evolution of viroids and viroid-like RNA agents. To achieve this goal, a comprehensive database of relevant sequences was curated and a specialized software pipeline for the detection of viroid-like sequences was built. The database provides a value-added resource of sequences, sample metadata, and annotations. The pipeline detects circular sequences from input transcriptomes or metatranscriptomes and then searches against the viroid database and a set of self-cleaving ribozyme profiles. Viroid-like sequences are then clustered into species-level clusters and their secondary structures predicted.

The pipeline was tested on a diverse set of plant transcriptomes to verify its ability to recover known viroid sequences. It was then applied to a large collection of global metatranscriptomes from varied ecosystems. The search yielded a fivefold increase of species-level diversity compared to the set of known viroids and viroid-like sequences. The results also included novel ribozyme content within the viroid-like sequences. Second, new clades of ribozyviruses were identified among the circular RNAs. Third, the search identified the viroid-like nature of two additional types of small RNA viruses, both of which infect fungi and/or protists and lack capsids. Finally, CRISPR spacer matches to viroid-like sequences were detected, suggesting that the host range of viroid-like agents extends to prokaryotes.

This project was the first large-scale metatranscriptomic investigation of viroids and viroid-like agents. The results reveal that their true diversity of these unique agents has been previously understated and opens up new questions into their origin and evolutionary history.

Diversity and Evolution of Viroids and Viroid-like RNA Agents

Benjamin D. Lee

Christ Church College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2024

To my parents, without whom I would not be where I am today

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisors, Prof. Peter Simmonds and Dr. Eugene Koonin, for their guidance and support throughout this project. Their vast knowledge and invaluable advice have been instrumental in the completion of this work.

I would also like to thank my colleagues at the National Center for Biotechnology Information (NCBI) for their collaboration and thought-provoking discussions. In particular, I would like to thank Drs. Yuri Wolf and Kira Makarova.

I also extend my deep appreciation to Uri Neri at Tel Aviv University for his assistance and friendship throughout the past several years. Without his help, this project would not have been possible.

I must also acknowledge the dozens of collaborators who have contributed their data from around the world to the Integrated Microbial Genomes and Microbiomes (IMG/M) database.

Institutional support for this work was provided by the National Institutes of Health Oxford-Cambridge Scholars Program and the National Center for Biotechnology Information of the National Library of Medicine. This unique program has enabled me to pursue my research interests on both sides of the Atlantic and has provided me with a wealth of opportunities for personal and professional growth.

On a personal level, I owe a debt of gratitude to my wife, Elizabeth, for her patience and understanding throughout my DPhil. Throughout the many late nights and weekends spent working on this project, she has been a constant source of support and encouragement.

Finally, I would like to thank my parents, Jeremy and Elisabeth, for their love and support throughout my life. From culturing phages at home in middle school to sending me to virus camp in high school, they have always encouraged my scientific curiosity. I dedicate this work to them.

Abstract

Viroids are the smallest and simplest known pathogens. Comprised of circular noncoding RNA, these agents can cause fatal diseases in plants. Although other viroid-like agents such as satellite RNAs, ribozyviruses, and retrozymes have been discovered, the evolution and true extent of viroids and viroid-like RNA agents remains poorly understood.

The primary aim of this project was to examine the diversity and evolution of viroids and viroid-like RNA agents. To achieve this goal, a comprehensive database of relevant sequences was curated and a specialized software pipeline for the detection of viroid-like sequences was built. The database provides a value-added resource of sequences, sample metadata, and annotations. The pipeline detects circular sequences from input transcriptomes or metatranscriptomes and then searches against the viroid database and a set of self-cleaving ribozyme profiles. Viroid-like sequences are then clustered into species-level clusters and their secondary structures predicted.

The pipeline was tested on a diverse set of plant transcriptomes to verify its ability to recover known viroid sequences. It was then applied to a large collection of global metatranscriptomes from varied ecosystems. The search yielded a fivefold increase of species-level diversity compared to the set of known viroids and viroid-like sequences. The results also included novel ribozyme content within the viroid-like sequences. Second, new clades of ribozyviruses were identified among the circular RNAs. Third, the search identified the viroid-like nature of two additional types of small RNA viruses, both of which infect fungi and/or protists and lack capsids. Finally, CRISPR spacer matches to viroid-like sequences were detected, suggesting that the host range of viroid-like agents extends to prokaryotes.

This project was the first large-scale metatranscriptomic investigation of viroids and viroid-like agents. The results reveal that their true diversity of these unique agents has been previously understated and opens up new questions into their origin and evolutionary history.

Table of contents

List of Figures xv

List of Tables xvii

Preface 1

1 Introduction 3

1.1 Viroids . . . 4

1.1.1 Taxonomy . . . 4

1.1.2 Pathogenicity and RNA silencing . . . 6

1.1.3 Secondary structure and replication . . . 7

1.1.4 Prevention, transmission, and treatment . . . 16

1.1.5 Distribution and host range . . . 17

1.1.6 Economic impact . . . 19

1.2 Satellite RNAs . . . 20

1.2.1 Large satRNAs and small linear satRNAs . . . 22

1.2.2 Pathogenicity and RNA silencing . . . 23

1.2.3 Replication . . . 24

1.2.4 Taxonomy . . . 25

1.2.5 Potential protein-coding capacity . . . 26

1.2.6 Economic impact . . . 27

1.3 Ribozyviruses . . . 27

1.3.1 Human hepatitis delta virus . . . 28

1.3.2 Structure and replication 31
1.3.3 Symptoms 34
1.3.4 Pathogenicity 35
1.4 Retrozymes 36
1.5 Retroviroids 37
1.6 Ribozymes 38
1.7 Origin and evolution of viroid-like RNAs 40
1.7.1 Possible mechanisms of emergence 41
1.7.2 Common ancestry versus multiple origins 45
1.7.3 Retrozymes as a possible ancestor 46
1.8 Conclusion 48

2 A database of viroids and viroid-like circular RNAs 51

Summary 51
2.1 Introduction 51
2.2 Database content 52
2.2.1 Sequence data 54
2.2.2 Secondary structures 55
2.2.3 Predicted ribozymes 55
2.2.4 Sequence clusters 56
2.3 Data storage and infrastructure 58
2.4 Data query and access 59
2.4.1 ViroidDB website 59
2.4.2 Data portal 65
2.5 Future directions 66
2.6 Conclusion 67

3 A method for identifying viroids and viroid-like RNAs **69**

Summary 69

3.1 Introduction 69

3.2 Prior work 71

3.3 Pipeline implementation 72

3.3.1 Circularity detection 73

3.3.2 cccRNA deduplication 76

3.3.3 Ribozyme-based filtering 78

3.3.4 Direct sequence search 79

3.3.5 RNA secondary structure prediction 79

3.3.6 Clustering 80

3.3.7 ORF prediction 81

3.4 Key features 82

3.4.1 Simplicity and ease of use 82

3.4.2 Scalability 83

3.4.3 Portability 84

3.5 Validation 84

3.6 Conclusion 87

4 A vast world of viroid-like circular RNAs **89**

Summary 89

4.1 Introduction 90

4.2 Methods 91

4.2.1 Data acquisition 91

4.2.2 Viroid-like cccRNA detection 92

4.2.3 Clustering 93

4.2.4 Protein searches 93

4.2.5 HDV antigen analysis 94

4.2.6 Read mapping 94
4.2.7 CRISPR spacer analysis 94
4.2.8 Software resources 96
4.3 Results 97
4.3.1 A five-fold expansion of the known diversity of viroid-like cccRNAs 97
4.3.2 Putative novel viroid-like cccRNAs 100
4.3.3 Geographic and ecological distribution of viroid-like cccRNAs 104
4.3.4 Virus-like elements blurring lines between riboviruses, ribozyviruses and viroids 108
4.3.5 Diverse deltavirus-like viruses 112
4.3.6 Diverse ribozymes and ribozyme combinations 117
4.3.7 CRISPR spacers matching cccRNAs 119
4.4 Conclusion 121

5 An advanced tool for circular sequence analysis 127
Summary 127
5.1 Introduction 127
5.2 Language choice: Rust 130
5.3 Algorithms implemented 131
5.3.1 Circularity detection and monomerisation 131
5.3.2 canonicalisation and deduplication 134
5.3.3 ORF prediction 135
5.3.4 Concatenation and deconcatenation 138
5.3.5 Sequence rotation 139
5.4 Test suite 139
5.5 Compressed data I/O 141
5.6 Conclusion 141

6 Summary and implications **143**

6.1 A new era of viroid-like RNA diversity 143

6.2 New vocabulary is needed 145

6.3 New taxonomy is needed 146

6.4 Can a replicator be both a viroid and a virus? 147

6.5 Ribozymes as the key to the subviral world 148

6.6 The need to analyse more (meta)transcriptomes 149

6.7 Viroids as paragons of minimal replicator emergence 151

6.8 Viroid-like RNAs and emergence of protein coding 154

6.9 An expanding host range 155

6.10 Evolution and possible origins of viroid-like agents 156

6.11 A new era of viroid research 157

References **159**

Appendices **203**

A Statement on re-use of published materials **203**

List of Figures

1.1 Schematic structures of distinct classes of viroid-like RNAs 5

1.2 Representative secondary structures of viroids, produced using RNAfold 8

1.3 The asymmetric rolling circle replication mechanism 10

1.4 The symmetric rolling circle replication mechanism 11

1.5 Representative secondary structures of satRNAs, produced using RNAfold 21

1.6 Secondary structure of human hepatitis delta virus (HDV) 30

1.7 Representative secondary structures of retroviroids, produced using RNAfold 37

1.8 Evolutionary scenarios for the emergence of viroid-like replicators . 42

1.9 Secondary structure of a satRNA that reportedly emerged *de novo* . 43

2.1 A flowchart of the complete ViroidDB data processing and release pipeline 53

2.2 Metadata panels on the sequence detail page 60

2.3 Sequence viewer on the sequence detail page 60

2.4 Ribozyme predictions on the sequence detail page 61

2.5 Secondary structures and predictions on the sequence detail page . 62

2.6 Cluster listing page of ViroidDB 63

2.7 Cluster detail page 64

2.8 The search page of ViroidDB demonstrates the site's fuzzy search, search-match highlighting, taxonomic refinement lists, and species sub-search 65

3.1 The `vdsearch` command-line interface 72
3.2 A CbVd-like sequence detected using `vdsearch` 85

4.1 Viroid-like cccRNA detection pipeline 91
4.2 Viroid-like cccRNAs identified in metatranscriptomes 99
4.3 Predicted secondary structures of representatives of the five largest clusters of distinct viroid-like cccRNAs 101
4.4 Map of sample locations from which viroid-like cccRNAs were detected 104
4.5 Proportion of samples in each ecosystem type that contained one or more distinct viroid-like cccRNA clusters 106
4.6 Relative frequencies of ribozyme combinations within symmetric cccRNA clusters in each ecosystem type. 107
4.7 Genomic and secondary structure of viroid-like viruses 110
4.8 Clustering of the HDV antigen (Ag)-like protein homologs based on their sequence similarity 114
4.9 Schematic representation of the HDVAg 115
4.10 Octameric structure of the conserved dimerisation domain of HDVAg 115
4.11 Comparison of the sequence conservation among reference HDVAg from GenBank and those from metatranscriptomic datasets 116
4.12 Maximum likelihood phylogeny of HDVAg-like sequences 116
4.13 Ribozyme diversity in viroid-like cccRNAs 117

5.1 A conceptual schematic of viroid identification pipeline 128
5.2 The `circKit` command-line interface 131
5.3 A jagged stop codon index array 136

List of Tables

1.1 The major types of previously known viroid-like cccRNAs 3

2.1 The genera and number of sequences in the 2021-09-07 release of ViroidDB. 55

2.2 Cluster counts for varying identity thresholds (75% minimum reciprocal overlap required) 57

4.1 Software used for viroid-like cccRNA detection and analysis 96

4.2 Summary of the five largest clusters of novel viroid-like cccRNAs . 100

4.3 Summary of viroid-like cccRNA detection in different ecosystems . 104

5.1 Next codon stop index search frame when wrapping around 137

6.1 The major types of known viroid-like cccRNAs (updated) 145

Preface

Parasites are ubiquitous in nature. They are found in every ecosystem, from the deepest oceans to the highest mountains. They infect every form of life, from the smallest bacteria to the largest mammals.

One general feature of parasitism is that parasites are smaller than their hosts. For example, the parasite that causes leishmaniasis is a single-celled protozoan that is only a few microns in size—so small that it is able to hide from the immune system by living inside the infected person's own cells. In addition to physical size, this principle also applies to the size of the parasite's genetic material. Parasites are under selective pressure to minimize their genome size because smaller, simpler genomes are easier to replicate and transmit. What are the limits of this genome minimisation?

Viroids are the answer to this question. They are the smallest known parasites, with RNA genomes of only a few hundred nucleotides. These genomes are so small that they do not encode any proteins, and instead rely on host enzymes for replication. Surprisingly, despite their small size and extreme simplicity, only a few dozen viroids had been discovered in the half century between their discovery by T. O. Diener[1] and when I began my thesis. This begged the question: are viroids really so rare or are they simply difficult to find?

In short, when searching with the right tools, viroid-like RNAs are everywhere. Not only is the absolute number of viroids and viroid-like RNAs much greater than previously known, but they are also much more diverse. This diversity has

[1] "Potato Spindle Tuber 'Virus'," *Virology* 45, no. 2 (August 1971): 411–428, doi:10.1016/0042-6822(71)90342-4.

significant implications for the origins of viroids and breathes new life into the hypothesis that viroids may be remnants from the earliest stages of life on Earth.

At the beginning of my DPhil, the field of subviral biology was a sleepy backwater compared to that of virology as a whole. While the number of known viruses was growing exponentially, the number of known viroids and viroid-like RNA agents was growing linearly. This is no hyperbole: in the 50 years since the discovery of the first viroid, the total number of known viroids and viroid-like species had increased by less than 50.[2]

In retrospect, I am pleased to note that this is no longer the case.

It is my sincere hope that this thesis will appear quaint and outdated in a few years' time, as the field of viroid-like RNA biology continues to grow and mature. Mentions of the several dozen viroid species will hopefully be replaced by references to the hundreds of species that have been discovered, characterized, and classified in the intervening years.

[2] Michela Chiumenti et al., "Reassessing Species Demarcation Criteria in Viroid Taxonomy by Pairwise Identity Matrices," *Virus Evolution* 7, no. 1 (January 2021): veab001, doi:10.1093/ve/veab001.

1 Introduction

This thesis concerns itself with a group of small, generally non-coding, covalently closed circular RNAs (cccRNAs) that are found in plants and other organisms. Together, these agents have been referred to as the viroid-like "brotherhood,"[1] summarized in Table 1.1 and depicted in Figure 1.1. Before proceeding further, I will briefly introduce the members of this brotherhood known at the commencement of work on this thesis and their key features. Unless otherwise noted, information presented here reflects the state of knowledge as of Michaelmas 2020 when the research for this thesis began. An updated review, featuring the results of this work and the work of others, is provided in Chapter 6.

Table 1.1: The major types of previously known viroid-like cccRNAs

Viroid-like RNA	Size	Host	Ribozymes	Coding Capacity
Viroids	246–450 nt	Plants	Hammerhead when present	None
Satellite RNAs	220–457 nt	Plants	Hammerhead or Hairpin	None (except satRYMV)

[1] Lynda Rocheleau and Martin Pelchat, "The Subviral RNA Database: A Toolbox for Viroids, the Hepatitis Delta Virus and Satellite RNAs Research," *BMC Microbiology* 6 (March 2006): 24, doi:10.1186/1471-2180-6-24.

Viroid-like RNA	Size	Host	Ribozymes	Coding Capacity
Ribozyviruses	1547–1735 nt	Metazoans	HDVR or Hammerhead	One conserved protein
Retrozymes	174–1116 nt	Eukaryotic genomes	Hammerhead	None
Retroviroids	275 nt	Carnations	Hammerhead	None

1.1 Viroids

Viroids were the first subviral agents to be discovered and remain the most diverse group in the brotherhood. First described by Diener in 1967[2] and named in 1971,[3] these agents are unencapsidated cccRNAs 200–400 nt in length that can cause diseases in plants, some of which are fatal. Indeed, the first viroid discovered, potato spindle tuber viroid (PSTVd), was initially supposed to be a virus until evidence of its unique properties emerged.[4] Viroids are the smallest and simplest RNA replicators, with the only smaller known pathogen being the prion,[5] a proteinaceous infectious agent.

1.1.1 Taxonomy

Over the years, two families of viroids have been recognized: *Pospiviroidae* and *Avsunviroidae*.[6] In the family *Pospiviroidae*, five genera have been recognized: *Pospiviroid*, *Hostuviroid*, *Apscaviroid*, *Coleviroid*, and *Pelamoviroid*. These genera

[2] T. O. Diener and W. B. Raymer, "Potato Spindle Tuber Virus: A Plant Virus with Properties of a Free Nucleic Acid," *Science* 158, no. 3799 (October 1967): 378–381, doi:10.1126/science.158.3799.378; W. B. Raymer and T. O. Diener, "Potato Spindle Tuber Virus: A Plant Virus with Properties of a Free Nucleic Acid," *Virology* 37, no. 3 (March 1969): 343–350, doi:10.1016/0042-6822(69)90218-9; T. O. Diener and W. B. Raymer, "Potato Spindle Tuber Virus: A Plant Virus with Properties of a Free Nucleic Acid," *Virology* 37, no. 3 (March 1969): 351–366, doi:10.1016/0042-6822(69)90219-0; B. David Stollar and T. O. Diener, "Potato Spindle Tuber Viroid: V. Failure of Immunological Tests to Disclose Double-Stranded RNA or RNA-DNA Hybrids," *Virology* 46, no. 1 (October 1971): 168–170, doi:10.1016/0042-6822(71)90017-1.

[3] Diener, "Potato Spindle Tuber 'Virus'."

[4] Diener.

[5] T. O. Diener, M. P. McKinley, and S. B. Prusiner, "Viroids and Prions," *Proceedings of the National Academy of Sciences of the United States of America* 79, no. 17 (September 1982): 5220–5224, doi:10.1073/pnas.79.17.5220.

[6] Francesco Di Serio et al., "Viroid Taxonomy," in *Viroids and Satellites* (Elsevier, 2017), 135–146, doi:10.1016/B978-0-12-801498-1.00013-9.

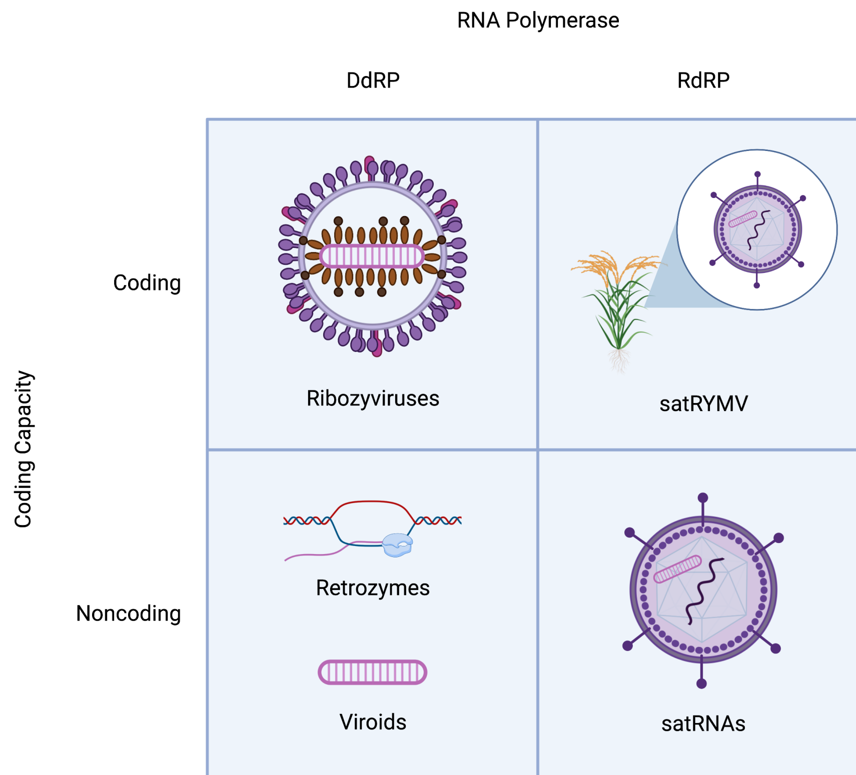


Figure 1.1: Schematic structures of distinct classes of viroid-like RNAs. Viroid-like agents are divided into groups depending on whether they encode proteins and are replicated by DNA-dependent or RNA-dependent RNA polymerase. The satRNAs (lower right quadrant) are encapsidated by the helper virus capsid proteins. In one case (top right quadrant), a satRNA of rice yellow mottle virus (satRYMV) appears to encode a protein.

currently contain 39 species, with the best known being PSTVd, the type species of the family. In the family *Avsunviroidae*, three genera have been recognized: *Avsunviroid*, *Elaviroid*, and *Pelamoviroid*. These genera currently contain 5 species, with the type species of the family being avocado sunblotch viroid (ASBVd). In sum, there are 44 recognized species of viroids known fifty years since their discovery, a rate of roughly one new species every year that stands in stark contrast to the dramatically higher rate of discovery of new viruses.[7]

[7]Yuri I. Wolf et al., "Doubling of the Known Set of RNA Viruses by Metagenomic Analysis of an Aquatic Virome," *Nature Microbiology*, July 2020, doi:10.1038/s41564-020-0755-4; Uri Neri et

The demarcation criteria for viroids differs from that of viruses. While viruses can be classified purely based on their nucleotide sequences,[8] viroids still require biological evidence for classification.[9] Sequence-based classification has received some consideration using pairwise sequence identity matrices to update the arbitrary 90% sequence identity threshold used to classify viroids.[10] One challenge sequence-based classification faces is that of recombinant viroids, which can have sequences from multiple viroids.[11] Ultimately, the biological criteria for classification remain the same: viroids must be able to demonstrably replicate autonomously in a host and have unique biological features.

1.1.2 Pathogenicity and RNA silencing

That viroids can cause disease at all is remarkable given that they do not encode any proteins. Indeed, had viroids not been pathogenic, they may have eluded discovery for much longer. Some viroids are in fact latent, causing no discernible symptoms in their hosts.[12] It must be noted that the distinction between latent and non-latent viroids is not always clear-cut. Viroids—some despite the appellation "latent"—can cause symptoms in *some* hosts, such as the peach latent mosaic viroid (PLMVd) in peach trees.[13] On the other end of the spectrum of pathogenicity, viroid infection can lead to the death of the host plant. For example, the infection of coconut palms with coconut cadang-cadang viroid (CCCVd) leads to the death of the tree.[14]

---

al., "Expansion of the Global RNA Virome Reveals Diverse Clades of Bacteriophages," *Cell* 185, no. 21 (October 2022): 4023–4037.e18, doi:10.1016/j.cell.2022.08.023.

[8] Peter Simmonds et al., "Virus Taxonomy in the Age of Metagenomics," *Nature Reviews Microbiology* 15, no. 3 (March 2017): 161–168, doi:10.1038/nrmicro.2016.177.

[9] Di Serio et al., "Viroid Taxonomy."

[10] Chiumenti et al., "Reassessing Species Demarcation Criteria in Viroid Taxonomy by Pairwise Identity Matrices."

[11] Di Serio et al., "Viroid Taxonomy."

[12] Jacobus Th.J. Verhoeven et al., "Dahlia Latent Viroid," in *Viroids and Satellites* (Elsevier, 2017), 211–216, doi:10.1016/B978-0-12-801498-1.00020-6; José-Antonio Daròs, "Eggplant Latent Viroid," in *Viroids and Satellites* (Elsevier, 2017), 339–344, doi:10.1016/B978-0-12-801498-1.00032-2.

[13] Ricardo Flores et al., "Peach Latent Mosaic Viroid in Infected Peach," in *Viroids and Satellites* (Elsevier, 2017), 307–316, doi:10.1016/B978-0-12-801498-1.00029-2.

[14] Ganesan Vadamalai et al., "Coconut Cadang-Cadang Viroid and Coconut Tinangaja Viroid," in *Viroids and Satellites* (Elsevier, 2017), 263–273, doi:10.1016/B978-0-12-801498-1.00025-5.

The mechanisms by which viroids cause disease are not fully understood, but they are thought to involve the induction of RNA silencing, a host defense mechanism against viruses and other parasitic nucleic acids.[15] Under this model, viroids are recognized by the host RNA silencing machinery, which looks for double-stranded RNA (dsRNA) as a sign of infection. The host machinery then cleaves the viroid RNA into small RNAs (sRNAs), which are then used to guide the host machinery to other viroid RNAs.[16] However, in addition to targeting host machinery against the viroids themselves, the viroid-derived sRNAs also target host mRNAs, leading to the symptoms of disease.

An example of this model is the mechanism by which PLMVd causes disease in peach trees. Peach calico, an extreme form of leaf and fruit albinism, is one symptom of the disease. Notably, PLMVd is a member of the *Avsunviroidae* family, which replicate in the chloroplasts of plant cells.[17] A specific 12 nt sequence within PLMVd is responsible for the symptoms of peach calico, as it is complementary to an mRNA encoding chloroplastic heat shock protein 90 (cHSP90).[18] This protein is critical for chloroplast development and its downregulation via RNA silencing leads to the symptoms of peach calico.[19]

In the case of RNA silencing, the sequence of the viroid is critical for its pathogenicity. However, viroid biology is also influenced by its secondary structure, which is critical for its replication and stability.

1.1.3 Secondary structure and replication

[15] Francesco Di Serio et al., "Role of RNA Silencing in Plant-Viroid Interactions and in Viroid Pathogenesis," *Virus Research* 323 (January 2023): 198964, doi:10.1016/j.virusres.2022.198964.

[16] Ricardo Flores et al., "Viroid Pathogenesis," in *Viroids and Satellites* (Elsevier, 2017), 93–103, doi:10.1016/B978-0-12-801498-1.00009-7.

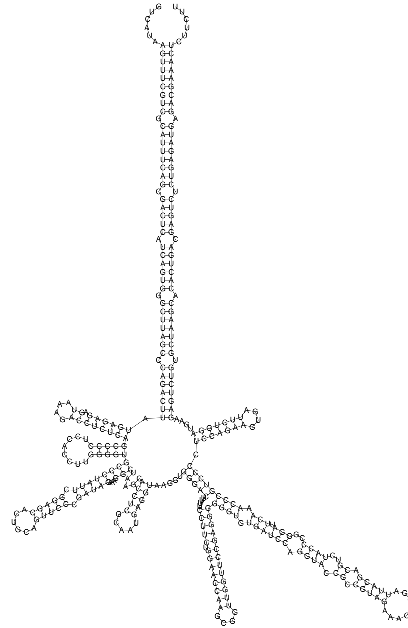
[17] Ricardo Flores, Sofia Minoia, et al., "Viroid Replication," in *Viroids and Satellites* (Elsevier, 2017), 71–81, doi:10.1016/B978-0-12-801498-1.00007-3.

[18] Beatriz Navarro et al., "Small RNAs Containing the Pathogenic Determinant of a Chloroplast-Replicating Viroid Guide the Degradation of a Host mRNA as Predicted by RNA Silencing," *The Plant Journal* 70, no. 6 (June 2012): 991–1003, doi:10.1111/j.1365-313X.2012.04940.x.

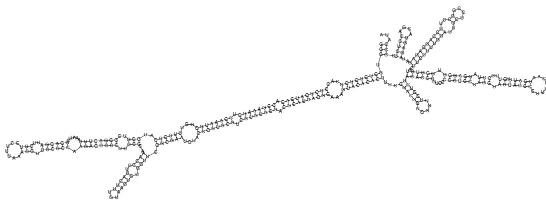
[19] Maria-Elena Rodio et al., "A Viroid RNA with a Specific Structural Motif Inhibits Chloroplast Development," *The Plant Cell* 19, no. 11 (December 2007): 3610–3626, doi:10.1105/tpc.106.049775.

[20] Ronny Lorenz et al., "ViennaRNA Package 2.0," *Algorithms for Molecular Biology* 6, no. 1 (December 2011): 26, doi:10.1186/1748-7188-6-26.

(a) Potato spindle tuber viroid (PSTVd)



(b) Peach latent mosaic viroid (PLMVd)



(c) Eggplant latent viroid (ELVd)

Figure 1.2: Representative secondary structures of viroids, produced using RNAfold[20]

Viroid genomes exhibit extensive self-complementarity resulting in compact, robust structures.[21] While many viroids are rod shaped such as PSTVd shown in Figure 1.2a, others viroids such as PLMVd fold into a branched conformation, shown in Figure 1.2b. PSTVd has 71.9% of its nucleotides involved in base pairing in the (+) polarity, while PLMVd has 71.8% and ASBVd has 68% pairing.[22] This robust secondary structure of viroids gives them considerable thermal stability[23] while simultaneously providing resistance to cleavage by endonucleases that target dsRNA.[24] It also grants PSTVd the ability to remain infective after 24 hours on a variety of surfaces.[25] Most importantly, the secondary structure of viroids is critical for their replication.

Viroids replicate via the rolling circle replication (RCR) mechanism, in which the circular RNA is reiteratively transcribed by hijacked host DNA-dependent RNA polymerase (DdRP) into multimeric intermediates of the opposite polarity.[26] The

[21] Tamara Giguère, Charith Raj Adkar-Purushothama, and Jean-Pierre Perreault, "Comprehensive Secondary Structure Elucidation of Four Genera of the Family Pospiviroidae," *PLOS One* 9, no. 6 (2014): e98655, doi:10.1371/journal.pone.0098655; Tamara Giguère et al., "Elucidation of the Structures of All Members of the Avsunviroidae Family," *Molecular Plant Pathology* 15, no. 8 (October 2014): 767–779, doi:10.1111/mpp.12130.

[22] Benjamin D. Lee et al., "ViroidDB: A Database of Viroids and Viroid-Like Circular RNAs," *Nucleic Acids Research*, November 2021, gkab974, doi:10.1093/nar/gkab974.

[23] H. L. Sanger et al., "Viroids Are Single-Stranded Covalently Closed Circular RNA Molecules Existing as Highly Base-Paired Rod-Like Structures." *Proceedings of the National Academy of Sciences* 73, no. 11 (November 1976): 3852–3856, doi:10.1073/pnas.73.11.3852.

[24] Jinhong Chang, Patrick Provost, and John M. Taylor, "Resistance of Human Hepatitis Delta Virus RNAs to Dicer Activity," *Journal of Virology* 77, no. 22 (November 2003): 11910–11917, doi:10.1128/JVI.77.22.11910-11917.2003.

[25] A. E. Mackie et al., "Potato Spindle Tuber Viroid: Stability on Common Surfaces and Inactivation With Disinfectants," *Plant Disease* 99, no. 6 (June 2015): 770–775, doi:10.1094/PDIS-09-14-0929-RE.

[26] Ricardo Flores et al., "Rolling-Circle Replication of Viroids, Viroid-Like Satellite RNAs and Hepatitis Delta Virus: Variations on a Theme," *RNA Biology* 8, no. 2 (March 2011): 200–206, doi:10.4161/rna.8.2.14238; Andrea D. Branch, Bonnie J. Benenfeld, and Hugh D. Robertson, "Evidence for a Single Rolling Circle in the Replication of Potato Spindle Tuber Viroid," *Proceedings of the National Academy of Sciences* 85, no. 23 (December 1988): 9128–9132, doi:10.1073/pnas.85.23.9128; J. A. Daròs et al., "Replication of Avocado Sunblotch Viroid: Evidence for a Symmetric Pathway with Two Rolling Circles and Hammerhead Ribozyme Processing," *Proceedings of the National Academy of Sciences of the United States of America* 91, no. 26 (December 1994): 12813–12817, doi:10.1073/pnas.91.26.12813; Thomas B. Macnaughton et al., "Rolling Circle Replication of Hepatitis Delta Virus RNA Is Carried Out by Two Different Cellular RNA Polymerases," *Journal of Virology* 76, no. 8 (April 2002): 3920–3927, doi:10.1128/jvi.76.8.3920-3927.2002; Ricardo Flores, María-Eugenia Gas, et al., "Viroid Replication: Rolling-Circles, Enzymes and Ribozymes," *Viruses* 1, no. 2 (September 2009): 317–334, doi:10.3390/v1020317; R. Flores et al., "Viroids, the Simplest RNA Replicons: How They

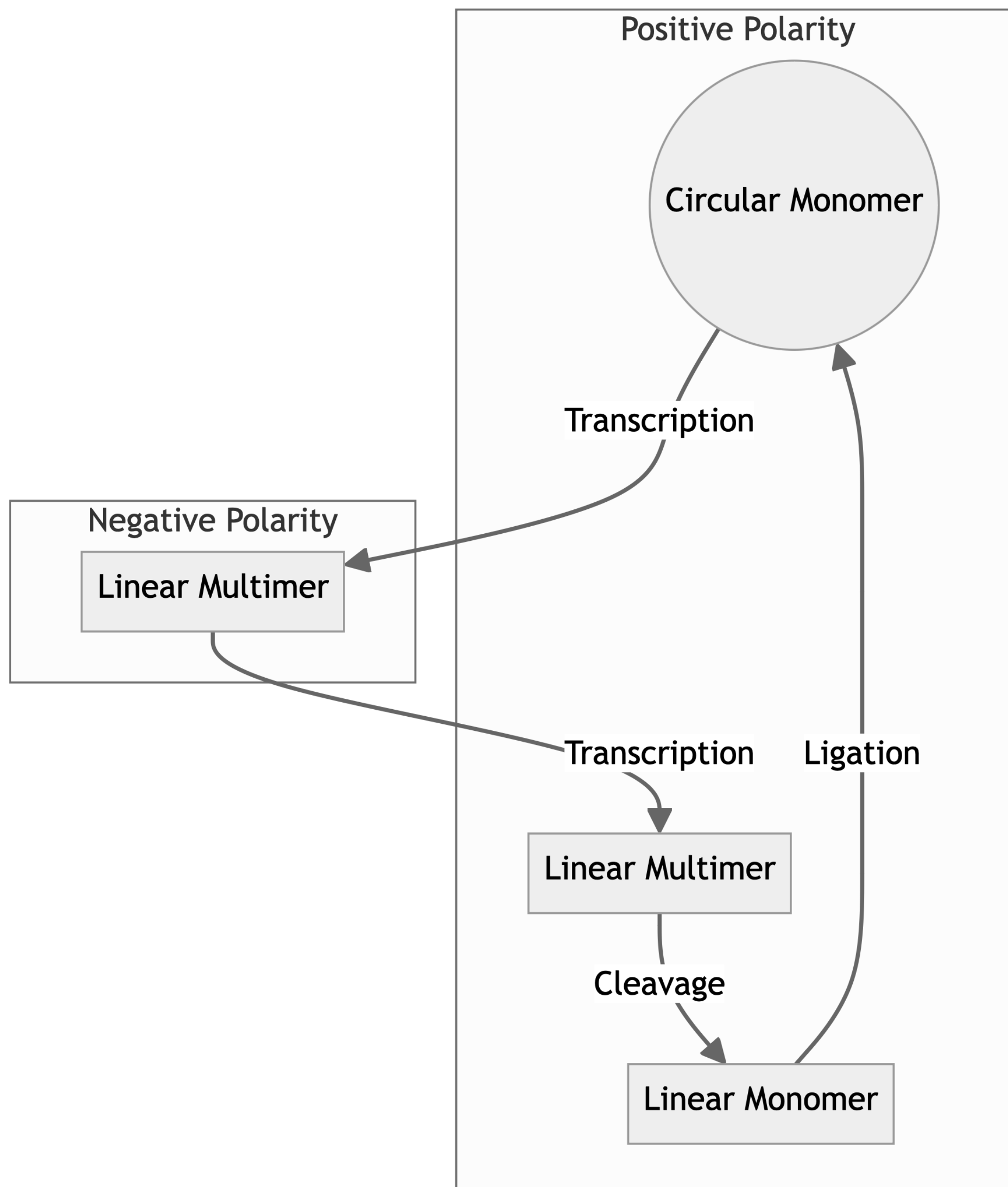


Figure 1.3: The asymmetric rolling circle replication mechanism

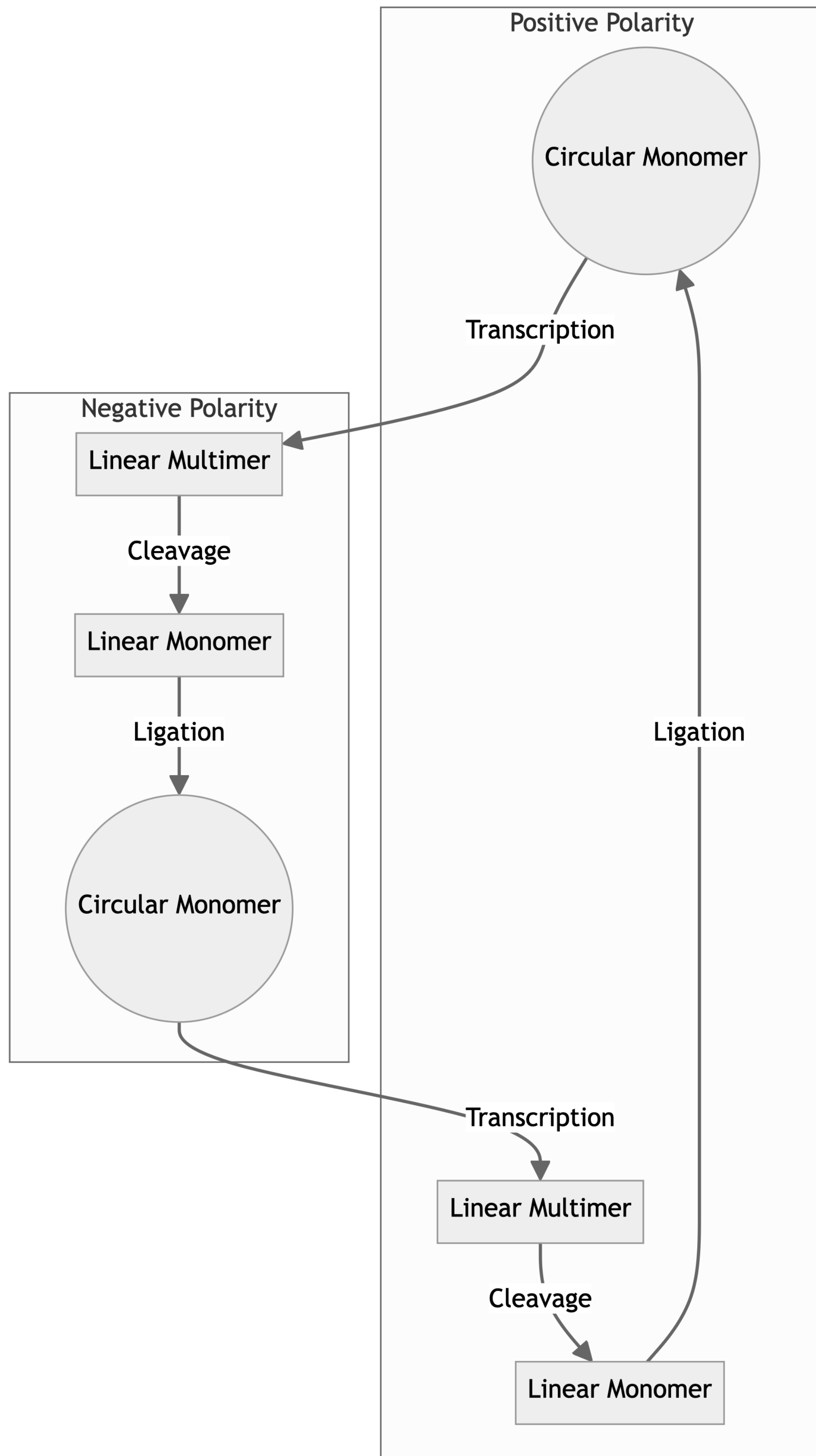


Figure 1.4: The symmetric rolling circle replication mechanism

specific polymerase used by viroids varies depending on the family. Pospiviroid replication occurs in the nucleus and is prevented by α-amanitin, a specific inhibitor of Pol II.[27] In contrast, avsunviroids replicate in chloroplasts,[28] where they are transcribed by the plastid nuclear-encoded polymerase (NEP). Chloroplasts have two polymerases, the plastid-encoded polymerase (PEP) and the phage-like NEP.[29] Evidence for the use of NEP by avsunviroids comes from the fact that their replication is not affected by tagetitoxin, a specific inhibitor of PEP.[30]

These multimeric intermediates are then transcribed themselves and cleaved to unit length, after which the original circular positive-polarity sequence is regenerated via ligation by a co-opted host ligase. The cleavage step varies between the two families of viroids. In members of the *Pospiviroidae* family, a host RNAse cleaves the intermediates, whereas in members of the family *Avsunviroidae*, cleavage of the intermediate is catalysed by an autocatalytic RNA enzyme (hereinafter referred to

---

Manipulate Their Hosts for Being Propagated and How Their Hosts React for Containing the Infection," *Virus Research* 209 (November 2015): 136–145, doi:10.1016/j.virusres.2015.02.027.

[27]Hans-P. Mühlbach and Heinz L. Sänger, "Viroid Replication Is Inhibited by α-Amanitin," *Nature* 278, no. 5700 (March 1979): 185–188, doi:10.1038/278185a0; I.-M. Schindler and H.-P. Mühlbach, "Involvement of Nuclear DNA-dependent RNA Polymerases in Potato Spindle Tuber Viroid Replication: A Reevaluation," *Plant Science* 84, no. 2 (January 1992): 221–229, doi:10.1016/0168-9452(92)90138-C.

[28]Roderick G. Bonfiglioli, Geoffrey I. McFadden, and Robert H. Symons, "In Situ Hybridization Localizes Avocado Sunblotch Viroid on Chloroplast Thylakoid Membranes and Coconut Cadang Cadang Viroid in the Nucleus," *The Plant Journal* 6, no. 1 (July 1994): 99–103, doi:10.1046/j.1365-313X.1994.6010099.x; M. I. Lima et al., "Detection of Avocado Sunblotch Viroid in Chloroplasts of Avocado Leaves by in Situ Hybridization," *Archives of Virology* 138, no. 3–4 (September 1994): 385–390, doi:10.1007/BF01379142; F. Bussière et al., "Subcellular Localization and Rolling Circle Replication of Peach Latent Mosaic Viroid: Hallmarks of Group A Viroids," *Journal of Virology* 73, no. 8 (August 1999): 6353–6360, doi:10.1128/JVI.73.8.6353-6360.1999.

[29]Boris Hedtke, Thomas Börner, and Andreas Weihe, "Mitochondrial and Chloroplast Phage-Type RNA Polymerases in Arabidopsis," *Science* 277, no. 5327 (August 1997): 809–811, doi:10.1126/science.277.5327.809; Luca Tadini et al., "The Plastid Transcription Machinery and Its Coordination with the Expression of Nuclear Genome: Plastid-Encoded Polymerase, Nuclear-Encoded Polymerase and the Genomes Uncoupled 1-Mediated Retrograde Communication," *Philosophical Transactions of the Royal Society B: Biological Sciences* 375, no. 1801 (May 2020): 20190399, doi:10.1098/rstb.2019.0399.

[30]José-Antonio Navarro, Antonio Vera, and Ricardo Flores, "A Chloroplastic RNA Polymerase Resistant to Tagetitoxin Is Involved in Replication of Avocado Sunblotch Viroid," *Virology* 268, no. 1 (March 2000): 218–225, doi:10.1006/viro.1999.0161; D E Mathews and R D Durbin, "Tagetitoxin Inhibits RNA Synthesis Directed by RNA Polymerases from Chloroplasts and Escherichia Coli." *Journal of Biological Chemistry* 265, no. 1 (January 1990): 493–498, doi:10.1016/S0021-9258(19)40258-5.

as a *ribozyme*[31]) formed by the viroid RNA itself.[32] Viroids of the *Avsunviroidae* engage in a symmetric process, in which intermediates are ligated into circular forms of the opposite polarity before undergoing the same RCR process again, to produce circular RNAs of the original polarity.[33] This process requires viroids to contain two ribozymes (one per RNA strand of each polarity). In members of the *Pospiviroidae* family, only one RCR cycle occurs to create the negative strand.[34] The linear, multimeric negative strand is transcribed, and the positive polarity transcript is cleaved to unit length by a host RNAse. The two families of viroids also vary in their choice of the ligase: members of *Pospiviroidae* use host DNA ligase 1 that they repurpose as RNA ligase,[35] whereas members of *Avsunviroidae*, which replicate within plastids, use chloroplast tRNA ligase.[36]

As a side effect of their replication mechanism, viroid "genomes" should be thought of as not just a single sequence but a pair of sequences of opposite polarity. These sequences may have different secondary structures[37] and intracellular trafficking.[38] The polarities are named (+) and (-) semi-arbitrarily, with the (+) polarity being

[31]It must be noted that there are other ribozymes, most notably the ribosome, but for the purpose of this thesis, the term *ribozyme* will exclusively refer to the autocatalytic type unless explicitly stated otherwise.

[32]Ying Wang, "Current View and Perspectives in Viroid Replication," *Current Opinion in Virology* 47 (April 2021): 32–37, doi:10.1016/j.coviro.2020.12.004; Flores et al., "Rolling-Circle Replication of Viroids, Viroid-Like Satellite RNAs and Hepatitis Delta Virus"; Flores, Gas, et al., "Viroid Replication."

[33]José-Antonio Daròs, Santiago F Elena, and Ricardo Flores, "Viroids: An Ariadne's Thread into the RNA Labyrinth," *EMBO Reports* 7, no. 6 (June 2006): 593–598, doi:10.1038/sj.embor.7400706; Diego Molina-Serrano et al., "Processing of RNAs of the Family Avsunviroidae in Chlamydomonas Reinhardtii Chloroplasts," *Journal of Virology* 81, no. 8 (April 2007): 4363–4366, doi:10.1128/JVI.02556-06.

[34]Branch, Benenfeld, and Robertson, "Evidence for a Single Rolling Circle in the Replication of Potato Spindle Tuber Viroid."

[35]María-Ángeles Nohales, Ricardo Flores, and José-Antonio Daròs, "Viroid RNA Redirects Host DNA Ligase 1 to Act as an RNA Ligase," *Proceedings of the National Academy of Sciences* 109, no. 34 (August 2012): 13805–13810, doi:10.1073/pnas.1206187109.

[36]María-Ángeles Nohales et al., "Involvement of the Chloroplastic Isoform of tRNA Ligase in the Replication of Viroids Belonging to the Family Avsunviroidae," *Journal of Virology* 86, no. 15 (August 2012): 8269–8276, doi:10.1128/JVI.00629-12.

[37]David N. Kuhn, Andrew D. W. Geering, and Jonathan Dixon, "Avocado Sunblotch Viroid," in *Viroids and Satellites* (Elsevier, 2017), 297–305, doi:10.1016/B978-0-12-801498-1.00028-0.

[38]Yijun Qi and Biao Ding, "Differential Subnuclear Localization of RNA Strands of Opposite Polarity Derived from an Autonomously Replicating Viroid," *The Plant Cell* 15, no. 11 (November 2003): 2566–2577, doi:10.1105/tpc.016576.

the polarity that accumulates most in the host.[39] A further side effect of their replication mechanism that must be noted is that viroids are able to form multimeric linear intermediates of either polarity, meaning that detection methods that rely exclusively on the presence of circular RNAs may miss some viroids. This problem is addressed in Chapter 3, which introduces a method that works on both circular and linear forms of the viroid genome.

Another side effect of their replication mechanism is that viroids have an extremely high mutation rate, estimated at 10^{-3} mutations per nucleotide in some avsunviroids.[40] This high mutation rate is due to the lack of proofreading by the host NEP, resulting in the highest mutation rate of any known biological agent. Indeed, it appears that avsunviroids skirt the edge of the error threshold, the point at which the mutation rate is so high that the population can no longer maintain its genetic information.[41] The nuclear-replicating pospiviroids have a lower mutation rate, estimated at 10^{-4} mutations per nucleotide,[42] a rate closer to that of RNA viruses.

Ribozymes are a key feature of viroids and will be a recurring theme in this thesis. In viroids that use ribozymes, the self-cleaving ribozyme of choice is the hammerhead ribozyme (HHR).[43] There are multiple known variants of the HHR, but the most common variant in viroids is the type III HHR. Within the HHR is a conserved core essential for catalysis, `CUGANGA` (where `N` is any nucleotide).[44] This

[39] J.-A. Navarro, "Characterization of the Initiation Sites of Both Polarity Strands of a Viroid RNA Reveals a Motif Conserved in Sequence and Structure," *The EMBO Journal* 19, no. 11 (June 2000): 2662–2670, doi:10.1093/emboj/19.11.2662; Beatriz Navarro, Luisa Rubino, and Francesco Di Serio, "Small Circular Satellite RNAs," in *Viroids and Satellites* (Elsevier, 2017), 659–669, doi:10.1016/B978-0-12-801498-1.00061-9.

[40] Selma Gago et al., "Extremely High Mutation Rate of a Hammerhead Viroid," *Science* 323, no. 5919 (March 2009): 1308–1308, doi:10.1126/science.1169202.

[41] Manfred Eigen, "Selforganization of Matter and the Evolution of Biological Macromolecules," *Die Naturwissenschaften* 58, no. 10 (October 1971): 465–523, doi:10.1007/BF00623322; discussed in depth in Eugene V. Koonin, *The Logic of Chance: The Nature and Origin of Biological Evolution* (Upper Saddle River, N.J: Pearson Education, 2012).

[42] Amparo López-Carrasco et al., "Different Rates of Spontaneous Mutation of Chloroplastic and Nuclear Viroids as Determined by High-Fidelity Ultra-Deep Sequencing," ed. Edward C. Holmes, *PLOS Pathogens* 13, no. 9 (September 2017): e1006547, doi:10.1371/journal.ppat.1006547.

[43] Flores, Minoia, et al., "Viroid Replication."

[44] Iris Eckert et al., "Discovery of Natural Non-Circular Permutations in Non-Coding RNAs," *Nucleic Acids Research* 51, no. 6 (April 2023): 2850–2861, doi:10.1093/nar/gkad137.

motif, as well as the Rfam models for HHRs,[45] will be used to identify circular RNAs as viroid-like in Chapter 3.

Regions of the viroid genome that are not catalytically active are nonetheless important for replication. In pospiviroids, multiple domains have been identified.[46] These domains include a pathogenicity domain, a central conserved region, and a terminal conserved region or hairpin (depending on the genus).[47] In addition, a variable domain is also present.[48] Recent research has shown that these domains may not be as clearly defined as previously thought[49] but the conservation of these non-enzymatic regions demonstrates that more than simple secondary structure is required for viroid replication. Mutational analyses of PSTVd has shown that disrupting nearly every loop in the secondary structure had adverse effects on replication or trafficking, with the variable domain most tolerant of mutation.[50] Similarly, Jian Wu et al.[51] demonstrated that mutations in seven of the the 17 G/U wobble pairs in the PSTVd genome inhibited replication or trafficking. One of these pairs in the region in which host Pol II binds was shown to be essential for replication. In sum, the secondary structure—and not just the ribozymes—of viroids is critical for their replication.

[45] Ioanna Kalvari et al., "Rfam 14: Expanded Coverage of Metagenomic, Viral and microRNA Families," *Nucleic Acids Research* 49, no. D1 (January 2021): D192–D200, doi:10.1093/nar/gkaa1047.

[46] P Keese and R H Symons, "Domains in Viroids: Evidence of Intermolecular RNA Rearrangements and Their Contribution to Viroid Evolution." *Proceedings of the National Academy of Sciences of the United States of America* 82, no. 14 (July 1985): 4582–4586, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC390429/.

[47] Kevin-Phil Wüsthoff and Gerhard Steger, "Conserved Motifs and Domains in Members of Pospiviroidae," *Cells* 11, no. 2 (January 2022): 230, doi:10.3390/cells11020230.

[48] Y. Hu et al., "Role of the Variable Domain in Modulating Potato Spindle Tuber Viroid Replication," *Virology* 219, no. 1 (May 1996): 45–56, doi:10.1006/viro.1996.0221.

[49] Wüsthoff and Steger, "Conserved Motifs and Domains in Members of Pospiviroidae."

[50] Xuehua Zhong et al., "A Genomic Map of Viroid RNA Motifs Critical for Replication and Systemic Trafficking," *The Plant Cell* 20, no. 1 (2008): 35–47, https://www.jstor.org/stable/25224211.

[51] "Functional Analysis Reveals G/U Pairs Critical for Replication and Trafficking of an Infectious Non-Coding Viroid RNA," *Nucleic Acids Research* 48, no. 6 (February 2020): 3134–3155, doi:10.1093/nar/gkaa100.

1.1.4 Prevention, transmission, and treatment

The prevention of viroid-induced diseases focuses primarily on the use of viroid-free planting material and the control of vectors. Prevention is particularly important given that effective methods for elimination of viroids from infected plants are challenging[52] and there are no effective methods for eliminating viroids from seeds.[53] Methods to ensure viroid-free planting material include the use of quarantines, certification programs, and the use of molecular methods to detect viroids in plants.[54]

As viroid transmission is primarily the result of mechanical transmission, disinfection of tools is another important method of prevention.[55] Transmission by vectors such as insects, hemiparasitic plants (*e.g.*, mistletoe),[56] and goats[57] has also been reported,[58] suggesting that the control of insect vectors may also be relevant in some cases. Viroid transmission may also occur via the same mechanisms as viral transmission due to transencapsidation, a process in which viroids are encapsidated by viruses and transmitted by them.[59] Such a process has been observed in the case of PSTVd, which was encapsidated with potato leafroll virus and transmitted by aphids.[60] As such, other methods of preventing disease spread include management

[52] Marina Barba et al., "Viroid Elimination by Thermotherapy, Cold Therapy, Tissue Culture, In Vitro Micrografting, or Cryotherapy," in *Viroids and Satellites* (Elsevier, 2017), 425–435, doi:10.1016/B978-0-12-801498-1.00040-1.

[53] Kai-Shu Ling, "Decontamination Measures to Prevent Mechanical Transmission of Viroids," in *Viroids and Satellites* (Elsevier, 2017), 437–445, doi:10.1016/B978-0-12-801498-1.00041-3.

[54] Marina Barba and Delano James, "Quarantine and Certification for Viroids and Viroid Diseases," in *Viroids and Satellites* (Elsevier, 2017), 415–424, doi:10.1016/B978-0-12-801498-1.00039-5.

[55] Ling, "Decontamination Measures to Prevent Mechanical Transmission of Viroids."

[56] Thomas Leichtfried et al., "Transmission Studies of the Newly Described Apple Chlorotic Fruit Spot Viroid Using a Combined RT-qPCR and Droplet Digital PCR Approach," *Archives of Virology* 165, no. 11 (2020): 2665–2671, doi:10.1007/s00705-020-04704-5.

[57] Oded Cohen et al., "Goat Horns: Platforms for Viroid Transmission to Fruit Trees?" *Phytoparasitica* 33, no. 2 (April 2005): 141–148, doi:10.1007/BF03029972.

[58] Reviewed in Rosemarie W. Hammond, "Seed, Pollen, and Insect Transmission of Viroids," in *Viroids and Satellites* (Elsevier, 2017), 521–530, doi:10.1016/B978-0-12-801498-1.00048-6; Ahmed Hadidi, Liying Sun, and John W. Randles, "Modes of Viroid Transmission," *Cells* 11, no. 4 (February 2022): 719, doi:10.3390/cells11040719.

[59] Jerzy Syller, Waldemar Marczewski, and Jerzy Pawłowicz, "Transmission by Aphids of Potato Spindle Tuber Viroid Encapsidated by Potato Leafroll Luteovirus Particles," *European Journal of Plant Pathology* 103, no. 3 (1997): 285–289, doi:10.1023/A:1008648822190.

[60] Syller, Marczewski, and Pawłowicz.

practices such as limiting worker movement between areas and disposal of infected plants, while not specific to viroids, are also relevant.

In the future, transgenic plants tolerant to viroids may be developed.[61] Recent evidence has shown that PSTVd tolerance present in wild tomato species can be introduced into cultivated tomatoes as a dominant trait by crossing.[62] This suggests that breeding for viroid tolerance may be a viable strategy for some crops, although the mechanisms of tolerance are not yet understood.

1.1.5 Distribution and host range

Viroid of both families are found on every continent except Antarctica.[63] Given the extent of global trade in plants, this widespread distribution is not surprising. For example, an outbreak of PSTVd infections within Cape Gooseberries in New Zealand was traced to seeds imported from Germany.[64] Even remote islands are not immune to viroid infection, as evidenced by the widespread presence of coconut tinangaja viroid in Guam.[65]

With respect to host range, viroids are found in a wide variety of plants. Some viroids exhibit a narrow host range both in the wild and in the lab, such as

[61] Rosemarie W. Hammond and Natalia Kovalskaya, "Strategies to Introduce Resistance to Viroids," in *Viroids and Satellites* (Elsevier, 2017), 447–455, doi:10.1016/B978-0-12-801498-1.00042-5.

[62] Takashi Naoi and Tatsuji Hataya, "Tolerance Even to Lethal Strain of Potato Spindle Tuber Viroid Found in Wild Tomato Species Can Be Introduced by Crossing," *Plants* 10, no. 3 (March 2021): 575, doi:10.3390/plants10030575.

[63] Edward V. Podleckis, "Geographical Distribution of Viroids in the Americas," in *Viroids and Satellites* (Elsevier, 2017), 459–472, doi:10.1016/B978-0-12-801498-1.00043-7; Francesco Faggioli et al., "Geographical Distribution of Viroids in Europe," in *Viroids and Satellites* (Elsevier, 2017), 473–484, doi:10.1016/B978-0-12-801498-1.00044-9; Khaled A. El-Dougdoug et al., "Geographical Distribution of Viroids in Africa and the Middle East," in *Viroids and Satellites* (Elsevier, 2017), 485–496, doi:10.1016/B978-0-12-801498-1.00045-0; Andrew D. W. Geering, "Geographical Distribution of Viroids in Oceania," in *Viroids and Satellites* (Elsevier, 2017), 497–506, doi:10.1016/B978-0-12-801498-1.00046-2; Dattaraj B. Parakh, Shuifang Zhu, and Teruo Sano, "Geographical Distribution of Viroids in South, Southeast, and East Asia," in *Viroids and Satellites* (Elsevier, 2017), 507–518, doi:10.1016/B978-0-12-801498-1.00047-4.

[64] L. I. Ward et al., "First Report of Potato Spindle Tuber Viroid in Cape Gooseberry (Physalis Peruviana) in New Zealand," *Plant Disease* 94, no. 4 (April 2010): 479, doi:10.1094/PDIS-94-4-0479A.

[65] R. A. J. Hodgson, G. C. Wall, and J. W. Randles, "Specific Identification of Coconut Tinangaja Viroid for Differential Field Diagnosis of Viroids in Coconut Palm," *Phytopathology®* 88, no. 8 (August 1998): 774–781, doi:10.1094/PHYTO.1998.88.8.774.

chrysanthemum chlorotic mottle viroid (CChMVd), which has only been reported to infect chrysanthemums.[66] Yet others, such as hop stunt viroid (HSVd), have a wide natural host range, infecting plants from multiple orders.[67] While plants are indubitably the primary hosts of viroids, there is evidence that viroids can replicate in other organisms.

Early evidence suggested that avsunviroids could be processed in algae[68] and replicated in yeast.[69] Delan-Forino, Maurel, and Torchet[70] demonstrated that ASBVd was able to self-cleave and replicate in *Saccharomyces cerevisiae*, although no evidence of pathogenicity was observed, nor were claims of infectivity in yeast made. Recently, Shuang Wei et al.[71] reported symptomatic infection of the phytopathogenic fungi *Fusarium graminearum*, *Cryphonectria parasitica*, and *Valsa mali* with HSVd. In addition, they demonstrated that iresine viroid 1 (IrVd-1) and ASBVd were able to stably replicate in *C. parasitica*. Furthermore, they claimed that HSVd was able to be transmitted bidirectionally between *F. graminearum* and plants and that HSVd infection led to decreased virulence in one fungus, *V. mali*. These findings were immediately challenged,[72] citing various methodological issues. The authors of the original study responded to these criticisms[73] with evidence of nucleotide substitutions consistent with viroid replication. The debate remains

[66] Ricardo Flores et al., "Chrysanthemum Chlorotic Mottle Viroid," in *Viroids and Satellites* (Elsevier, 2017), 331–338, doi:10.1016/B978-0-12-801498-1.00031-0.

[67] Tatsuji Hataya, Taro Tsushima, and Teruo Sano, "Hop Stunt Viroid," in *Viroids and Satellites* (Elsevier, 2017), 199–210, doi:10.1016/B978-0-12-801498-1.00019-X.

[68] Molina-Serrano et al., "Processing of RNAs of the Family Avsunviroidae in Chlamydomonas Reinhardtii Chloroplasts."

[69] C. Delan-Forino, M.-C. Maurel, and C. Torchet, "Replication of Avocado Sunblotch Viroid in the Yeast Saccharomyces Cerevisiae," *Journal of Virology* 85, no. 7 (April 2011): 3229–3238, doi:10.1128/JVI.01320-10.

[70]

[71] "Symptomatic Plant Viroid Infections in Phytopathogenic Fungi," *Proceedings of the National Academy of Sciences* 116, no. 26 (June 2019): 13042–13050, doi:10.1073/pnas.1900762116.

[72] Pedro Serra et al., "Symptomatic Plant Viroid Infections in Phytopathogenic Fungi: A Request for a Critical Reassessment," *Proceedings of the National Academy of Sciences* 117, no. 19 (May 2020): 10126–10128, doi:10.1073/pnas.1922249117.

[73] Shuang Wei et al., "Reply to Serra Et Al.: Nucleotide Substitutions in Plant Viroid Genomes That Multiply in Phytopathogenic Fungi," *Proceedings of the National Academy of Sciences* 117, no. 19 (May 2020): 10129–10130, doi:10.1073/pnas.2001670117.

ongoing[74] but evidence presented in Section 4.3.4 does support the expanded host range of viroids to include phytopathogenic fungi.

1.1.6 Economic impact

The economic impact of viroids is substantial.[75] These impacts are split between direct costs (*i.e.*, damage to crops) and indirect costs (*i.e.*, control measures). No literature is forthcoming regarding the total economic impacts of this entire class of agents, but the impacts of individual viroids have been variously estimated. Together, these estimates provide a sense of the total economic impact of viroids.

A recent study has estimated the economic impact of hop stunt viroid at between $432 and $26,795 per acre of hops, depending on the level of yield reduction.[76] The top end of that range assumed a 62% reduction in yield, which has been observed in some cultivars.[77] A similar reduction in yield has been reported in some cases for PSTVd,[78] which has had its total economic impact in the European Union estimated at €4.4 million in potatoes and €5.7 million in tomatoes for 2012.[79] In contrast, PSTVd has been all but eliminated within potatoes of the United States,[80]

[74]Summarized in Liying Sun and Ahmed Hadidi, "Mycoviroids: Fungi as Hosts and Vectors of Viroids," *Cells* 11, no. 8 (April 2022): 1335, doi:10.3390/cells11081335.

[75]Reviewed in M. Judith B. Rodriguez, Ganesan Vadamalai, and John W. Randles, "Economic Significance of Palm Tree Viroids," in *Viroids and Satellites* (Elsevier, 2017), 39–49, doi:10.1016/B978-0-12-801498-1.00004-8; Jacobus Th.J. Verhoeven, Rosemarie W. Hammond, and Giuseppe Stancanelli, "Economic Significance of Viroids in Ornamental Crops," in *Viroids and Satellites* (Elsevier, 2017), 27–38, doi:10.1016/B978-0-12-801498-1.00003-6; Ahmed Hadidi, Georgios Vidalakis, and Teruo Sano, "Economic Significance of Fruit Tree and Grapevine Viroids," in *Viroids and Satellites* (Elsevier, 2017), 15–25, doi:10.1016/B978-0-12-801498-1.00002-4; Rosemarie W. Hammond, "Economic Significance of Viroids in Vegetable and Field Crops," in *Viroids and Satellites* (Elsevier, 2017), 5–13, doi:10.1016/B978-0-12-801498-1.00001-2.

[76]Trent J. Davis et al., "The Economic Impact of Hop Stunt Viroid and Certified Clean Planting Materials," *HortScience* 56, no. 12 (December 2021): 1471–1475, doi:10.21273/HORTSCI15975-21.

[77]Madhu Kappagantu et al., "Hop Stunt Viroid: Effects on Vegetative Growth and Yield of Hop Cultivars, and Its Distribution in Central Washington State," *Plant Disease* 101, no. 4 (April 2017): 607–612, doi:10.1094/PDIS-06-16-0884-RE.

[78]T. O. Diener, ed., *The Viroids* (Boston, MA: Springer US, 1987), doi:10.1007/978-1-4613-1855-2.

[79]T. Soliman et al., "Quantitative Economic Impact Assessment of an Invasive Plant Disease Under Uncertainty – A Case Study for Potato Spindle Tuber Viroid (PSTVd) Invasion into the European Union," *Crop Protection* 40 (October 2012): 28–35, doi:10.1016/j.cropro.2012.04.019.

[80]Mike Sun et al., "Survey of Potato Spindle Tuber Viroid in Seed Potato Growing Areas of the United States," *American Journal of Potato Research* 81, no. 3 (May 2004): 227–231, doi:10.1007/BF02871753.

although scattered incidents of tomato infection resulting in crop losses have been reported on both coasts.[81] Even apparently latent infections may reduce yield, as has been observed in the case of ASBVd in avocado trees which had 95% yield reduction.[82]

The host range of viroids is actively expanding as agriculture expands, resulting in greater economic impacts. One prime example is the recent discovery of hop latent viroid (HLVd) in cannabis plants in the United States.[83] This discovery corresponded to only the fourth reported host of HLVd. Again, despite the appellation "latent", HLVd caused changes in the terpene profile of the cannabis plants as well as a variety of symptoms such as stunting, chlorosis, and malformation. Given the high economic value of cannabis and the recent advent of large-scale legal cultivation in the United States, the discovery of HLVd in cannabis plants is likely to have a substantial economic impact into the future.

1.2 Satellite RNAs

Broadly similar to viroids in terms of structure, replication, and range, small circular satellite RNAs (satRNAs) differ in that they are encapsidated and use the helper virus's polymerase rather than the host's polymerase. This encapsidation is not performed by any proteins that they encode within their genomes but rather by means of a helper virus.[84] These viral satellites replicate within plants, just like

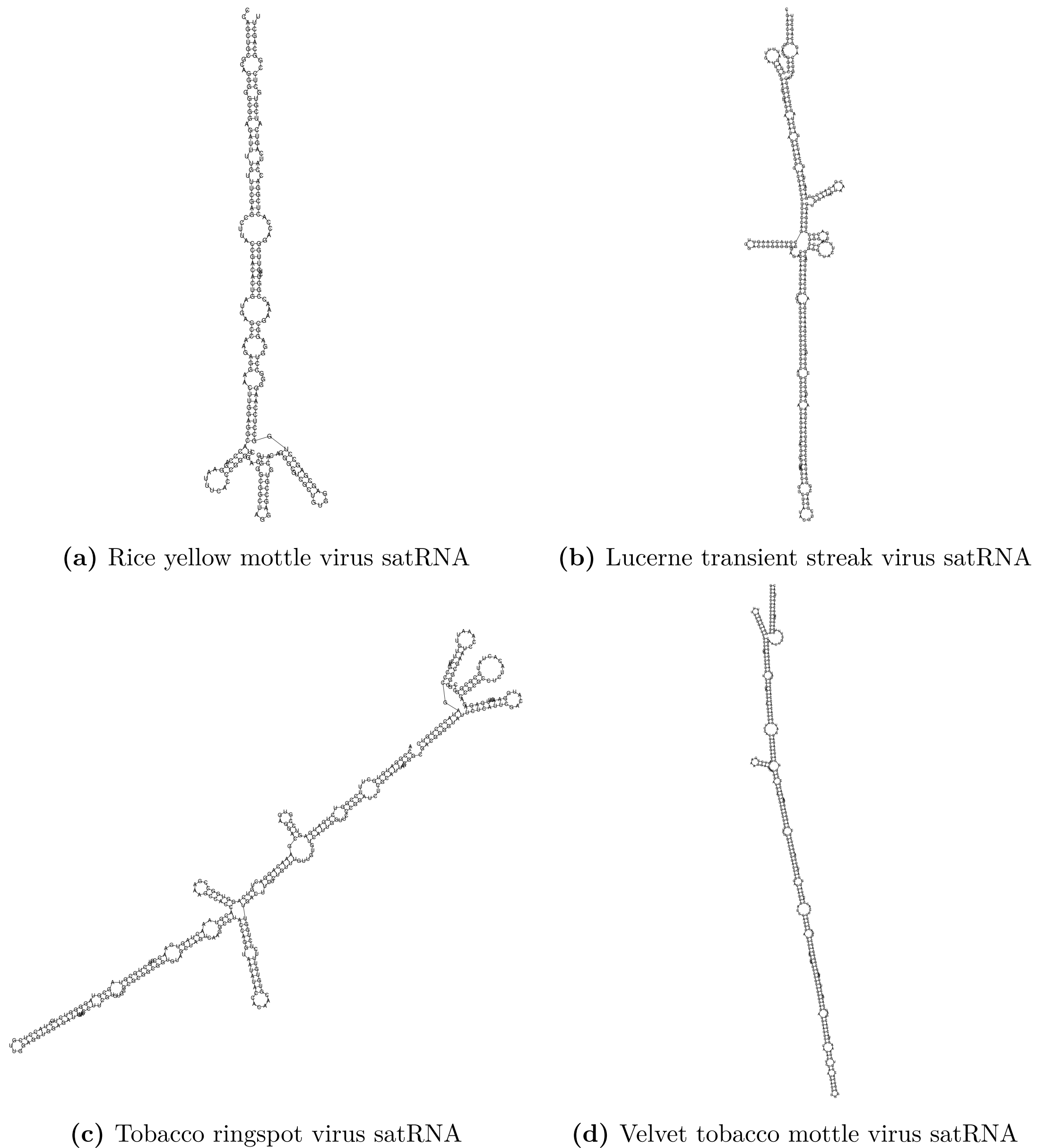
[81] K.-S. Ling and D. Sfetcu, "First Report of Natural Infection of Greenhouse Tomatoes by Potato Spindle Tuber Viroid in the United States," *Plant Disease* 94, no. 11 (November 2010): 1376, doi:10.1094/PDIS-07-10-0516; K.-S. Ling et al., "First Report of Potato Spindle Tuber Viroid Naturally Infecting Greenhouse Tomatoes in North Carolina," *Plant Disease* 97, no. 1 (January 2013): 148, doi:10.1094/PDIS-07-12-0679-PDN.

[82] Diener, *The Viroids*.

[83] J. G. Warren, J. Mercado, and D. Grace, "Occurrence of Hop Latent Viroid Causing Disease in Cannabis Sativa in California," *Plant Disease* 103, no. 10 (October 2019): 2699–2699, doi:10.1094/PDIS-03-19-0530-PDN; A. Bektaş et al., "Occurrence of Hop Latent Viroid in Cannabis Sativa with Symptoms of Cannabis Stunting Disease in California," *Plant Disease* 103, no. 10 (October 2019): 2699, doi:10.1094/PDIS-03-19-0459-PDN; Charith Raj Adkar-Purushothama, Teruo Sano, and Jean-Pierre Perreault, "Hop Latent Viroid: A Hidden Threat to the Cannabis Industry," *Viruses* 15, no. 3 (March 2023): 681, doi:10.3390/v15030681.

[84] Uzma Badar et al., "Molecular Interactions of Plant Viral Satellites," *Virus Genes* 57, no. 1 (February 2021): 1–22, doi:10.1007/s11262-020-01806-9; Prabu Gnanasekaran and Supriya Chakraborty, "Biology of Viral Satellites and Their Role in Pathogenesis," *Current Opinion in Virology* 33 (December 2018): 96–105, doi:10.1016/j.coviro.2018.08.002.

viroids, and are also small (about 300 nt) and circular.



(a) Rice yellow mottle virus satRNA **(b)** Lucerne transient streak virus satRNA

(c) Tobacco ringspot virus satRNA **(d)** Velvet tobacco mottle virus satRNA

Figure 1.5: Representative secondary structures of satRNAs, produced using RNAfold[85]

However, unlike viroids, satRNAs do not rely on host DdRP for replication, but instead employ the RNA-dependent RNA polymerase (RdRP) of the helper virus.[86]

[85] Lorenz et al., "ViennaRNA Package 2.0."

[86] Flores et al., "Rolling-Circle Replication of Viroids, Viroid-Like Satellite RNAs and Hepatitis Delta Virus"; Navarro, Rubino, and Di Serio, "Small Circular Satellite RNAs"; A. L. N. Rao and Kriton Kalantidis, "Virus-Associated Small Satellite RNAs and Viroids Display Similarities in Their Replication Strategies," *Virology* 479–480 (May 2015): 627–636, doi:10.1016/j.virol.2015.02.018; Ying-Wen Huang et al., "Replication of Satellites," in *Viroids and Satellites* (Elsevier, 2017), 577–586, doi:10.1016/B978-0-12-801498-1.00053-X.

Additionally, of note, satRNAs vary in their replication mechanisms, beyond using a different polymerase, in that they do not always contain one ribozyme per strand polarity and, when they do, each strand can contain a distinct type of ribozyme.[87] Unlike viroids, satRNAs are known to encode both HHRs and hairpin ribozymes (HPR), representing a different autocatalytic motif.[88] Given that viroids and satRNAs share so many features and are present in the same types of hosts, these agents are thought to share a common ancestor. Phylogenetic reconstructions have shown evidence of such an evolutionary relationship although the exact nature of the common ancestor remains unclear.[89]

1.2.1 Large satRNAs and small linear satRNAs

It must be noted that the term *satellite RNA* can refer to other agents that are not viroid-like. These agents, while they do have RNA genomes and are satellites, are here excluded from the term *satRNA* for the purposes of this thesis unless explicitly stated otherwise. Such agents include large RNA satellites, which exceed 800 nt in length and can encode nonstructural proteins,[90] and small linear satellites, which do not use a rolling circle replication mechanism.[91] Some small satRNAs, such as the satellite of tobacco ringspot virus (satTRSV) depicted in Figure 1.5c, are encapsidated as linear intermediates but still replicate via the rolling circle mechanism.[92] These are included in the term *satRNA* for the purposes of this thesis due to their fundamental similarity to viroids and other viroid-like agents.

[87]Navarro, Rubino, and Di Serio, "Small Circular Satellite RNAs."

[88]Adrian R. Ferré-D'Amaré and William G. Scott, "Small Self-Cleaving Ribozymes," *Cold Spring Harbor Perspectives in Biology* 2, no. 10 (October 2010): a003574, doi:10.1101/cshperspect.a003574.

[89]Santiago F. Elena et al., "Phylogenetic Analysis of Viroid and Viroid-Like Satellite RNAs from Plants: A Reassessment," *Journal of Molecular Evolution* 53, no. 2 (August 2001): 155–159, doi:10.1007/s002390010203.

[90]Mazen Alazem and Na-Sheng Lin, "Large Satellite RNAs," in *Viroids and Satellites* (Elsevier, 2017), 639–648, doi:10.1016/B978-0-12-801498-1.00059-0.

[91]Mikyeong Kim and Marilyn J. Roossinck, "Small Linear Satellite RNAs," in *Viroids and Satellites* (Elsevier, 2017), 649–658, doi:10.1016/B978-0-12-801498-1.00060-7.

[92]Navarro, Rubino, and Di Serio, "Small Circular Satellite RNAs."

1.2.2 Pathogenicity and RNA silencing

Just as with viroids, satRNAs modulate RNA expression by means of RNA silencing.[93] However, satRNAs have two other biological entities with which they must interact: the helper virus and the host. There are examples of satRNAs that suppress the symptoms of the helper virus and others that enhance them.[94] An example of the former is the smaller satRNA of chicory yellow mottle virus (satCYMV)[95], which causes symptomless infection.[96] An example of the latter is the smaller satRNA of the arabis mosaic virus, which worsens the symptoms of infected *Chenopodium quinoa* plants.[97] Evidence of RNA silencing in circular satRNAs is scarce. It is likely that the interaction between satRNAs and the helper virus is mediated by RNA silencing, as is the case with viroids.

While not a viroid-like circular satRNA, the remarkable case of the satellite of cucumber mosaic virus (satCMV) is worth mentioning as a demonstration of RNA silencing. This satRNA, which is small and linear, is able to control the host plant, the helper virus, and the insect vector (aphids) to ensure its own transmission.[98] It does so using RNA silencing, which it uses to suppress the symptoms of the helper virus by markedly reducing the levels of the helper virus's RNA. This is a form of hyperparasitism, in which the satRNA parasitizes the helper virus, which in turn parasitizes the host plant. A naive assumption would be that reducing the levels of the helper virus should reduce the transmission of satCMV. This is true, but not by as much as one might expect, as the CMV levels in the vector insect were

[93] Navarro, Rubino, and Di Serio.

[94] Navarro, Rubino, and Di Serio.

[95] There are in fact multiple satellite RNAs of this virus and the arabis mosaic virus as well. The ones that are referred to here are of course the circular one, rather than the larger linear ones.

[96] P. Piazzolla, C. Vovlas, and L. Rubino, "Symptom Regulation Induced by Chicory Yellow Mottle Virus Satellite-Like RNA," *Journal of Phytopathology* 115, no. 2 (February 1986): 124–129, doi:10.1111/j.1439-0434.1986.tb00868.x.

[97] D. L. Davies and M. F. Clark, "A Satellite-Like Nucleic Acid of Arabis Mosaic Virus Associated with Hop Nettlehead Disease," *Annals of Applied Biology* 103, no. 3 (December 1983): 439–448, doi:10.1111/j.1744-7348.1983.tb02781.x.

[98] Wikum H. Jayasinghe et al., "A Plant Virus Satellite RNA Directly Accelerates Wing Formation in Its Insect Vector for Spread," *Nature Communications* 12, no. 1 (December 2021): 7087, doi:10.1038/s41467-021-27330-4.

not reduced as much as the levels in the host plant. Moreover, satCMV reduces the symptoms of the helper virus in the plant host while also turning the leaves of the host plant yellow, which attracts the vector insect. It does so using siRNAs derived from satCMV and without seriously impacting the photosynthetic capacity of the plant. The yellowing of the leaves makes them more attractive to aphids, which are then manipulated by satCMV. When consumed by immature nymphs, satCMV induces wing formation indirectly by forming PIWI-interacting RNAs (piRNAs) that target the miRNAs of the aphid, which in turn target the genes that control wing formation. By competitive inhibition of the aphid's miRNAs, satCMV upregulates wing formation in the aphid, promoting its own transmission.[99] While no such complex interactions have been observed in viroids or viroid-like satRNAs, it is not unreasonable to assume that such interactions could exist, given that the size of satCMV is the same as that of viroids and viroid-like satRNAs.

1.2.3 Replication

The replication of satRNAs is assisted by the helper virus, which provides the RdRP that is used to replicate the satRNA. Evidence for the use of RdRP rather than host DdRP comes from the fact that satRNA replication is not inhibited by α-amanitin,[100] a specific inhibitor of DdRP that does inhibit the replication of viroids.[101] As with viroids, satRNAs may use either the symmetric or asymmetric RCR mechanism. In the case of the asymmetric mechanism, only one polarity is cleaved by an autocatalytic ribozyme, a feature unique to satRNAs among viroid-like agents. A different ribozyme, the HPR, is used by some satRNAs to cleave multimeric intermediates.[102] In other cases, the satRNA uses a more traditional HHR for cleavage.[103] Indeed, the satRNAs associated with viruses of the *Nepovirus* genus have both HHRs and HPRs in opposite polarities.

[99] Jayasinghe et al.
[100] Jian-Guo Wu, Wen-Jun Lu, and Po Tien, "Multiplication of Velvet Tobacco Mottle Virus in Nicotiana Clevelandii Protoplasts Is Resistant to α-Amanitin," *Journal of General Virology* 67, no. 12 (December 1986): 2757–2762, doi:10.1099/0022-1317-67-12-2757.
[101] Mühlbach and Sänger, "Viroid Replication Is Inhibited by α-Amanitin."
[102] Ferré-D'Amaré and Scott, "Small Self-Cleaving Ribozymes."
[103] Navarro, Rubino, and Di Serio, "Small Circular Satellite RNAs."

With respect to the ligase, less is known about satRNAs than viroids. The HPR is capable of performing ligation as well as cleavage,[104] particularly when frozen (but interestingly, not in supercooled water) or dehydrated.[105] It is therefore possible that the same ribozyme that cleaves the multimeric intermediates also ligates the unit length RNAs into circular forms. Evidence for this activity comes from the fact that the mismatched ribozymes of satTRSV yield different levels of ligation.[106] The HPR displays a preference towards ligation while the HHR prefers cleavage.[107] The weak activity of the HHR with respect to ligation implies that host ligase is required for stages of RCR where the monomers contain HHRs. Whether host ligase is also used for ligation of HPR-containing polarities (perhaps in conjunction with the HPR) remains to be determined.

1.2.4 Taxonomy

While small circular satRNAs are under the purview of the International Committee on Taxonomy of Viruses (ICTV), they are not classified in the same way as viroids.[108] Currently, satellites are in general classified into standard virus taxonomy ranks, with suffixes noting their status as satellites. However, no small circular satRNA has been classified, and their evolutionary relationships are not well understood. Indeed, whether they are monophyletic is an open question (see Section 1.7 for a discussion of viroid-like RNA evolution). There is a conserved `GAUUUU` motif within all small circular satRNAs of members of the *Sombemovirus*

---

[104] M. J. Fedor, "Tertiary Structure Stabilization Promotes Hairpin Ribozyme Ligation," *Biochemistry* 38, no. 34 (August 1999): 11040–11050, doi:10.1021/bi991069q; Martha J Fedor, "Structure and Function of the Hairpin Ribozyme," *Journal of Molecular Biology* 297, no. 2 (March 2000): 269–291, doi:10.1006/jmbi.2000.3560; Michelle K Nahas et al., "Observation of Internal Cleavage and Ligation Reactions of a Ribozyme," *Nature Structural & Molecular Biology* 11, no. 11 (November 2004): 1107–1113, doi:10.1038/nsmb842.

[105] S. A. Kazakov, "Ligation of the Hairpin Ribozyme in Cis Induced by Freezing and Dehydration," *RNA* 12, no. 3 (March 2006): 446–456, doi:10.1261/rna.2123506.

[106] See Navarro, Rubino, and Di Serio, "Small Circular Satellite RNAs" for a review.

[107] Fedor, "Structure and Function of the Hairpin Ribozyme."

[108] Peter Palukaitis, "Satellite Taxonomy," in *Viroids and Satellites* (Elsevier, 2017), 615–622, doi:10.1016/B978-0-12-801498-1.00057-7.

genus,[109] although the motif does not appear to be present in all small circular satRNAs.

1.2.5 Potential protein-coding capacity

One satRNA of particular note is the 220 nt satRNA of rice yellow mottle virus (satRYMV), which does encode a protein.[110] This encoded protein is unique in that it is formed by a self-overlapping ORF. The satRYMV genome is not divisible by three, meaning that a polymerase creating an mRNA would wrap around three times before returning to the same reading frame. Each nucleotide is therefore a part of three different codons as the first, second, and third nucleotide position. As a result, the length of the amino acid sequence is therefore approximately equal to the length of the RNA sequence. The encoded reading frame does contain a stop codon after two complete wraps around the circular RNA, but experimental evidence suggests the stop codon is leaky, resulting in a longer read-through protein. In addition, products suggesting five-round translation were detected *in vivo*. The conservation of the reading frame across all 37 known isolates of satRYMV suggests that the protein is essential. Even more surprising is the fact that contained within this sequence are two hammerhead ribozymes, one in each polarity, that are capable of cleaving the RNA into unit length.[111] This remarkably compact "nanogenome" has not been observed in any other satRNA or viroid and remains a unique feature of satRYMV. Methods for identifying such ORFs will be discussed in Section 5.3.3.

[109] Paul Keese, George Bruening, and Robert H. Symons, "Comparative Sequence and Structure of Circular RNAs from Two Isolates of Lucerne Transient Streak Virus," *FEBS Letters* 159, no. 1–2 (August 1983): 185–190, doi:10.1016/0014-5793(83)80443-8; Navarro, Rubino, and Di Serio, "Small Circular Satellite RNAs."

[110] Mounir Georges AbouHaidar et al., "Novel Coding, Translation, and Gene Expression of a Replicating Covalently Closed Circular RNA of 220 Nt," *Proceedings of the National Academy of Sciences* 111, no. 40 (October 2014): 14542–14547, doi:10.1073/pnas.1402814111.

[111] AbouHaidar et al.

1.2.6 Economic impact

Compared to viroids, little is known about the impact of small circular satRNAs on agriculture.[112] While the impact of other types of satRNAs such as satCMV is better understood, the impact of small circular satRNAs is hard to gauge. Given that they generally appear to reduce the symptoms of the helper virus,[113] it seems plausible to assume that they could have a positive net economic impact, but this remains to be demonstrated. Indeed, satCMV has been proposed as a biocontrol agent for cucumber mosaic virus.[114] Should the same be true for small circular satRNAs, they could also prove a useful tool for controlling plant viruses.

1.3 Ribozyviruses

Neither is the range of viroid-like RNA agents limited to plants nor are they all strictly non-coding. Members of the realm *Ribozyviria*, of which the only well-characterized one is human hepatitis delta virus (HDV), infect animals and encode a single protein that undergoes post-translational modification, producing two distinct forms that perform various functions in virus reproduction including the role of nucleocapsid inside the virions.[115] Apart from this distinction, ribozyviruses exhibit the key features of viroids and satRNAs. They too replicate via the rolling circle mechanism catalysed by host DdRP and employ a virus-encoded ribozyme to cleave the replication intermediates.[116] Until recently, HDV was the sole member

[112] Tiziana Mascia and Donato Gallitelli, "Economic Significance of Satellites," in *Viroids and Satellites* (Elsevier, 2017), 555–563, doi:10.1016/B978-0-12-801498-1.00051-6.

[113] Mascia and Gallitelli; Chikara Masuta and Hanako Shimura, "Satellite RNAs: Their Involvement in Pathogenesis and RNA Silencing," in *Viroids and Satellites* (Elsevier, 2017), 587–596, doi:10.1016/B978-0-12-801498-1.00054-1; Chung-Chi Hu, Yau-Heiu Hsu, and Na-Sheng Lin, "Satellite RNAs and Satellite Viruses of Plants," *Viruses* 1, no. 3 (December 2009): 1325–1350, doi:10.3390/v1031325.

[114] Mascia and Gallitelli, "Economic Significance of Satellites."

[115] W. H. Huang et al., "Post-Translational Modification of Delta Antigen of Hepatitis D Virus," *Current Topics in Microbiology and Immunology* 307 (2006): 91–112, doi:10.1007/3-540-29802-9_5; Camille Sureau and Francesco Negro, "The Hepatitis Delta Virus: Replication and Pathogenesis," *Journal of Hepatology* 64, no. 1 (April 2016): S102–S116, doi:10.1016/j.jhep.2016.02.013.

[116] Lucy E. Modahl et al., "RNA-Dependent Replication and Transcription of Hepatitis Delta Virus RNA Involve Distinct Cellular RNA Polymerases," *Molecular and Cellular Biology* 20, no. 16 (August 2000): 6030–6039, doi:10.1128/MCB.20.16.6030-6039.2000; T. B. Macnaughton and M. M. C. Lai, "HDV RNA Replication: Ancient Relic or Primer?" *Current Topics in Microbiology and Immunology* 307 (2006): 25–45, doi:10.1007/3-540-29802-9_2; Chung-Hsin Tseng and Michael

of the *Deltavirus* genus, but the discovery of distantly related viruses in a variety of vertebrates and invertebrates, including rodents,[117] snakes,[118] termites,[119] fish,[120] birds,[121] and bats,[122] suggests a considerable uncharacterized diversity among the ribozyviruses. Similarly to satRNAs, ribozyviruses use different types of ribozymes, one of which is unique to members of *Ribozyviria*.[123]

1.3.1 Human hepatitis delta virus

Given that HDV was the first ribozyvirus to be discovered and that it causes symptomatic infection in humans, it is the best characterized of the ribozyviruses. Identified by M. Rizzetto et al.[124] in 1977, HDV is a satellite virus of hepatitis B virus (HBV), generally requiring the presence of HBV for its packaging and transmission.[125]

HBV is itself a member of an entirely different realm, *Riboviria*, and is a DNA virus that replicates via an RNA intermediate.[126] As such, it is a member of the *Revtraviricetes* class, which also includes other reverse-transcribing viruses

M. Lai, "Hepatitis Delta Virus RNA Replication," *Viruses* 1, no. 3 (November 2009): 818–831, doi:10.3390/v1030818.

[117] Sofia Paraskevopoulou et al., "Mammalian Deltavirus Without Hepadnavirus Coinfection in the Neotropical Rodent Proechimys Semispinosus," *Proceedings of the National Academy of Sciences* 117, no. 30 (July 2020): 17977–17983, doi:10.1073/pnas.2006750117.

[118] Udo Hetzel et al., "Identification of a Novel Deltavirus in Boa Constrictors," ed. Anne Moscona, *mBio* 10, no. 2 (April 2019): e00014–19, doi:10.1128/mBio.00014-19; Leonora Szirovicza et al., "Snake Deltavirus Utilizes Envelope Proteins of Different Viruses To Generate Infectious Particles," *mBio* 11, no. 2 (March 2020): e03250–19, doi:10.1128/mBio.03250-19.

[119] Wei-Shan Chang et al., "Novel Hepatitis D-like Agents in Vertebrates and Invertebrates," *Virus Evolution* 5, no. 2 (July 2019): vez021, doi:10.1093/ve/vez021.

[120] Chang et al.

[121] Michelle Wille et al., "A Divergent Hepatitis D-Like Agent in Birds," *Viruses* 10, no. 12 (December 2018): 720, doi:10.3390/v10120720.

[122] Laura M. Bergner et al., "Diversification of Mammalian Deltaviruses by Host Shifting," *Proceedings of the National Academy of Sciences* 118, no. 3 (January 2021): e2019907118, doi:10.1073/pnas.2019907118.

[123] Marcos de la Peña et al., "Hepatitis Delta Virus-Like Circular RNAs from Diverse Metazoans Encode Conserved Hammerhead Ribozymes," *Virus Evolution* 7, no. 1 (January 2021): veab016, doi:10.1093/ve/veab016.

[124] "Immunofluorescence Detection of New Antigen-Antibody System (Delta/Anti-Delta) Associated to Hepatitis B Virus in Liver and in Serum of HBsAg Carriers," *Gut* 18, no. 12 (December 1977): 997–1003, doi:10.1136/gut.18.12.997.

[125] Tseng and Lai, "Hepatitis Delta Virus RNA Replication."

[126] Jianming Hu, "Hepatitis B Virus Virology and Replication," in *Hepatitis B Virus in Human Diseases*, ed. Yun-Fan Liaw and Fabien Zoulim (Cham: Springer International Publishing, 2016), 1–34, doi:10.1007/978-3-319-22330-8_1.

such as retroviruses. It is extremely infectious, with a single virion capable of causing infection.[127] HBV has a small circular DNA genome of 3.2 kb that encodes four proteins. During infection, the DNA is transcribed into RNA, which is then reverse transcribed into DNA. Impressively, the HBV polymerase is capable of transcription, reverse transcription, and RNA cleavage.[128] A full discussion of HBV is beyond the scope of this thesis, but it is important to note that the envelope of HDV includes the surface antigens of HBV, which are essential for the entry of HDV into hepatocytes.

Preliminary evidence that HDV may not require HBV has emerged but requires further study. A recent study was able to demonstrate non-HBV infection both *in vitro* and *in vivo*.[129] In this work, HDV particles were able to be produced in cells expressing non-HBV envelope glycoproteins. Moreover, HDV RNA was only detectable extracellularly in the presence of viral envelope proteins while intracellular HDV RNA was detectable in the absence of viral envelope proteins, suggesting that envelope protein coexpression is necessary for the release of HDV RNA. With respect to the HDV particles, they too were able to be released in the absence of HBV envelope proteins, suggesting that the release of HDV particles is independent of HBV. These non-HBV enveloped particles were able to infect human hepatocytes but were inhibited by lonafarnib, which prevents HDV assembly. Infectivity in the presence of Hepatitis C virus (HCV) was also demonstrated in mice. Whether HDV is able to infect humans in the absence of HBV remains to be determined.

Figure 1.6: Secondary structure of human hepatitis delta virus (HDV)

1.3.2 Structure and replication

Ribozyviruses have an approximately 1.7 kb covalently closed circular RNA genome[130] that encodes a single protein, the δ antigen (δAg), which is essential for replication.[131] Size variation among the non-human ribozyviruses range from the 1,547 nt Chusan Island toad virus 1 to the 1,735 nt Chinese fire belly newt virus 1.[132] The secondary structure of the ribozyvirus genome, shown in Figure 1.6, is similar to that of pospiviroids in that it is rod-shaped and contains a high degree of self-complementarity. Despite the high levels of self-complementarity in the genome's secondary structure, it is not targeted by the host RNA silencing machinery,[133] likely due to the bulges when the self-complementarity is not perfect. Ribozyvirus RNA contains a unique ribozyme, the HDV ribozyme (HDVR), to cleave the multimeric intermediates of the RCR process.[134] Recently, a second ribozyme, the HHR, was discovered in HDV in homologous locations within the genome.[135]

Within the virion, the RNA forms a ribonucleoprotein (RNP) complex with the δAg, which is essential for the replication of the virus.[136] The δAg has two isoforms,

[127] Shinichi Asabe et al., "The Size of the Viral Inoculum Contributes to the Outcome of Hepatitis B Virus Infection," *Journal of Virology* 83, no. 19 (October 2009): 9652–9662, doi:10.1128/JVI.00867-09.

[128] Hu, "Hepatitis B Virus Virology and Replication."

[129] Jimena Perez-Vargas et al., "Enveloped Viruses Distinct from HBV Induce Dissemination of Hepatitis D Virus in Vivo," *Nature Communications* 10 (May 2019): 2098, doi:10.1038/s41467-019-10117-z.

[130] A. Kos et al., "The Hepatitis Delta (δ) Virus Possesses a Circular RNA," *Nature* 323, no. 6088 (October 1986): 558–560, doi:10.1038/323558a0; Kang-Sheng Wang et al., "Structure, Sequence and Expression of the Hepatitis Delta (δ) Viral Genome," *Nature* 323, no. 6088 (October 1986): 508–514, doi:10.1038/323508a0.

[131] J. M. Taylor, "Structure and Replication of Hepatitis Delta Virus RNA," *Current Topics in Microbiology and Immunology* 307 (2006): 1–23, doi:10.1007/3-540-29802-9_1.

[132] Chang et al., "Novel Hepatitis D-like Agents in Vertebrates and Invertebrates."

[133] Chang, Provost, and Taylor, "Resistance of Human Hepatitis Delta Virus RNAs to Dicer Activity."

[134] Adrian R. Ferré-D'Amaré, Kaihong Zhou, and Jennifer A. Doudna, "Crystal Structure of a Hepatitis Delta Virus Ribozyme," *Nature* 395, no. 6702 (October 1998): 567–574, doi:10.1038/26912.

[135] De la Peña et al., "Hepatitis Delta Virus-Like Circular RNAs from Diverse Metazoans Encode Conserved Hammerhead Ribozymes."

[136] W S Ryu et al., "Ribonucleoprotein Complexes of Hepatitis Delta Virus." *Journal of Virology* 67, no. 6 (June 1993): 3281–3287, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC237669/; Brittany L. Griffin et al., "Hepatitis Delta Antigen Requires a Flexible Quasi-Double-Stranded

the 24 kDa small δAg and the 27 kDa large δAg, which are 195 and 214 amino acids long, respectively.[137] RNA editing of the antigenome, in which a UAG stop codon is deaminated to a UGG tryptophan codon, produces the large δAg.[138] The δAg plays a role in HDV virion assembly,[139] nuclear import,[140] inhibition of RNA editing,[141] and replication regulation.[142] Its role in replication and RNA editing regulation is notable, as expression of the large δAg is able to inhibit HDV replication and therefore its own expression.[143] Extensive post-translational modification of the δAg is also important for its function.

In HDV, the 19nm RNP is within a 36 nm envelope that contains the HBV surface antigens, which are essential for the entry into hepatocytes.[144] These proteins are the only contributions of HBV to HDV's infection cycle. Only HDV RNA of the positive polarity is packaged into the virion, with the negative polarity being replicated within the host cell. The precise mechanism is still not fully understood.

RNA Structure To Bind and Condense Hepatitis Delta Virus RNA in a Ribonucleoprotein Complex," *Journal of Virology* 88, no. 13 (July 2014): 7402–7411, doi:10.1128/JVI.00443-14.

[137] Zaigham Abbas and Rafia Afzal, "Life Cycle and Pathogenesis of Hepatitis D Virus: A Review," *World Journal of Hepatology* 5, no. 12 (December 2013): 666–675, doi:10.4254/wjh.v5.i12.666.

[138] Andrew G. Polson et al., "Hepatitis Delta Virus RNA Editing Is Highly Specific for the Amber/W Site and Is Suppressed by Hepatitis Delta Antigen," *Molecular and Cellular Biology* 18, no. 4 (April 1998): 1919, doi:10.1128/mcb.18.4.1919.

[139] F L Chang et al., "The Large Form of Hepatitis Delta Antigen Is Crucial for Assembly of Hepatitis Delta Virus." *Proceedings of the National Academy of Sciences of the United States of America* 88, no. 19 (October 1991): 8490–8494, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC52534/.

[140] Huei-Chi Chou et al., "Hepatitis Delta Antigen Mediates the Nuclear Import of Hepatitis Delta Virus RNA," *Journal of Virology* 72, no. 5 (May 1998): 3684–3690, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC109589/.

[141] Polson et al., "Hepatitis Delta Virus RNA Editing Is Highly Specific for the Amber/W Site and Is Suppressed by Hepatitis Delta Antigen."

[142] M Chao, S Y Hsieh, and J Taylor, "Role of Two Forms of Hepatitis Delta Virus Antigen: Evidence for a Mechanism of Self-Limiting Genome Replication." *Journal of Virology* 64, no. 10 (October 1990): 5066–5069, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC247998/.

[143] Chao, Hsieh, and Taylor; Shuji Sato, Cromwell Cornillez-Ty, and David W. Lazinski, "By Inhibiting Replication, the Large Hepatitis Delta Antigen Can Indirectly Regulate Amber/W Editing and Its Own Expression," *Journal of Virology* 78, no. 15 (August 2004): 8120–8134, doi:10.1128/JVI.78.15.8120-8134.2004; Lucy E. Modahl and Michael M. C. Lai, "The Large Delta Antigen of Hepatitis Delta Virus Potently Inhibits Genomic but Not Antigenomic RNA Synthesis: A Mechanism Enabling Initiation of Viral Replication," *Journal of Virology* 74, no. 16 (August 2000): 7375–7380, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC112257/.

[144] Sureau and Negro, "The Hepatitis Delta Virus."

There is controversy over which polymerases are used by HDV for replication. It is possible that, during RCR, the two polarities are transcribed by different host polymerases as evidenced by their reported differential inhibition by α-amanitin[145] and localisation.[146] However, there remains debate over whether antigenomic transcription is indeed inhibited by α-amanitin, as other studies have found that it is, thereby suggesting that both polarities are transcribed by the same polymerase.[147] Use of Pol II is also supported by its interaction with the δAg.[148] The δAg likely can bind to Pol II, stimulating elongation[149] at the cost of fidelity,[150]

[145] Macnaughton et al., "Rolling Circle Replication of Hepatitis Delta Virus RNA Is Carried Out by Two Different Cellular RNA Polymerases"; Modahl et al., "RNA-Dependent Replication and Transcription of Hepatitis Delta Virus RNA Involve Distinct Cellular RNA Polymerases."

[146] Yi-Jia Li et al., "RNA-Templated Replication of Hepatitis Delta Virus: Genomic and Antigenomic RNAs Associate with Different Nuclear Bodies," *Journal of Virology* 80, no. 13 (July 2006): 6478–6486, doi:10.1128/JVI.02650-05; Wen-Hung Huang, Yen-Shun Chen, and Pei-Jer Chen, "Nucleolar Targeting of Hepatitis Delta Antigen Abolishes Its Ability to Initiate Viral Antigenomic RNA Replication," *Journal of Virology* 82, no. 2 (January 2008): 692–699, doi:10.1128/JVI.01155-07.

[147] Gloria Moraleda and John Taylor, "Host RNA Polymerase Requirements for Transcription of the Human Hepatitis Delta Virus Genome," *Journal of Virology* 75, no. 21 (November 2001): 10161–10169, doi:10.1128/JVI.75.21.10161-10169.2001; Jinhong Chang et al., "Action of Inhibitors on Accumulation of Processed Hepatitis Delta Virus RNAs," *Journal of Virology* 80, no. 7 (April 2006): 3205–3214, doi:10.1128/JVI.80.7.3205-3214.2006; Jinhong Chang et al., "Transcription of Hepatitis Delta Virus RNA by RNA Polymerase II," *Journal of Virology* 82, no. 3 (February 2008): 1118–1127, doi:10.1128/JVI.01758-07; T B Fu and J Taylor, "The RNAs of Hepatitis Delta Virus Are Copied by RNA Polymerase II in Nuclear Homogenates." *Journal of Virology* 67, no. 12 (December 1993): 6965–6972, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC238155/.

[148] Natali Abeywickrama-Samarakoon et al., "Hepatitis Delta Virus Histone Mimicry Drives the Recruitment of Chromatin Remodelers for Viral RNA Replication," *Nature Communications* 11, no. 1 (January 2020): 419, doi:10.1038/s41467-020-14299-9; Huang, Chen, and Chen, "Nucleolar Targeting of Hepatitis Delta Antigen Abolishes Its Ability to Initiate Viral Antigenomic RNA Replication"; Dan Cao et al., "Combined Proteomic–RNAi Screen for Host Factors Involved in Human Hepatitis Delta Virus Replication," *RNA* 15, no. 11 (November 2009): 1971–1979, doi:10.1261/rna.1782209; Chunfen Zhang, Honggao Yan, and Zachary F. Burton, "Combinatorial Control of Human RNA Polymerase II (RNAP II) Pausing and Transcript Cleavage by Transcription Factor IIF, Hepatitis Delta Antigen, and Stimulatory Factor II," *The Journal of Biological Chemistry* 278, no. 50 (December 2003): 50101–50111, doi:10.1074/jbc.M307590200; reviewed in Yuki Yamaguchi and Hiroshi Handa, "Hepatitis Delta Antigen and RNA Polymerase II," in *Madame Curie Bioscience Database* (Landes Bioscience, 2013), https://www.ncbi.nlm.nih.gov/books/NBK6547/; reviewed in Valerie Greco-Stewart and Martin Pelchat, "Interaction of Host Cellular Proteins with Components of the Hepatitis Delta Virus," *Viruses* 2, no. 1 (January 2010): 189–212, doi:10.3390/v2010189.

[149] Yuki Yamaguchi et al., "Stimulation of RNA Polymerase II Elongation by Hepatitis Delta Antigen," *Science* 293, no. 5527 (July 2001): 124–127, doi:10.1126/science.1057925.

[150] Yuki Yamaguchi et al., "Hepatitis Delta Antigen Binds to the Clamp of RNA Polymerase II and Affects Transcriptional Fidelity," *Genes to Cells* 12, no. 7 (2007): 863–875, doi:10.1111/j.1365-2443.2007.01094.x.

although evidence to the contrary exists.[151] The use of two different polymerases would be less parsimonious and, in the presence of evidence to the contrary,[152] should be considered unlikely. It is universally agreed that mRNA synthesis is performed by host RNA polymerase II, as evidenced by the presence of 5' caps and 3' polyadenylation.[153]

1.3.3 Symptoms

Hepatitis D can present as either an acute or chronic infection. The acute form is generally indistinguishable from other acute viral hepatitis infections,[154] with symptoms including fatigue, nausea, lethargy, anorexia, and jaundice. Acute liver failure may occur despite no prior liver disease. The chronic form of the disease is more severe, with symptoms including cirrhosis, ascites, and hepatocellular carcinoma (HCC).[155] Over 90% of cases of acute superinfection progress to chronic infection,[156] compared to the 95% of cases of coinfection that result in complete clearance of the virus in less than 6 months.[157] 70% of those with chronic hepatitis D will develop cirrhosis within 5–10 years.[158] A 28-year study of 299 patients with chronic hepatitis D found that 25% (46 patients) had developed hepatocellular

[151] Ziying Han et al., "Intracellular Localization of Hepatitis Delta Virus Proteins in the Presence and Absence of Viral RNA Accumulation," *Journal of Virology* 83, no. 13 (July 2009): 6457–6463, doi:10.1128/JVI.00008-09; T. B. MacNaughton et al., "Hepatitis Delta Antigen Is Necessary for Access of Hepatitis Delta Virus RNA to the Cell Transcriptional Machinery but Is Not Part of the Transcriptional Complex," *Virology* 184, no. 1 (September 1991): 387–390, doi:10.1016/0042-6822(91)90855-6; Shiao-Ya Hong and Pei-Jer Chen, "Phosphorylation of Serine 177 of the Small Hepatitis Delta Antigen Regulates Viral Antigenomic RNA Replication by Interacting with the Processive RNA Polymerase II," *Journal of Virology* 84, no. 3 (February 2010): 1430–1438, doi:10.1128/JVI.02083-09.

[152] Chang et al., "Transcription of Hepatitis Delta Virus RNA by RNA Polymerase II."

[153] S. Y. Hsieh et al., "Hepatitis Delta Virus Genome Replication: A Polyadenylated mRNA for Delta Antigen," *Journal of Virology* 64, no. 7 (July 1990): 3192, doi:10.1128/jvi.64.7.3192-3198.1990; Severin Gudima et al., "Characterization of the 5′ Ends for Polyadenylated RNAs Synthesized During the Replication of Hepatitis Delta Virus," *Journal of Virology* 73, no. 8 (August 1999): 6533, doi:10.1128/jvi.73.8.6533-6539.1999.

[154] Patrizia Farci and Grazia Niro, "Clinical Features of Hepatitis D," *Seminars in Liver Disease* 32, no. 3 (August 2012): 228–236, doi:10.1055/s-0032-1323628.

[155] Farci and Niro; Heiner Wedemeyer and Michael P. Manns, "Epidemiology, Pathogenesis and Management of Hepatitis D: Update and Challenges Ahead," *Nature Reviews Gastroenterology & Hepatology* 7, no. 1 (January 2010): 31–40, doi:10.1038/nrgastro.2009.205.

[156] Farci and Niro, "Clinical Features of Hepatitis D."

[157] Calvin Pan et al., "Diagnosis and Management of Hepatitis Delta Virus Infection," *Digestive Diseases and Sciences* 68, no. 8 (August 2023): 3237–3248, doi:10.1007/s10620-023-07960-y.

[158] Farci and Niro, "Clinical Features of Hepatitis D."

carcinoma (HCC) and 21% (63 patients) had died, with liver failure the cause of death in the majority of cases.[159]

In non-human hosts, the symptoms of ribozyviruses are much less well understood. Snake deltavirus was found in snakes with suspected inclusion body disease, a fatal disease of snakes that is characterized by the presence of intracytoplasmic inclusions in the cells of the central nervous system and other tissues.[160] Sequencing various tissues from infected snakes revealed the presence of the virus in the brain, liver, lung, spleen, and kidney, with the highest viral loads in the brain and liver. Whether other ribozyviruses cause similar symptoms in their hosts is unknown, as they were discovered via metatranscriptomics and not through the observation of symptoms.[161]

1.3.4 Pathogenicity

Evidence with respect to pathogenicity mechanisms is scarce. In human hepatitis D, pathogenicity is believed to be due to the interaction with host immune system.[162] The eight HDV genotypes exhibit different pathogenicities, although the mechanisms behind this are not well understood.[163] Both innate and adaptive immune responses are believed to be involved in the pathogenesis of HDV.[164] Within the cell, expression of the small δAg is able to downregulate the p53 protein, which is a tumor suppressor.[165] How this downregulation contributes to the pathogenesis of HDV is not well understood, nor is whether it is related to the development of HCC. It is certainly a tempting hypothesis, given that p53 is a key regulator of

[159] Raffaella Romeo et al., “A 28-Year Study of the Course of Hepatitis Delta Infection: A Risk Factor for Cirrhosis and Hepatocellular Carcinoma,” *Gastroenterology* 136, no. 5 (May 2009): 1629–1638, doi:10.1053/j.gastro.2009.01.052.

[160] Hetzel et al., “Identification of a Novel Deltavirus in Boa Constrictors.”

[161] Wille et al., “A Divergent Hepatitis D-Like Agent in Birds.”

[162] Grazia Anna Niro and Antonina Smedile, “Current Concept in the Pathophysiology of Hepatitis Delta Infection,” *Current Infectious Disease Reports* 14, no. 1 (February 2012): 9–14, doi:10.1007/s11908-011-0233-5; Wedemeyer and Manns, “Epidemiology, Pathogenesis and Management of Hepatitis D.”

[163] Farci and Niro, “Clinical Features of Hepatitis D.”

[164] Eirini D. Tseligka, Sophie Clément, and Francesco Negro, “HDV Pathogenesis: Unravelling Ariadne’s Thread,” *Viruses* 13, no. 5 (April 2021): 778, doi:10.3390/v13050778.

[165] Marta Mendes et al., “Proteomic Changes in HEK-293 Cells Induced by Hepatitis Delta Virus Replication,” *Journal of Proteomics* 89 (August 2013): 24–38, doi:10.1016/j.jprot.2013.06.002.

the cell cycle, apoptosis, and DNA repair. However, a variety of other possible, perhaps non-exclusive, mechanisms for HCC pathogenesis have been proposed.[166]

1.4 Retrozymes

Recently, retrotransposons containing hammerhead ribozymes (HHR) similar to those present in viroids, satRNAs, and some ribozyviruses were discovered in plant genomes.[167] These ribozyme-containing retrotransposons, named retrozymes, are not autonomous and apparently depend on the reverse transcriptase activity of Ty3-like retrotransposons. The retrozymes consist of a 300–700 nt non-coding sequence flanked by long terminal repeats (LTRs) and, apart from the presence of the HHR, resemble other small, non-autonomous retrotransposons that are abundant in plants, such as terminal-repeat retrotransposons in miniature (TRIMs)[168] or small LTR retrotransposons (SMARTs).[169] The retrozymes are actively transcribed, the transcripts are self-cleaved by HHR and form abundant cccRNAs of the positive polarity.[170] Notably, the circularisation is catalysed *in vitro* by chloroplast tRNA ligase as is the case also in avsunviroids. The RNA sequences of retrozymes are predicted to fold into stable, branched structures resembling those of avsunviroids.

---

[166] Reviewed in Abbas and Afzal, "Life Cycle and Pathogenesis of Hepatitis D Virus"; Patrizia Farci et al., "Hepatitis D Virus and Hepatocellular Carcinoma," *Viruses* 13, no. 5, 5 (May 2021): 830, doi:10.3390/v13050830.

[167] Amelia Cervera, Denisse Urbina, and Marcos de la Peña, "Retrozymes Are a Unique Family of Non-Autonomous Retrotransposons with Hammerhead Ribozymes That Propagate in Plants Through Circular RNAs," *Genome Biology* 17, no. 1 (December 2016): 135, doi:10.1186/s13059-016-1002-4; Marcos de la Peña, "Circular RNAs Biogenesis in Eukaryotes Through Self-Cleaving Hammerhead Ribozymes," *Advances in Experimental Medicine and Biology* 1087 (2018): 53–63, doi:10.1007/978-981-13-1426-1_5; Amelia Cervera and Marcos de la Peña, "Cloning and Detection of Genomic Retrozymes and Their circRNA Intermediates," in *Ribozymes: Methods and Protocols*, ed. Robert J. Scarborough and Anne Gatignol, Methods in Molecular Biology (New York, NY: Springer US, 2021), 27–44, doi:10.1007/978-1-0716-0716-9_3.

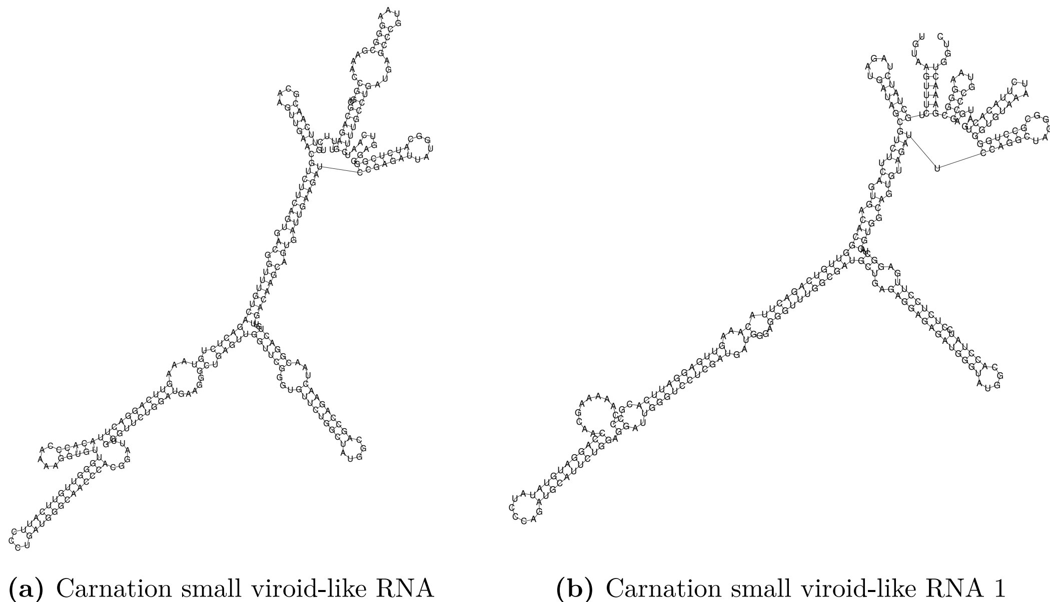
[168] C.-P. Witte et al., "Terminal-Repeat Retrotransposons in Miniature (TRIM) Are Involved in Restructuring Plant Genomes," *Proceedings of the National Academy of Sciences* 98, no. 24 (November 2001): 13778–13783, doi:10.1073/pnas.241341898.

[169] Dongying Gao et al., "A Highly Conserved, Small LTR Retrotransposon That Preferentially Targets Genes in Grass Genomes," ed. Mark A. Batzer, *PLoS ONE* 7, no. 2 (February 2012): e32010, doi:10.1371/journal.pone.0032010.

[170] Marcos de la Peña and Amelia Cervera, "Circular RNAs with Hammerhead Ribozymes Encoded in Eukaryotic Genomes: The Enemy at Home," *RNA Biology* 14, no. 8 (August 2017): 985–991, doi:10.1080/15476286.2017.1321730; Marcos de la Peña, Raquel Ceprián, and Amelia Cervera, "A Singular and Widespread Group of Mobile Genetic Elements: RNA Circles with Autocatalytic Ribozymes," *Cells* 9, no. 12 (November 2020): E2555, doi:10.3390/cells9122555.

In these structures, over 70% of the nucleotides are paired, with the HHR located within LTRs.[171] Even smaller, non-LTR retrozymes of 170–400 nt have been discovered in genomes of diverse vertebrate and invertebrate animals and also shown to form abundant cccRNAs.[172]

1.5 Retroviroids



(a) Carnation small viroid-like RNA **(b)** Carnation small viroid-like RNA 1

Figure 1.7: Representative secondary structures of retroviroids, produced using RNAfold[173]

One last viroid-like agent is the retroviroid-like element (hereafter retroviroid), carnation small viroid-like RNA (CarSV) and its DNA form.[174] CarSV, the only

[171] Cervera, Urbina, and de la Peña, "Retrozymes Are a Unique Family of Non-Autonomous Retrotransposons with Hammerhead Ribozymes That Propagate in Plants Through Circular RNAs"; Amelia Cervera and Marcos de la Peña, "Small circRNAs with Self-Cleaving Ribozymes Are Highly Expressed in Diverse Metazoan Transcriptomes," *Nucleic Acids Research* 48, no. 9 (May 2020): 5054–5064, doi:10.1093/nar/gkaa187.

[172] Cervera and de la Peña, "Small circRNAs with Self-Cleaving Ribozymes Are Highly Expressed in Diverse Metazoan Transcriptomes."

[173] Lorenz et al., "ViennaRNA Package 2.0."

[174] José-Antonio Daròs and Ricardo Flores, "Identification of a Retroviroid-Like Element from Plants," *Proceedings of the National Academy of Sciences* 92, no. 15 (July 1995): 6856–6860, doi:10.1073/pnas.92.15.6856; A. Vera et al., "The DNA of a Plant Retroviroid-Like Element Is Fused to Different Sites in the Genome of a Plant Pararetrovirus and Shows Multiple Forms with Sequence Deletions," *Journal of Virology* 74, no. 22 (November 2000): 10390–10400, doi:10.1128/jvi.74.22.10390-10400.2000; Krisztina Hegedus, Géza Dallmann, and Ervin Balázs, "The DNA Form of a Retroviroid-Like Element Is Involved in Recombination

retroviroid so far described, is similar to viroids in size, circularity, and the presence of hammerhead ribozymes in both polarities. However, it differs from viroids in that there exists a homologous extrachromosomal DNA sequence that can be integrated into the plant genome by means of a helper pararetrovirus. A variant of CarSV, named CarSV-1, has recently been found, although it remains unstudied.[175] Unlike viroids however, CarSV does not appear to transmit horizontally from plant to plant.

1.6 Ribozymes

The key unifying motif in all of these agents (save only pospiviroids) is the catalytic RNA enzyme that enables RCR. These ribozymes are ubiquitous not only in viroids and viroid-like RNAs but also among the genomes of bacteria, plants, and animals.[176] The primary ribozyme in the subviral world is the hammerhead. *In vitro* evidence suggests that this ribozyme may have multiple independent origins, which would help explain some of the ubiquity.[177] Indeed, in viroids, no other ribozyme is used. However, in satRNAs, another ribozyme, motif is found. The hairpin ribozyme, introduced in Section 1.2.3, is capable of both cleavage and ligation.[178] In some cases, a satRNA may have an HPR in one polarity and an HHR in the other, demonstrating that this motif is modular. A variety of viroid-like circular RNAs has been identified with diverse HPRs and in combination with

Events with Itself and with the Plant Genome," *Virology* 325, no. 2 (August 2004): 277–286, doi:10.1016/j.virol.2004.04.035.

[175] Timo M. Breit et al., "Genome Sequence of a New Carnation Small Viroid-Like RNA, CarSV-1," ed. Jelle Matthijnssens, *Microbiology Resource Announcements* 12, no. 3 (March 2023): e01219–22, doi:10.1128/mra.01219-22.

[176] M. de la Pena and I. Garcia-Robles, "Ubiquitous Presence of the Hammerhead Ribozyme Motif Along the Tree of Life," *RNA* 16, no. 10 (October 2010): 1943–1950, doi:v; Jonathan Perreault et al., "Identification of Hammerhead Ribozymes in All Domains of Life Reveals Novel Structural Variations," ed. Wyeth W. Wasserman, *PLoS Computational Biology* 7, no. 5 (May 2011): e1002031, doi:10.1371/journal.pcbi.1002031; Randi M. Jimenez, Eric Delwart, and Andrej Lupták, "Structure-Based Search Reveals Hammerhead Ribozymes in the Human Microbiome," *Journal of Biological Chemistry* 286, no. 10 (March 2011): 7737–7743, doi:10.1074/jbc.C110.209288.

[177] Kourosh Salehi-Ashtiani and Jack W. Szostak, "In Vitro Evolution Suggests Multiple Origins for the Hammerhead Ribozyme," *Nature* 414, no. 6859 (November 2001): 82–84, doi:10.1038/35102081.

[178] Fedor, "Tertiary Structure Stabilization Promotes Hairpin Ribozyme Ligation"; Fedor, "Structure and Function of the Hairpin Ribozyme"; Kazakov, "Ligation of the Hairpin Ribozyme in Cis Induced by Freezing and Dehydration."

other riboyzmes.[179] This ribozyme modularity is also the case for ribozyviruses, in which the exact same place on the genome where the unique HDVR is found (the end of the rod-shaped secondary structure) instead harbors an HHR in some cases.[180] Methods to identify novel ribozymes have focused on looking in the neighborhood of known ribozymes such as the HHR to identify common motifs.[181] These approaches have yielded other ribozymes, some of which had been identified viroid-like RNAs although others (such as the twister, CPEB3, and twister-sister) had not at the start of this thesis research.

These ribozymes contain helical segments linked by unpaired, conserved nucleotides that facilitate tertiary interactions, compacting the RNA and forming the catalytic center crucial for cleavage reaction rate enhancement.[182] Four classical principles (α-δ) define the catalytic mechanism: in-line alignment for nucleophilic attack (α), neutralization of non-bridging oxygens (β), deprotonation of the 2' OH (γ), and charge neutralization at the 5 oxygen (δ). Two additional principles (γ' and γ'') have been identified, both involving the activation of the 2' OH nucleophile via guanine interaction, either by hydrogen bond donation (γ') or via competitive hydrogen bonding reducing inhibitory interactions γ''.[183] However, each ribozyme has mechanistic differences, with the HDV and twister sister ribozymes requiring a divalent metal ion for catalysis.[184] As additional ribozymes are discovered, the

[179] Christina E Weinberg et al., "Identification of over 200-Fold More Hairpin Ribozymes Than Previously Known in Diverse Circular RNAs," *Nucleic Acids Research* 49, no. 11 (June 2021): 6375–6388, doi:10.1093/nar/gkab454.

[180] De la Peña et al., "Hepatitis Delta Virus-Like Circular RNAs from Diverse Metazoans Encode Conserved Hammerhead Ribozymes."

[181] Adam Roth et al., "A Widespread Self-Cleaving Ribozyme Class Is Revealed by Bioinformatics," *Nature Chemical Biology* 10, no. 1 (January 2014): 56–60, doi:10.1038/nchembio.1386; Zasha Weinberg et al., "New Classes of Self-Cleaving Ribozymes Revealed by Comparative Genomics Analysis," *Nature Chemical Biology* 11, no. 8 (August 2015): 606–610, doi:10.1038/nchembio.1846.

[182] Christina E Weinberg, Zasha Weinberg, and Christian Hammann, "Novel Ribozymes: Discovery, Catalytic Mechanisms, and the Quest to Understand Biological Function," *Nucleic Acids Research* 47, no. 18 (October 2019): 9480–9494, doi:10.1093/nar/gkz737.

[183] Daniel D. Seith et al., "Elucidation of Catalytic Strategies of Small Nucleolytic Ribozymes From Comparative Analysis of Active Sites," *ACS Catalysis* 8, no. 1 (January 2018): 314–327, doi:10.1021/acscatal.7b02976.

[184] Yijin Liu, Timothy J. Wilson, and David M. J. Lilley, "The Structure of a Nucleolytic Ribozyme That Employs a Catalytic Metal Ion," *Nature Chemical Biology* 13, no. 5 (May 2017): 508–513, doi:10.1038/nchembio.2333.

diversity of ribozyme mechanisms is likely to increase as well.

1.7 Origin and evolution of viroid-like RNAs

Thus far, I have discussed the various types of viroid-like RNAs, their replication mechanisms, and their impact on their hosts. Omitted from this discussion is their origin and evolution. Indeed, among the most puzzling questions about viroids and viroid-like RNAs is where these diminutive replicators came from. While conclusive evidence is lacking, hypotheses abound, more or less mirroring the main scenarios considered for the origin of viruses.[185] These hypotheses are:

1. Origin from the pre-cellular replicator pool
2. Regression from existing replicators
3. Escape from existing genomes

The pre-cellular origin hypothesis has a twist in the case of viroids. Unlike the virus-first pre-cellular origin scenario, which postulates that virus-like replicators evolved from pre-cellular replicators, the viroid version holds that viroids are not only direct descendants of pre-cellular replicators but also holdovers from the pre-protein RNA world, as evidenced by their use of ribozymes for (some stages of) replication.[186] Later, ribozymes would again provide the engine for the transition to the RNA-protein world by serving as the catalytic core of the ribosome. In some of these "viroid early" scenarios, viroids or viroid-like RNAs come across not only as relics from the RNA world but also descendants of the first RNA world replicators.[187] Hence, there is significant special interest in viroid-like agents as a unique window into the hypothetical pre-cellular RNA world.

[185] Mart Krupovic, Valerian V. Dolja, and Eugene V. Koonin, "Origin of Viruses: Primordial Replicators Recruiting Capsids from Hosts," *Nature Reviews Microbiology* 17, no. 7 (July 2019): 449–458, doi:10.1038/s41579-019-0205-6.

[186] T. O. Diener, "Circular RNAs: Relics of Precellular Evolution?" *Proceedings of the National Academy of Sciences* 86, no. 23 (December 1989): 9370–9374, doi:10.1073/pnas.86.23.9370.

[187] Karin Moelling and Felix Broecker, "Viroids and the Origin of Life," *International Journal of Molecular Sciences* 22, no. 7 (March 2021): 3476, doi:10.3390/ijms22073476.

1.7.1 Possible mechanisms of emergence

A broad variety of possible alternative scenarios for viroid emergence have been proposed (Figure 1.8). The most radical idea is that viroids emerged *de novo*.[188] Under this model, viroids are not the descendants of any known replicator but rather emerged spontaneously from a pool of random RNA sequences. Variations on this hypothesis include different timelines for the emergence of viroids, ranging from the pre-cellular RNA world to a potentially ongoing process in cells.

Hard evidence for this mechanism is scant and difficult to obtain. However, both computational modelling[189] and *in vitro* experiments[190] indicate that RNAs with properties similar to those of viroids might be produced *de novo*. Specifically, simulations suggest that rod-shaped and branched circular structures can form when short random RNA sequences fold into hairpins, which then catalyse ligation with other random hairpins, finally producing circular structures. These cccRNAs could then grow in length by both recombination and random insertions. More importantly, such RNAs have the potential to evolve such that sequences with greater stability, robustness, and polymerase affinity being selected for. Thus, Catalán et al.[191] conclude that *de novo* emergence of viroid-like RNA molecules through a stepwise evolutionary process seeded by any of the diverse small RNAs present in cells, such as microRNAs, is plausible.

Experimental evidence bears a similar picture, albeit with a different replication mechanism. Inspired by the role of DdRPs in viroid and ribozyvirus replication, Jain et al.[192] demonstrated that high concentrations of T7 phage DdRP can generate populations of replicating structured linear RNAs from DNA seeds in vitro. Unlike the standard model of viroid replication, which is based on RCR, with monomers ligating to form antigenomic cccRNAs, the mechanism these replicators

[188] Diener, "Circular RNAs."

[189] Pablo Catalán et al., "Parsimonious Scenario for the Emergence of Viroid-Like Replicons De Novo," *Viruses* 11, no. 5 (May 2019): 425, doi:10.3390/v11050425.

[190] Nimit Jain et al., "Transcription Polymerase–Catalyzed Emergence of Novel RNA Replicons," *Science* 368, no. 6487 (April 2020): eaay0688, doi:10.1126/science.aay0688.

[191] "Parsimonious Scenario for the Emergence of Viroid-Like Replicons De Novo."

[192] "Transcription Polymerase–Catalyzed Emergence of Novel RNA Replicons."

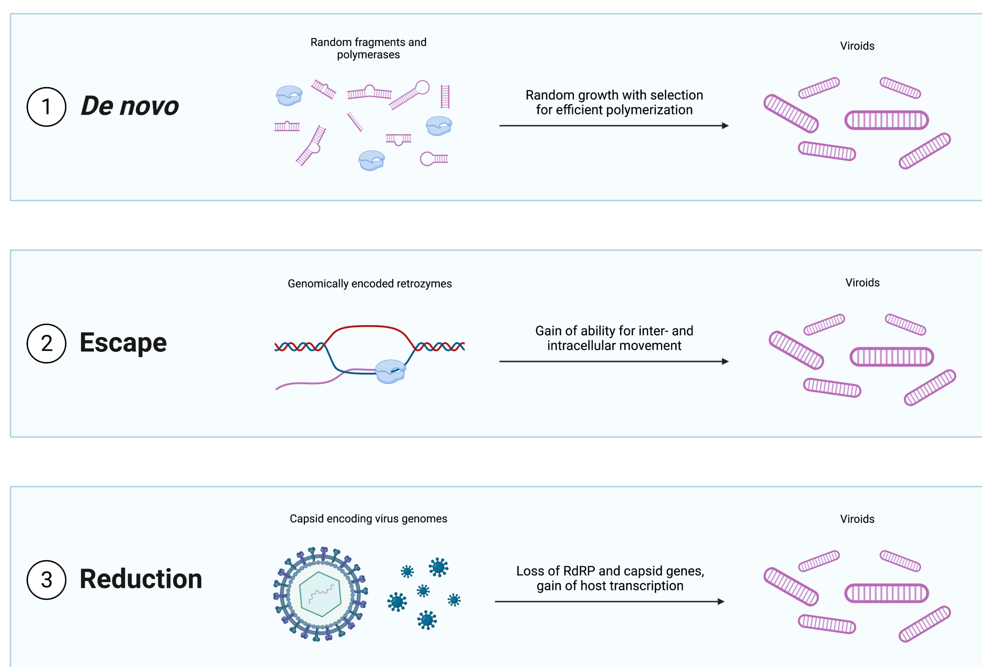


Figure 1.8: Evolutionary scenarios for the emergence of viroid-like replicators

used bypasses ligation altogether: instead, the polymerase jumps from the 5' end of the template back to the 3' end. Recent research also demonstrates that viroid-like self-complementary cccRNAs are readily formed in autocatalytic RNA reaction networks.[193] These lines of research suggest that replicators with the structural hallmarks of viroids could potentially emerge *de novo*, at least in the presence of a suitable polymerase.

With respect to satRNAs, at least one case of *de novo* satRNA emergence has been reported.[194] CMV absent satCMV was shown to have gained a satRNA with similarity to known satCMVs after serial passages in greenhouse-grown plants,

[193] Cyrille Jeancolas et al., "RNA Diversification by a Self-Reproducing Ribozyme Revealed by Deep Sequencing and Kinetic Modelling," *Chemical Communications* 57, no. 61 (2021): 7517–7520, doi:10.1039/D1CC02290C.

[194] M. R. Hajimorad, S. A. Ghabrial, and M. J. Roossinck, "De Novo Emergence of a Novel Satellite RNA of Cucumber Mosaic Virus Following Serial Passages of the Virus Derived from RNA Transcripts," *Archives of Virology* 154, no. 1 (January 2009): 137–140, doi:10.1007/s00705-008-0280-x.

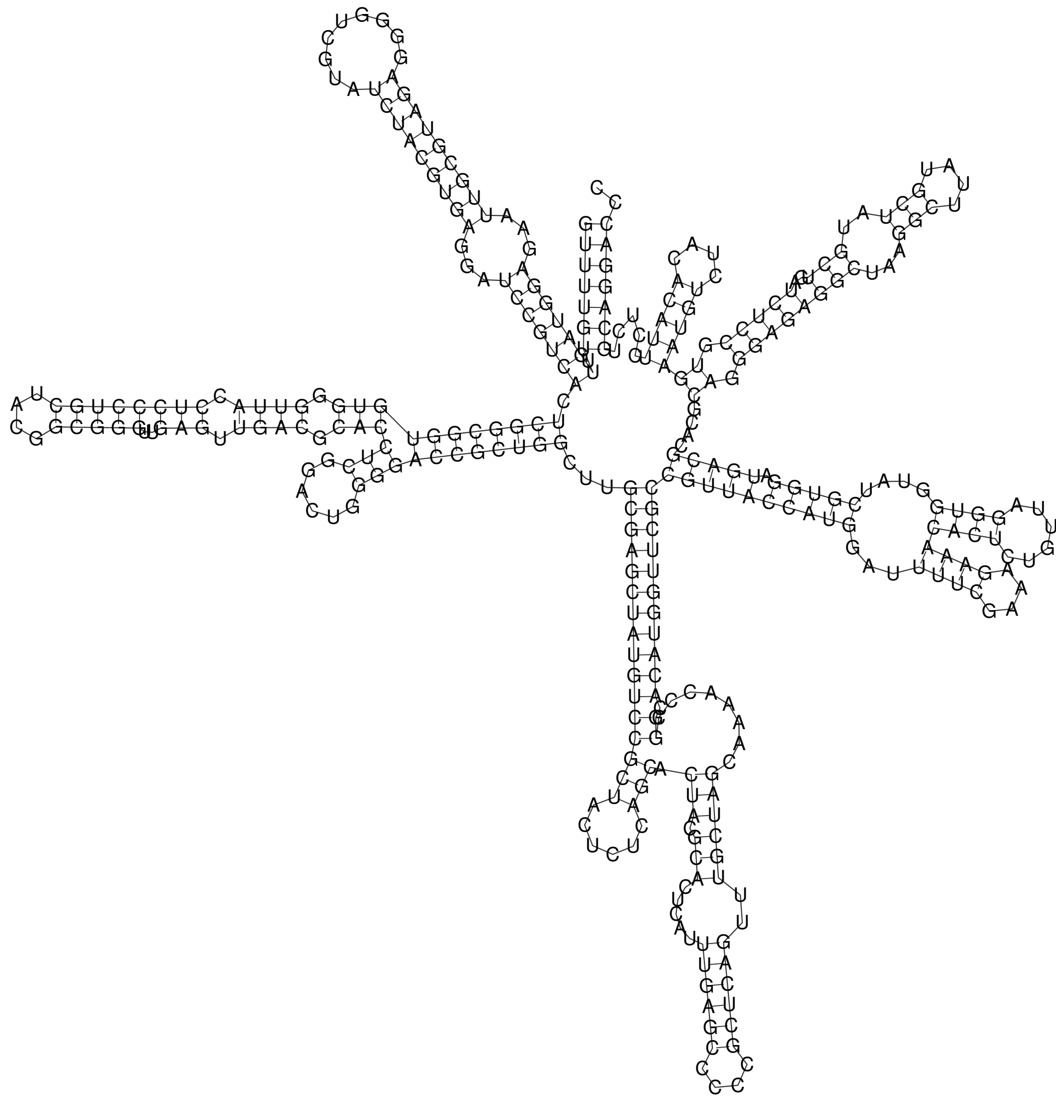


Figure 1.9: Secondary structure of a satRNA that reportedly emerged *de novo*

with the secondary structure predicted by `RNAfold`[195] shown in Figure 1.9. These satRNAs appear to be derived from host sRNAs.[196] Although not especially viroid-like in its absence of ribozymes, it does provide evidence that a replicator can emerge from small RNA fragments in a cellular context.[197]

Other proposed mechanisms for the emergence of viroids are also of considerable

[195] Lorenz et al., "ViennaRNA Package 2.0."

[196] Kiran Zahid et al., "Nicotiana Small RNA Sequences Support a Host Genome Origin of Cucumber Mosaic Virus Satellite RNA," *PLoS Genetics* 11, no. 1 (January 2015): e1004906, doi:10.1371/journal.pgen.1004906.

[197] Hu, Hsu, and Lin, "Satellite RNAs and Satellite Viruses of Plants."

interest when viewed in the broader context of replicator evolution. The genome escape hypothesis posits that viroids are genomic sequences that managed to achieve semi-autonomous replication.[198] Although arguably less exotic than the *de novo* scenario, this model nonetheless provides an example of the emergence of a replicator from a non-replicator. In this case, potential ancestors of viroids could be, for example, group I self-splicing introns, the simplest of which form viroid-sized cccRNAs after splicing catalysed by a ribozyme present in the intron itself.[199]

In the genome reduction model, viroids emerge from pre-existing replicators such as viruses, by shedding all coding regions and relying entirely on host enzymes. The classic *in vitro* evolution experiments of Spiegelman and colleagues[200] clearly demonstrate the phenomenon of viral genome reduction to viroid size by completely eliminating coding capacity. In these paradigmatic experiments, the 3800 nt genome of RNA bacteriophage Qβ ultimately shrank to a 218 nt non-coding RNA (known as "Spiegelman monster") under selection for replication speed in the presence of an excess of phage RdRP.[201] A similar process could have conceivably occurred in vivo. Genome reduction is a common phenomenon in obligate parasites of all kinds ranging from bacteria and archaea to worms to arthropods.[202] In some

[198]T. O. Diener, "Are Viroids Escaped Introns?" *Proceedings of the National Academy of Sciences* 78, no. 8 (August 1981): 5014–5015, doi:10.1073/pnas.78.8.5014.

[199]Paula J. Grabowski, Arthur J. Zaug, and Thomas R. Cech, "The Intervening Sequence of the Ribosomal RNA Precursor Is Converted to a Circular RNA in Isolated Nuclei of Tetrahymena," *Cell* 23, no. 2 (February 1981): 467–476, doi:10.1016/0092-8674(81)90142-2; Kelly Kruger et al., "Self-Splicing RNA: Autoexcision and Autocyclization of the Ribosomal RNA Intervening Sequence of Tetrahymena," *Cell* 31, no. 1 (November 1982): 147–157, doi:10.1016/0092-8674(82)90414-7.

[200]I. Haruna and S. Spiegelman, "Autocatalytic Synthesis of a Viral RNA in Vitro," *Science* 150, no. 3698 (November 1965): 884–886, doi:10.1126/science.150.3698.884; S. Spiegelman et al., "Studies in the Replication of Viral RNA," *Journal of Cellular Physiology* 70, no. S1 (1967): 35–64, doi:10.1002/jcp.1040700405; S. Spiegelman, "An Approach to the Experimental Analysis of Precellular Evolution," *Quarterly Reviews of Biophysics* 4, no. 2–3 (August 1971): 213–253, doi:10.1017/S0033583500000639.

[201]D. R. Mills, F. R. Kramer, and S. Spiegelman, "Complete Nucleotide Sequence of a Replicating RNA Molecule: The Sequence Suggests How Nucleic Acids Exhibit Phenotypes for Selection and Can Evolve to Greater Complexity." *Science* 180, no. 4089 (June 1973): 916–927, doi:10.1126/science.180.4089.916.

[202]Jennifer J Wernegreen, "For Better or Worse: Genomic Consequences of Intracellular Mutualism and Parasitism," *Current Opinion in Genetics & Development* 15, no. 6 (December 2005): 572–583, doi:10.1016/j.gde.2005.09.013; Eugene V. Koonin and Yuri I. Wolf, "Genomics of

such cases, the genome reduction can be extreme, with the genomes of some obligate parasitic and symbiotic bacteria shrinking more than an order of magnitude, and mitochondrial genomes more than two orders of magnitude.[203] Reductive evolution also readily occurs in viruses, in particular, giving rise to defective interfering particles, and to dramatic contraction of virus genomes upon passages in cell cultures.[204] Thus, it appears entirely plausible that viroids are ultimately reduced viruses. Identification of sequences homologous to viroids within viral or cellular genomes could lend credence to either of these hypotheses.

1.7.2 Common ancestry versus multiple origins

Whether all viroid-like agents are monophyletic remains an open question, although parsimony suggests a common ancestry.[205] The attributes of the putative common ancestor are hard to pin down in detail but, given that viroids are both the simplest and most autonomous agents within the brotherhood, it seems likely that this ancestor was either viroid-like or satRNA-like.[206] The hypothesis on the emergence of these agents within the RNA world is generally predicated on that assumption, as the other members of the brotherhood require a cellular context for replication. The alternative possibility, that the ancestor was a ribozy-like virus that gave rise

---

Bacteria and Archaea: The Emerging Dynamic View of the Prokaryotic World," *Nucleic Acids Research* 36, no. 21 (December 2008): 6688–6719, doi:10.1093/nar/gkn668; Yuri I. Wolf and Eugene V. Koonin, "Genome Reduction as the Dominant Mode of Evolution," *BioEssays* 35, no. 9 (September 2013): 829–837, doi:10.1002/bies.201300037.

[203] Nancy A. Moran, "Microbial Minimalism," *Cell* 108, no. 5 (March 2002): 583–586, doi:10.1016/S0092-8674(02)00665-7; John P. McCutcheon and Nancy A. Moran, "Extreme Genome Reduction in Symbiotic Bacteria," *Nature Reviews. Microbiology* 10, no. 1 (November 2011): 13–26, doi:10.1038/nrmicro2670.

[204] Mickaël Boyer et al., "Mimivirus Shows Dramatic Genome Reduction After Intraamoebal Culture," *Proceedings of the National Academy of Sciences of the United States of America* 108, no. 25 (June 2011): 10296–10301, doi:10.1073/pnas.1101118108; Rafael Sanjuán, "The Social Life of Viruses," *Annual Review of Virology* 8, no. 1 (September 2021): 183–199, doi:10.1146/annurev-virology-091919-071712; Tatiana G. Senkevich et al., "Ancient Gene Capture and Recent Gene Loss Shape the Evolution of Orthopoxvirus-Host Interaction Genes," *mBio* 12, no. 4 (July 2021): e01495–21, doi:10.1128/mBio.01495-21.

[205] Theodor O. Diener, "Viroids: 'Living Fossils' of Primordial RNAs?" *Biology Direct* 11, no. 1 (December 2016): 15, doi:10.1186/s13062-016-0116-7.

[206] Eugene V. Koonin and Valerian V. Dolja, "Virus World as an Evolutionary Network of Viruses and Capsidless Selfish Elements," *Microbiology and Molecular Biology Reviews* 78, no. 2 (June 2014): 278–303, doi:10.1128/MMBR.00049-13.

to viroids through reductive evolution, is clearly less parsimonious given their larger size and higher complexity.

Alternatively, if viroids and viroid-like agents are not monophyletic, then they must have emerged independently on multiple occasions. If this were the case, a mixture of the above scenarios could be at play, with some members of the brotherhood emerging *de novo*, others from retrotransposons, and yet others from viruses. Given the fact that HHRs appear to have multiple independent origins,[207] a scenario in which ribozyviruses and viroids do not share a common ancestor is not entirely implausible. Convergent evolution of viroid-like agents from different starting points would be a fascinating example of the power of natural selection to shape replicators into similar forms under similar selective pressures. If multiple origins were the true scenario, then the question of the origin of viroids would be even more complex than it currently appears. However, in the absence of evidence for multiple origins, the most parsimonious explanation remains that viroids and viroid-like agents share a common ancestor.

1.7.3 Retrozymes as a possible ancestor

Of the three principal models of emergence for *bona fide* viroids, two implicitly answer the question of when viroids emerged: both the escape and genome reduction hypotheses, by design, require viroids to have emerged in a cellular context. However, what type of cells? At the start of this thesis, viroids and satRNAs were known to be limited in range to higher plants, whereas ribozyviruses were thought to be limited to animals.[208] Assuming that viroids and viroid-like agents share a common ancestry, horizontal transfer between plants and animals could be an explanation.[209] Hard evidence for the direction of the transfer is lacking, but the greater complexity of ribozyviruses compared to viroids and their limited spread

[207]Salehi-Ashtiani and Szostak, "In Vitro Evolution Suggests Multiple Origins for the Hammerhead Ribozyme."
[208]This claim is later shown to likely be incorrect in Section 4.3.7.
[209]Diener, "Viroids."

in animals, taken together, would seem to point towards a likely origin in plants, with acquisition of the protein-coding gene via recombination in an animal host.

The broad representation of retrozymes in plant genomes is compatible with the origin of viroids via retrozyme escape in plants. Given their apparently limited host range, such escape seemed to be the most likely scenario for the emergence of viroids at the start of this thesis work, in accord with the prescient early hypothesis of Diener.[210] Conversely, the animal retrozymes potentially might have independently given rise to ribozyviruses, as an alternative to the horizontal transfer of viroids discussed above. The evolution of viroid-like replicators from retrozymes seems to combine elements of the escape and reduction scenarios, assuming the retrozymes themselves are products of reductive evolution of autonomous retrotransposons.

In contrast, the viroid-first hypothesis would seem to face multiple difficulties. Probably the most damning difficulty for this scenario has been the limited known host range of viroid-like agents, especially their apparent absence in bacteria or archaea. Like spliceosomal introns and many other varieties of non-coding RNAs, viroid-like cccRNAs belong to the expanded "new RNA world" of eukaryotic cells,[211] and thus, their origin most likely postdates the origin of eukaryotes. Evidence demonstrating a broader host range for viroid-like agents would be a significant challenge to the retrozyme escape hypothesis and a boon to the viroid-early hypothesis.

Beyond the host-range problem, replication of viroid-like RNAs in the hypothetical RNA world also seems problematic. Indeed, the extensive, robust secondary structures of these cccRNAs appear to be adapted to mimic DNA and so fool DdRPs and, possibly, also to avoid RNAi host response.[212] Neither of these factors that

[210]Jeancolas et al., "RNA Diversification by a Self-Reproducing Ribozyme Revealed by Deep Sequencing and Kinetic Modelling."

[211]Allison Jandura and Henry M. Krause, "The New RNA World: Growing Evidence for Long Noncoding RNA Functionality," *Trends in Genetics* 33, no. 10 (October 2017): 665–676, doi:10.1016/j.tig.2017.08.002; Stephen R. Quake, "The Cell as a Bag of RNA," *Trends in Genetics* 37, no. 12 (December 2021): 1064–1068, doi:10.1016/j.tig.2021.08.003.

[212]Beatriz Navarro, Ricardo Flores, and Francesco Di Serio, "Advances in Viroid-Host Interactions," *Annual Review of Virology* 8, no. 1 (September 2021): 305–325, doi:10.1146/annurev-virology-091919-092331.

apparently shape the evolution of viroid-like RNAs is relevant in the primordial setting. On the contrary, replication of such highly structured RNAs by the hypothetical ribozyme polymerases would be hampered, even beyond the more generic problems faced by such polymerases despite extensive experimental attempts to evolve accurate and processive ones.[213] However, the secondary structure would be beneficial in that it could protect the RNA from degradation and yield greater thermal stability, which would be advantageous in the prebiotic world.

Thus, at the start of this thesis, the retrozyme escape hypothesis appeared to be the most parsimonious scenario for the origin of viroid-like cccRNAs. Refutation of this hypothesis would imply that retrozymes either are integrated viroids or a case of convergence. Although HHRs appear to have evolved on multiple occasions independently,[214] the similarities between retrozymes and viroids are so pronounced that the most likely explanation remains that viroids are escaped retrozymes.

As fully described in Chapter 4, the discovery of a vast diversity of viroid-like RNAs in metatranscriptomic data from a wide variety of hosts has dramatically changed the picture of the viroid-like RNA universe. Specifically, evidence of viroid-like RNAs in bacteria has been found, which, if confirmed, would lend credence to the viroid-first hypothesis. The other challenges with the viroid-first hypothesis remain to be addressed, but this finding breathes new life into the hypothesis that viroids may be remnants from the earliest stages of life on Earth.

1.8 Conclusion

In this chapter, I have sketched out the contours of the subviral world. Over the course of this thesis, I will fill in the details of this picture.

In Chapter 2, I lay the groundwork for the discovery of new viroids by creating a database of all known viroids and viroid-like RNAs. Surprisingly, no such database

[213] Katrina F. Tjhung et al., “An RNA Polymerase Ribozyme That Synthesizes Its Own Ancestor,” *Proceedings of the National Academy of Sciences* 117, no. 6 (February 2020): 2906–2913, doi:10.1073/pnas.1914282117.

[214] Salehi-Ashtiani and Szostak, “In Vitro Evolution Suggests Multiple Origins for the Hammerhead Ribozyme.”

was accessible when I began my thesis, and so I created one from scratch. This database serves as the foundation for the work in Chapter 4, which describes the results of the a systematic search for viroids and viroid-like RNAs using metatranscriptomic data.

Then, in Chapter 3, I introduce the methodological advances that made the search for viroids at scale possible. Specifically, I describe the development and testing of a new computational pipeline for the analysis of transcriptomic and metatranscriptomic data. This pipeline, called `vdsearch`, forms the basis of the work in Chapter 4.

In Chapter 4, I describe the results of the search for viroids and viroid-like RNAs using `vdsearch` on large-scale metatranscriptomic data. This work represents the first systematic search for viroids and viroid-like RNAs in metatranscriptomic data, and resulted in a vast expansion of the known viroid and viroid-like RNA universe.

Then, in Chapter 5, I discuss the development of software for the analysis of circular RNAs with far greater performance than existing tools. This software, called circKit, was developed to ameliorate the performance issues encountered during the search for viroids and viroid-like RNAs in Chapter 4. However, it is a general-purpose tool that can be used for the analysis of any circular sequence data that I hope will be useful to the broader scientific community.

Finally, in Chapter 6, the summary and implications of this work as a whole are discussed. The future direction of viroid and viroid-like RNA biology is discussed in the context of this work.

2

A database of viroids and viroid-like circular RNAs

Summary

A database of viroid and viroid-like circular RNAs is presented. The database, ViroidDB, contains almost 10,000 sequences and provides a comprehensive resource for the analysis and classification of these agents. In addition to sequence data, the database also includes secondary structure predictions, ribozyme predictions, and sequence clusters. This database is freely accessible at https://viroids.org and serves as the basis of the analyses presented later in this thesis.

2.1 Introduction

At the inception of this thesis, there were no databases that would cover the panoply of the viroid-like agents. In fact, there were no databases that collated information purely about viroids, let alone the greater viroid-like RNA world. Previously developed databases[1] were out of date, not maintained, and inaccessible.

[1] D. A. Lafontaine et al., "The Viroid and Viroid-Like RNA Database," *Nucleic Acids Research* 26, no. 1 (January 1998): 190–191, doi:10.1093/nar/26.1.190; Rocheleau and Pelchat, "The Subviral RNA Database."

A complete database was a necessity for an efficient study of these increasingly diverse agents.

For example, performing read mapping to a reference database is a common step in the analysis of high-throughput sequencing data. Indeed, it was for just such a purpose that the database was initially developed as an *ad hoc* internal resource. Over time, the database grew to include more and more sequences and analyses, and it became clear that it was a valuable resource for the community at large. It was also evident that versioning and sharing the database would be beneficial to the community, as it would allow for reproducibility and the ability to track changes over time. Thus, ViroidDB was developed to provide the community with a consistent, usable, and up-to-date resource for the analysis and classification of viroid-like RNA agents.

ViroidDB is in fact several sub-projects and it is important to distinguish between them. First and foremost, ViroidDB is a large amount of data (*i.e.,* sequences, metadata, derived data, *etc.*), discussed in Section 2.2. Second, the database is a collection of serverless data storage systems that enable ViroidDB's data and website to be maintained in a sustainable manner, as discussed in Section 2.3. Third, ViroidDB has an interactive web interface that allows users to explore the data, search for sequences, and download the data in bulk, as discussed in Section 2.4.

2.2 Database content

ViroidDB is a comprehensive database of viroids and viroid-like RNA agents, comprising almost 10,000 genomic sequences. ViroidDB draws all primary sequence data from the National Center for Biotechnology (NCBI) GenBank[2] database. A custom pipeline then produces multiple downstream analyses that are incorporated into the database (Figure 2.1).

[2]Eric W Sayers et al., "GenBank," *Nucleic Acids Research* 49, no. D1 (January 2021): D92–D96, doi:10.1093/nar/gkaa1023.

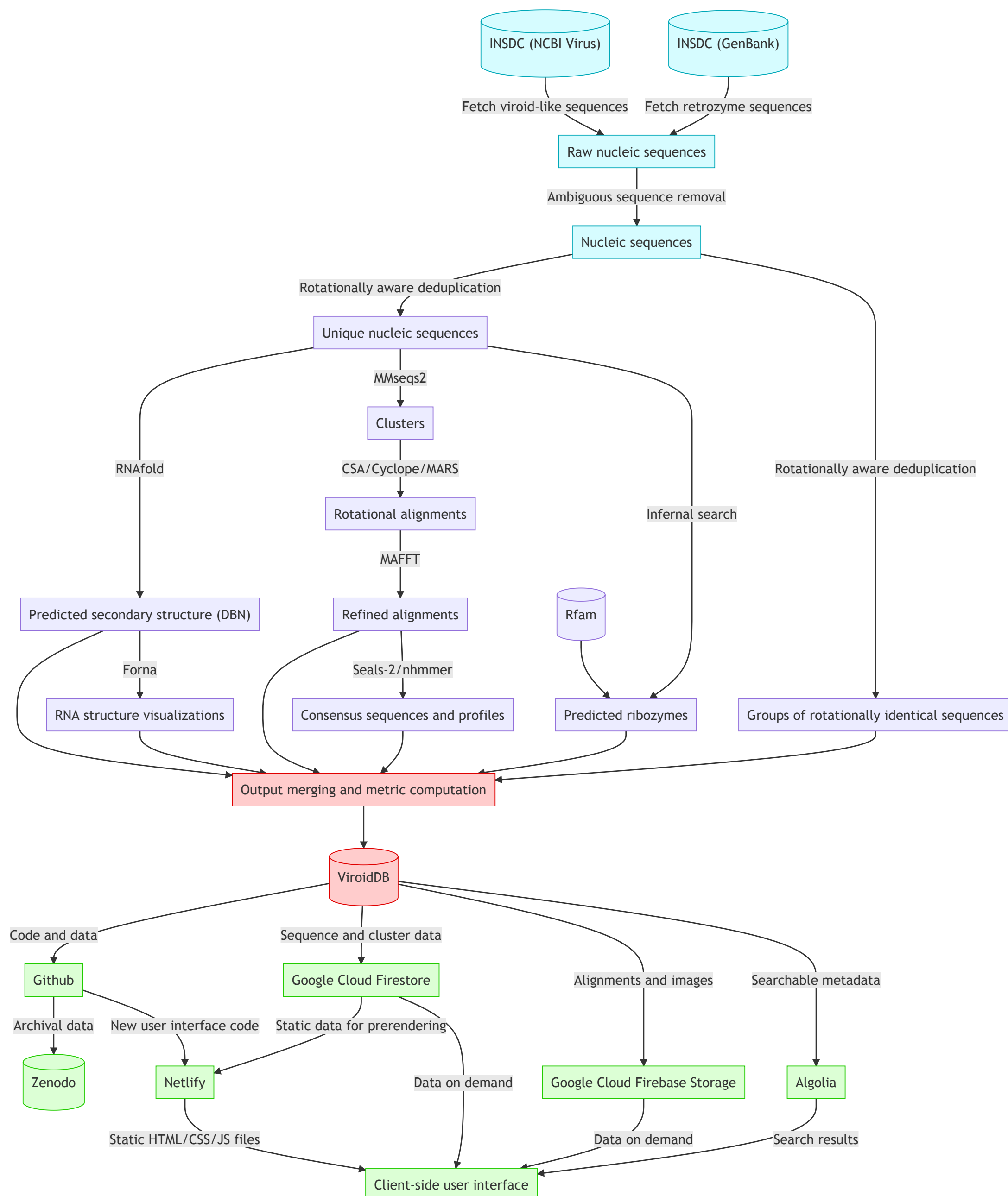


Figure 2.1: A flowchart of the complete data processing and release pipeline. Blue steps denote data fetching and pre-processing procedures while purple steps denote the analysis phase. Red steps correspond to post-processing organisational steps. The data release pipeline (green) is initiated upon every change to the underlying data or code ensures that the data are archived and that the website is up to date and efficient.

2.2.1 Sequence data

Initially, all sequences marked as complete and associated with viroid-like taxa (retrozymes excepted) were downloaded from the NCBI Virus resource[3] and formatted consistently. Retrozymes, which are not taxonomically assigned, were identified by searching for the term "retrozyme" in GenBank. All available metadata for these sequences were also downloaded.

Deduplication is first performed with respect not only to the exact sequences themselves but also to all rotations in both polarities. To this end, a canonical representation of the sequence is selected using the lexicographically minimal rotation (both polarities included). All subsequent sequences with the same canonical rotation are grouped together. From each group of rotationally identical sequences, a single reference sequence is chosen arbitrarily if no RefSeq-derived sequence[4] is a member. The other rotationally identical sequences are discarded and their identifiers are noted on the record of the reference sequence. In addition, sequences with ambiguous nucleotides are removed in this phase of the pipeline.

The current version of ViroidDB contains 9,691 sequences, of which the vast majority (9,353) are from viroids. Among viroids, peach latent mosaic viroid is the most abundantly represented, with 4,891 samples. The vast amount of these nearly identical sequences are the result of deep sequencing studies into the heterogeneity of quasispecies during infection.[5] Ribozyviruses comprise the next most populous sequence type (n=243), followed by retrozymes (n=73), retroviroids (n=12), and satRNAs (n=10). A more detailed breakdown of the sequence content by genus is given in Table 2.1.

---

[3]Eneida L. Hatcher et al., "Virus Variation Resource – Improved Response to Emergent Viral Outbreaks," *Nucleic Acids Research* 45, no. D1 (January 2017): D482–D490, doi:10.1093/nar/gkw1065.

[4]Wenjun Li et al., "RefSeq: Expanding the Prokaryotic Genome Annotation Pipeline Reach with Protein Family Model Curation," *Nucleic Acids Research* 49, no. D1 (January 2021): D1020–D1028, doi:10.1093/nar/gkaa1105.

[5]Jean-Pierre Sehi Glouzon et al., "Deep-Sequencing of the Peach Latent Mosaic Viroid Reveals New Aspects of Population Heterogeneity," ed. Yury E. Khudyakov, *PLoS ONE* 9, no. 1 (January 2014): e87297, doi:10.1371/journal.pone.0087297.

Table 2.1: The genera and number of sequences in the 2021-09-07 release of ViroidDB.

Genus	Count
Pelamoviroid	5,080
Pospiviroid	2,363
Apscaviroid	939
Hostuviroid	488
Deltavirus	243
Cocadviroid	184
Elaviroid	102
Avsunviroid	60
Coleviroid	30

2.2.2 Secondary structures

The secondary structure for both polarities of each sequence are predicted using RNAfold.[6] Additional metrics (*e.g.* percent of bases folded) are also computed and consolidated with visualisations generated by Forna.[7] In the visualisations, each base is colored depending on its base-pairing interaction (such as stems, hairpins, interior loops) for easier examination (shown on the website in Figure 2.5).

2.2.3 Predicted ribozymes

Beyond the secondary structure, all sequences were scanned for signature motifs of auto-catalytic ribozymes because many viroid-like RNAs are known to employ self-cleaving ribozymes for replication.[8] This scan was performed using the Infernal software,[9] against selected ribozyme families from the Rfam database.[10] All detected results with E values less than 0.1 are considered significant and reported. The database stores the location, type, E value, alignment, and truncation status of each hit. Altogether, the database currently reports 11,713 putative ribozymes. The large number of ribozymes is a result of the database containing so many agents

[6] Lorenz et al., "ViennaRNA Package 2.0."

[7] Peter Kerpedjiev, Stefan Hammer, and Ivo L. Hofacker, "Forna (Force-Directed RNA): Simple and Effective Online RNA Secondary Structure Diagrams," *Bioinformatics* 31, no. 20 (October 2015): 3377–3379, doi:10.1093/bioinformatics/btv372.

[8] Flores, Minoia, et al., "Viroid Replication."

[9] E. P. Nawrocki and S. R. Eddy, "Infernal 1.1: 100-Fold Faster RNA Homology Searches," *Bioinformatics* 29, no. 22 (November 2013): 2933–2935, doi:10.1093/bioinformatics/btt509.

[10] Kalvari et al., "Rfam 14."

with symmetric rolling circle replication mechanisms which require one ribozyme per polarity combined with the E value threshold allowing alternative topologies for a given motif to be reported as well.[11]

2.2.4 Sequence clusters

A custom, standalone pipeline was devised to cluster the sequences gathered from the reference database and generate multiple sequence alignments (MSAs) (Figure 2.1). Although these clusters are not intended to reflect taxonomic ranks, they greatly reduce the number of sequences in a consistent manner, streamlining advanced analyses, such as the identification of translocating motifs or conserved motif ordering. Initially, all deduplicated sequences are clustered using MMseqs2[12] at multiple identity thresholds: 75%, 80%, 85%, 90%, 95% with a fixed minimum alignment coverage of 75% for both the query and the target sequences. Then, rotational alignment (in which the optimal starting position is identified) is attempted within every cluster using CSA.[13] This step is necessary as similar sequences may have spuriously low alignment scores when the sequences are not oriented with consistent origins. CSA was unable to produce meaningful output for all clusters for several technical reasons (namely, built-in limitation preventing handling inputs with more than 64 sequences). If CSA fails to produce an output, cyclic comparison is attempted using Cyclope.[14] If Cyclope is also unsuccessful in producing an output, rotation is attempted using MARS.[15]

[11] While it would have been possible to filter the ribozyme predictions to only include the most significant hit for each sequence per polarity, this was not done to prevent presenting potentially misleading data to the user. If a sequence has multiple significant hits, the one with the lowest E value may not be the biologically correct one due to the location of the truncation of the circular sequence. Circularity-aware ribozyme prediction is not yet implemented by any tools.

[12] Martin Steinegger and Johannes Söding, "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets," *Nature Biotechnology* 35, no. 11 (November 2017): 1026–1028, doi:10.1038/nbt.3988.

[13] Francisco Fernandes, Luísa Pereira, and Ana T Freitas, "CSA: An Efficient Algorithm to Improve Circular DNA Multiple Alignment," *BMC Bioinformatics* 10, no. 1 (2009): 230, doi:10.1186/1471-2105-10-230.

[14] Axel Mosig, Ivo Hofacker, and Peter F Stadler, "Comparative Analysis of Cyclic Sequences: Viroids and Other Small Circular RNAs," in *German Conference on Bioinformatics*, vol. 83, 2006, 93–102.

[15] Lorraine A. K. Ayad and Solon P. Pissis, "MARS: Improving Multiple Circular Sequence Alignment Using Refined Sequences," *BMC Genomics* 18, no. 1 (December 2017): 86,

The rotated clusters were then re-aligned using MAFFT (v7.475) using its automatic mode.[16] The minority of the clusters, for which no cyclic comparison software successfully produced meaningful output, were forwarded directly to MAFFT, without rotational modifications. For each resulting MSA (one per cluster), a consensus sequence was generated using Seals-2 (D. Roland Walker and Eugene V. Koonin[17], available at https://github.com/YuriWolf-ncbi/seals-2). The consensus sequence was then inserted into the alignment as the first sequence in the multi-FASTA file to act as a 'master sequence.' This final MSA was used to construct an nhmm profile using the HMMER suite v.3.3.2.[18]

The number of clusters at each identity threshold is given in Table 2.2. Methodological improvements to circular sequence clustering are discussed in Section 3.3.6. These improvements will be incorporated into the ViroidDB pipeline in the future. The present architecture makes this less straightforward than it could be, but it is expected that the pipeline will be updated to use the vdsearch pipeline in the future (see Section 2.5 for more details).

Table 2.2: Cluster counts for varying identity thresholds (75% minimum reciprocal overlap required)

Threshold	Clusters
70%	226
75%	224
80%	252
85%	311
90%	458
95%	958

doi:10.1186/s12864-016-3477-5.

[16]K. Katoh and D. M. Standley, "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability," *Molecular Biology and Evolution* 30, no. 4 (April 2013): 772–780, doi:10.1093/molbev/mst010.

[17]"SEALS: A System for Easy Analysis of Lots of Sequences," in *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology* (AAAI Press, 1997), 333–339.

[18]Travis J. Wheeler and Sean R. Eddy, "Nhmmer: DNA Homology Search with Profile HMMs," *Bioinformatics (Oxford, England)* 29, no. 19 (October 2013): 2487–2489, doi:10.1093/bioinformatics/btt403.

2.3 Data storage and infrastructure

To ensure the continuous functionality of ViroidDB, all code (both for processing and the user interface) and data are permanently archived externally in CERN's Zenodo database (https://zenodo.org). All software is available on GitHub under the open-source MIT license. The data is versioned by release date and releases are immutable. No special identifiers are presently assigned for sequences in ViroidDB; the existing accessions are used for simplicity. Because new data could change the cluster assignments, clusters are given a date- and threshold-dependent identifier.

The live version of the website was optimized for speed, scalability, and sustainability. At a high level, ViroidDB's web infrastructure consists of a browser-based application executed locally in JavaScript (albeit written in its strongly typed derivative, TypeScript) and served as static files via the Netlify content delivery network (CDN) (https://netlify.com) supported by a serverless database, file storage, and search services. The front-end of the browser application is built using the Vue (https://vuejs.org) and Nuxt (https://nuxtjs.org) frameworks, which are configured to automatically update the site upon every accepted change in the version-controlled repository hosted on GitHub. When a user loads the website, they download pre-generated HTML, CSS, and JavaScript files stored entirely on the CDN's servers, which in turn fetch only the data needed to display the given page. Although the user interface relies heavily on dynamic JavaScript to retrieve the content being displayed, the compilation process is able to prefetch and statically render pages' default states (such as the truncated list of members of the family *Avsunviroidae*). This process completely eliminates the waiting period the user experiences after the page's scripts have loaded but before the data from the database has been loaded by the scripts, which in turn use it to render the document object model (DOM). It also reduces the number of database queries, which ensures that there is sufficient capacity for numerous users with minimal to no expenses resulting from operating the service. This compilation process is also

repeated automatically upon every change to the codebase, ensuring that the site is always up to date and efficient.

Importantly, no servers must be maintained specifically for ViroidDB, thereby easing the development and maintenance burden. The downloaded client-side code simply queries and displays data stored in the database and file storage layer, which is built upon Google Cloud's Firebase platform (https://firebase.google.com/). Because the platform provides a free tier of service with a 50,000 query daily limit, the site is able to operate with no expected operating costs due to the database. Small metadata, such as records associating the accessions of sequences with their identified ribozymes, are stored in a serverless NoSQL database for fast retrieval while larger data, such as full-length alignments of clusters and images, are placed in an object store.

2.4 Data query and access

The ViroidDB database supports the exploration and use of its data both via a graphical web interface and via downloadable raw data.

2.4.1 ViroidDB website

ViroidDB offers an interactive website that enables users to inspect, visualize, search, and download all the information stored in the database.

The site is organized around the sequence display page, with other pages of sequence collections linked to each sequence's page. These pages contain all the sequence-specific metadata, organized into panels. The first row of panels displays basic information such as length and GC-content, taxonomy information, and sample collection information (Figure 2.2). Also noted are any rotationally identical sequences. It also contains quick-access buttons to download or copy the sequence in FASTA format.

The next row of panels shows the sequence's nucleotide sequence in FASTA format and resource links (Figure 2.3). Users can toggle the nucleotide sequence to display

Hepatitis delta virus isolate QD-22

Copy FASTA

Download FASTA

Sequence information

Basic information about the sequence

Accession MK890232.1

Length 1679 bp

GC Content 59.0%

Release Date 2019-12-09

Database GenBank

Rotationally Identical None found

Collection information

Provided by the depositors

Submitters Naz,N., Qureshi,H., Khan,O.Y.

Host *Homo sapiens*

Source Blood

Location Pakistan

Collection Date 2018-09

Taxonomy

According to NCBI Taxonomy

Type Virus

Family *Kolmioviridae*

Genus *Deltavirus*

Species *Hepatitis delta virus*

Figure 2.2: Metadata panels on the sequence detail page

Nucleotide sequence

FASTA-formatted sequence data

Reverse complement

Display the reverse complement of the sequence

Secondary structure

Include the secondary structure of the sequence in dot-bracket notation

RNA

Display RNA bases instead of DNA (i.e. U instead of T)

```
>MK890232.1 Hepatitis delta virus isolate QD-22, complete genome [ViroidDB]
ATGAGCCACAGTTCCGAACGAGGAGACGCGGGTGGAAGGGATCAGCGCCCGAGAGGGAGTGGACGGTAAAGAGCATTGGA
ACGCCGAGGGGATGCCTCCCAAGAAGGAAAAAAGAGGTATCAAAAGCGGACGTGGTCCCCACGACGCTGGAGACATCCTG
GAAGGGGAAAGAGGATGGGTGGAAAAGAAGGGGCGGGCCTCCCGATCCGAGGGGCCCAACGCCAGCTTTGGGAGGCACTC
CGGCCCGAAGGGTTGAGAAGCACCCAGAGGGAGGAATCCACTCGGAGTAGAACAGAGAAATCACCTCCAGAGGACCCCTT
CAGCGAACGAAAGGGCGCTGAGCCGCGAGGAGTAAGACCATAGCGATGGGTGGAGATGCTAGGAGTTAGAGGAGACCGAA
GCGAGGAGGAAAGCAAAGAAAGCAACGGGGCTAGTCGGCAGGTGTTCCGCCCCCCGAGAGGGGACGAGTGAGGCTTATCC
CGGAGCATTCGACTTATCGTCCCCACATAGCAGAATCCCGGACCCCCTCCAAAAAGTGACCGAGGGAGGCGCTAGGATCG
CGGGGGAAACCGGTGGAACCATGGGATGCTCCTCCCGATTCCGTCCATCCCTCCTCCCAGGGGGTCGCCCAGGAATGGCG
GGACCCCACTTCATCAGGGTCCGCGTTCCATCCTTTCTTACCTGATGGCCGGCATGGTCCCAGCCTCCTCGCTGGCGCCG
GCTGGGCAACATTCCGAGGGGACCGTCCCTCGGTAATGGCGAATGGGACCCAGAAATCTCTCTAGATTCCCAGAAGGAAT
CGAGAGAAAACTGGCTCTCCCTTAGCCATCCGAGTGGACGTCTGTCCTCCTTCGGATGCCCAGGTCGGACCGCGAGGAGG
TGGAGATGCCATGCCGACCCGAAGAGGAAAGAAGGACACGGACGCGAACCTGTGAGTGGAAACCCGCTTTATTCACTGGG
GTCGACAACTCTGAGGAGAAAAGGGAGGATCGGATGGGAAGAGTAGATCCTATGGGAATCCCCGTTCTCCCCTCGTGTCC
AGCCCGTCCCCGGTCCGATGGAAGGGGGACTCCGGGACTCCTTGCATGCTGGGGACGAAGCCGCCTCCGGGCGCTCCCCT
CGGTCCACCTTCGAGGGGGTTCACACCTCCAGTCGACGGGCCGGCTACTCTTCTTTCCCTTCTCTCGTCTTCCTCGGTCA
ACCTCTTGAGTTCCTCTTCTTCCTCCCTGCTGAGGCTCTTCCCTCCCGCGCTCAACTGCCTCTTTTTGTTCTCGAGGGCC
TTCCTTCTTCGGTGATCCTGCCTCTCCTTGTCGGTGAATCCTCCCCTGAGAGGCCTCTTCCTAGGTCCGGAGTCTACTTC
CATCTGGTCTGTTCGGGCCCTCTTCGCCGGGGGAGCCCCCTCTCCATCCTTATCCTTCTTTCCGAGAATTCCTTTGATGT
TTCCCAGCCAGGGATTTTCGTCCTCTAGCTTTTTGATTTTCTTCTTAACCTTCCGGAGATCCCTCTCGAGTTCCTCTAAC
TTCTTTCTTCCGTTCACCCACTGCTCGAGAACCTCTTCTCTTCCTCCGCGGTTCTTCCTCGATTCGGACCGGCTCATCTC
GACAAGAGGCGGCAGTCCTCAGTACTCTTACTCTTTTCTGTAAAGAGGAGACTGCTGGACTCGCCGCCCAAGTTCGAGC
```

Links

External resources for this sequence

NCBI Nucleotide

Taxonomy

PubMed

Figure 2.3: Sequence viewer on the sequence detail page

the reverse complement, textual secondary structure, and/or convert between DNA and RNA bases. Users can also copy or download the nucleotide sequence in FASTA format through built-in buttons. Links to external resources, such as associated PubMed articles and GenBank pages, are included as well.

Ribozymes

Infernal search results for known self-cleaving ribozymes

```
Description: Hepatitis delta virus isolate QD-22, complete genome
Hit scores:
 rank     E-value  score  bias  modelname          start    end   mdl trunc   gc  description
 ----   --------- ------ -----  ----------------- ------ ------   --- ----- ----  -----------
  (1) !   4.5e-15   92.0   0.0  HDV_ribozyme         902    810 -  cm    no 0.65  Hepatitis delta virus ribozyme
  (2) !   1.1e-11   73.0   0.0  HDV_ribozyme         685    772 +  cm    no 0.68  Hepatitis delta virus ribozyme
 ------ inclusion threshold ------
  (3) ?     0.031   15.2   0.0  HDV-F-prausnitzii    900    815 -  cm    no 0.65  HDV ribozyme from F. prausnitzii
Hit alignments:
>> HDV_ribozyme  Hepatitis delta virus ribozyme
 rank     E-value  score  bias mdl mdl from   mdl to       seq from      seq to       acc trunc   gc
 ----   --------- ------ ----- --- -------- --------    ----------- -----------      ---- ----- ----
  (1) !   4.5e-15   92.0   0.0  cm        1       91 []         902         810 - .. 0.98    no 0.65

                                                                                                    NC
                   ::<<<<<<<------------<<<_______>>>>>>>>>>---<<<<<<<<-<<<-<.___.>->>>..>>>>>>>>::::::::::::::::: CS
  HDV_ribozyme   1 uugGgcgGgaauGuCAuCuCCgCCUCCuCGcGGccCcgcCuGGGCauuCcGaAGGgGg.ccG.cgcCC..uCgGaauGGCuAAGGGaGaGaCacu 91
                   U+GGG::GG+AUG+CAUCUCC:CCUCCUCGCGG:CC::CCUGGGCAU CCGAAGG:GG C G CG:CC  UCGG AUGGCUAAGGGAGAG+CA U
    MK890232.1 902 UCGGGUCGGCAUGGCAUCUCCACCUCCUCGCGGUCCGACCUGGGCAU-CCGAAGGAGGaCAGaCGUCCacUCGG-AUGGCUAAGGGAGAGCCAGU 810
                   ***********************************************.**********76669***********.******************** PP
```

Figure 2.4: Ribozyme predictions on the sequence detail page

The following panels displays alignments for any predicted self-cleaving ribozymes (Figure 2.4). These alignments are presented as the raw output from Infernal. Future versions of the site will include this scrollable panel as well as parsed metadata from the tabular output of Infernal, formatted similarly to the sequence metadata panels.

Below the ribozyme predictions, the secondary structure of the sequence is displayed (Figure 2.5). Both polarities are shown side by side, allowing for easy comparison. Minimum free energy and percent of bases folded are also displayed.

The sequences are listed on a sequence index page where the user can opt to only have sequences of a given type displayed for easier browsing.

Clusters are treated analogously to sequences. A cluster index page lists each cluster and its reference sequence as well as the number of members (Figure 2.6). Clicking into a cluster page displays information about a single cluster (Figure 2.7).

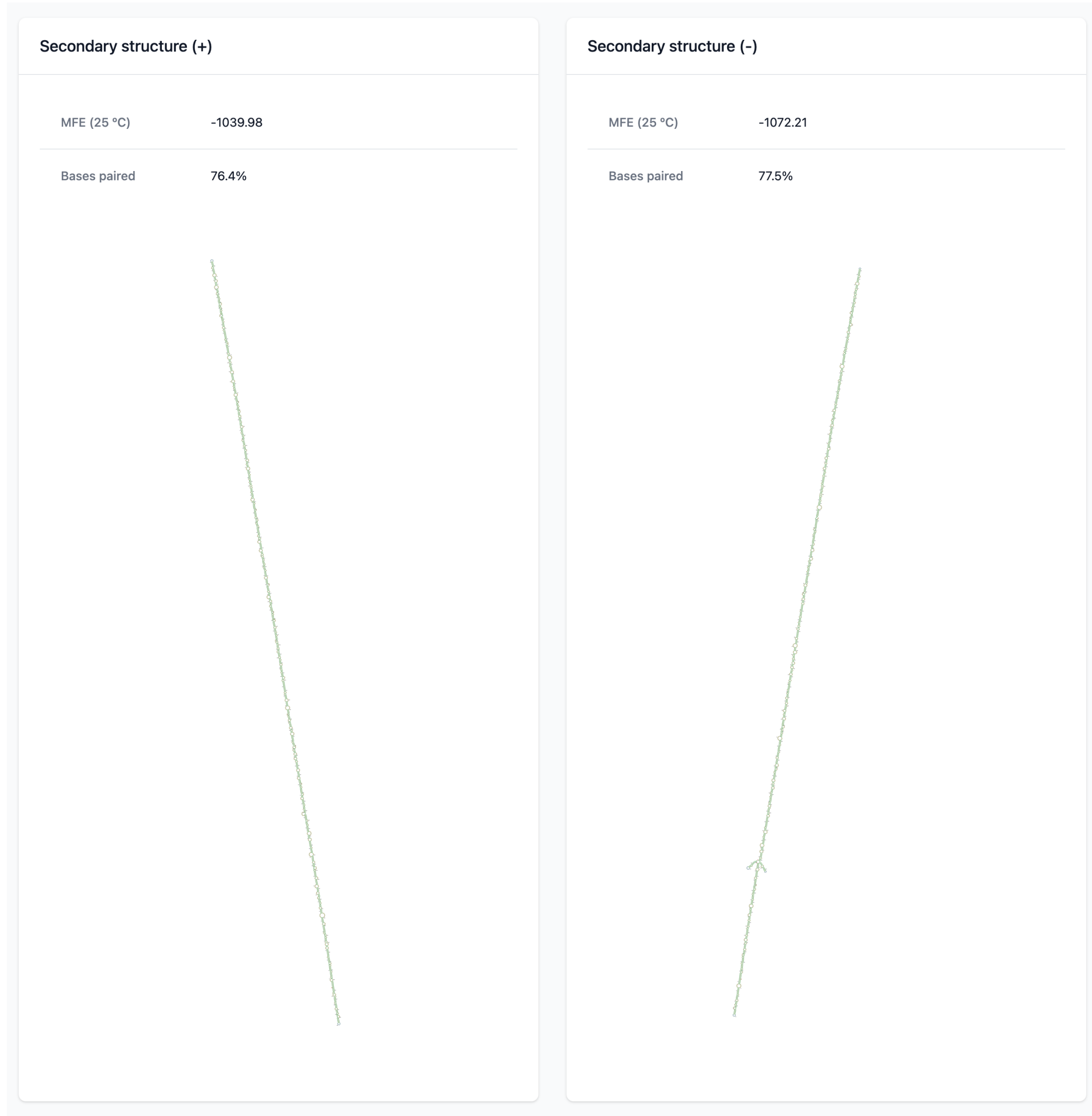


Figure 2.5: Secondary structures and predictions on the sequence detail page

Currently, this page is only available for clusters from the current release, but future versions of the site will allow users to view clusters from previous releases as well.

The first panel shows multiple sequence alignments generated from the reference sequence after rotational alignment with the resultant nhmm profile. Visualisations of the alignments are immediately below. These visualisations, which are produced by CSA, are not always able to be displayed to the user (see Section 2.2.4 for an explanation). If the visualisations are not available, the user is instead presented with a link to an explanation of the reason for the absence. The final panel is a list

ViroidDB About Sequences Clusters Search Download Releases Submit Community

Clusters

As with other biological entities, viroid-like RNAs form groups of related sequences, which most likely originate from a relatively recent common ancestor. The clusters below, and their derived information, are the results of a custom, automatic pipeline we wrote to consistently process all sequences. Note that these clusters are not necessarily representative of the refined taxonomy of the sequences they contain but are instead optimized to make it easier to examine groups of (potentially) related sequences. To browse sequences based on their taxonomic classification, use the search page.

Learn more about clustering →

Identity Threshold

90%

Viroid species demarcation historically used 90% as a cutoff.

CLUSTER ID	REFERENCE SEQUENCE	MEMBERS
2021-09-07-ID90-1	Hop stunt viroid RNA, isolate:hAIw36-5	2
2021-09-07-ID90-10	Chrysanthemum stunt viroid genomic RNA, strain: SK2	1
2021-09-07-ID90-100	Pear blister canker viroid clone PBCVd-tun8	45
2021-09-07-ID90-101	Hepatitis D Virus genotype 3, clone 010-OBCCI2	5
2021-09-07-ID90-102	Mexican papita viroid, isolate 12/0463/VI.04	5
2021-09-07-ID90-103	Columnea latent viroid isolate 12 clone 2	74
2021-09-07-ID90-104	Australian grapevine viroid isolate F3	2
2021-09-07-ID90-105	Hepatitis delta virus isolate 6	1

Figure 2.6: Cluster listing page of ViroidDB

of all sequences that are contained within the cluster. Users can jump directly to each sequence's individual page from this list.

A search page (Figure 2.8) is available to enable users to find sequences of interest. The Algolia service (https://algolia.com) is used to provide typo-tolerant approximate searching of all sequences and associated metadata in the database. Searchable fields include the accession, sequence name, full taxonomy, authors, geographic location, host, and isolation site. This integration allows the user to instantly (<10 ms) perform full-text search with optional type (*e.g.* viroid), family, genus, and species filters. If a match to the input query is found, the exact part of the record for which the match is found is highlighted in the results. As the user types a query into the search area, the number of results in each taxon is displayed dynamically.

Because this site is designed to serve as a resource for the community, it also hosts a community section of the website. This section enables members of the community to post upcoming events and notices of interest to a virtual bulletin

ViroidDB About Sequences Clusters Search Download Submit Community

2021-06-06-ID80-10

The alignment below was constructed using MAFFT's automatic mode and overlayed with a consensus on top to act as the master sequence. Whenever possible, we first performed a cyclic rotation of all cluster members. Learn more →

Alignment

Built with MAFFT, Cyclope, and CSA

```
>Cls_ID80_10_con
GCACCACAATGCCAATGCCGAAGGACGAAGGGAGGTTTCCCCATGGACCGCTCAACAATGCTCCGCGCCGGACTGTAAGATTACCTCCCCGCAGCACCTC
>BK009950.1
GCACCACAATGCCAATGCCGAAGGACGAAGGGAGGTTTCCCCATGGACCGCTCAACAATGCTCCGCGCCGGACTGTAAGATTACCTCCCCGCAGCACCTC
>BK009949.1
GCACCACAATGCCAATACCGAAGGATGAAGGGAGGTTTCCCCATGGACCGCTCAACAATGCTCCGCGCCGGACTGTAAGATTACCTCCCCGCAGCACCTC
>BK009948.1
GCACCACAATGCCAATACCGAAGGATGAAGGGAGGTTTCCCCATGGACCGCTCAACAATGCTCCGCGCCGGACTGTAAGATTACCTCCCCGCAGCACCTC
>BK009947.1
GCACCA------CAATGCCGAAGGACGAAGGGAGGTTTCCCCATGGACCGCTCAACAATGCTCCGCGCCGGACTGTAAGATTACCTCCCCGCAGCACCTC
>BK009945.1
GCACCACAATGCCAATGTCGAAGGTCGAAGGGAGGTTTCCCCATGGACCGCTCAACAATGCTCCGCGCCGGACTGTAAGATTATCTCCCCGCAGCACCTC
```

Circular Alignment

Generated by CSA

1416 177 354 531 708 885 1062 1239

BK009950.1 @ 420
BK009949.1 @ 424
BK009948.1 @ 335
BK009947.1 @ 311
BK009945.1 @ 179

Conservation
GapFrequency

Blocks

Which aligned sequence regions are where

0 50 100 150 200 250 300 350 400 450 500 550 600 650 700 750 800 850 900 950 1000

NAME	ACCESSION	BASES	TYPE	FAMILY	GENUS
TPA_inf: Fragaria x ananassa retrotransposon retrozyme Rtz_Fa2	BK009945.1	789	Retrozyme		

Figure 2.7: Cluster detail page

Figure 2.8: The search page of ViroidDB demonstrates the site's fuzzy search, search-match highlighting, taxonomic refinement lists, and species sub-search

board. Additionally, email forwarding addresses using the viroids.org domain name are available upon demand at no cost.[19] Finally, the site's community page provides access to an internet chat room to facilitate collaboration.

2.4.2 Data portal

In addition to facilitating the exploration of individual sequences and clusters, ViroidDB is also meant to enable computationally oriented researchers the ability to rapidly access high-quality data on viroid-like RNAs. To do so, the website also includes a data portal enabling users to download the cleaned sequence data and downstream analysis results in bulk form. The data download portal includes sequence data in FASTA format, RNA structure data in dot-bracket notation format, tabular and semi-structured output from Infernal, PDF documents containing

[19] I am quite proud of my viroids.org domain.

structure visualisations, bitmap visualisations and alignments for each cluster, and a JavaScript Object Notation (JSON) file containing metadata. These files can be readily used for subsequent bioinformatic research.

2.5 Future directions

Since its inception, viroids.org has been accessed 2,600 times by users from around the world. The most used site on the page (save the homepage) is the download page, which has been accessed over 200 times. These numbers are expected to increase as the site becomes more widely known and as viroid-like agents are discovered more frequently.

Presently, the raw data ingestion and processed data deployment stages are performed manually. Although not particularly cumbersome (taking less than ten minutes each), relying on human intervention to create new database releases limits the frequency of releases. Both of these stages can be automated: the manual effort is limited in scope to data transfer rather than analysis. As the pace of viroid-like agent discovery increases, the ability to perform fully automatic updates will become critical. While the database is currently updated on an *ad-hoc* basis, it will be optimal to create regularly scheduled releases in the future.

Additionally, while all ViroidDB application and pipeline software is available under a permissive open-source license, not all of the software upon which the site depends is similarly licensed. Ideally, upgrades to the site will reduce the amount of non-open dependencies, thereby ensuring that the interactive aspect of the site will remain usable in the future even if the data infrastructure is no longer available. Algolia and Firebase are both proprietary software and therefore less trustworthy as long-term dependencies. Replacing these dependencies with Elasticsearch and a self-hosted database such as PostgreSQL, respectively, would be a step in the right direction. Indeed, a future rewrite of the website in a more modern full-stack framework such as RedwoodJS (https://redwoodjs.com) would make the use of open-source software throughout the stack easier to implement. Currently, the

pipeline and user interface source is available under the permissive MIT license at https://github.com/Benjamin-Lee/viroiddb and is archived at Zenodo (https://doi.org/10.5281/zenodo.5202945).

It is hoped that, eventually, the ViroidDB creation pipeline will be a part of the vdsearch pipeline introduced in Chapter 3. This design would have several advantages. It would allow for the use of the same codebase for both the database and the search tool, thereby unifying two similar but distinct pipelines. Doing so would therefore make it easier take advantage of methodological improvements (*e.g.*, enhanced clustering or search methods) to improve the database. Because vdsearch is already modular, these improvements could be translated to the ViroidDB pipeline with minimal effort. Building the ViroidDB creation pipeline into vdsearch would also allow users to readily update their local databases with the latest data, regardless of whether the public database at viroids.org is updated.

2.6 Conclusion

ViroidDB is the only currently available database of viroid-like RNAs. In addition to a comprehensive collection of sequences, it provides users with a wealth of information on these agents as well as analytical tools, making it a platform for bioinformatic studies. Recent metatranscriptome mining efforts have led to a dramatic expansion of the known diversity of RNA viruses. As is described elsewhere in this thesis (Chapter 4), the viroid and viroid-like RNA world is no exception to this trend. A catalogue of viroid-like agents is therefore essential for the study of these agents, and ViroidDB, which is expected to be regularly updated, should become an indispensable resource for researchers studying this remarkable class of mobile elements. ViroidDB is freely accessible at https://viroids.org.

3

A method for identifying viroids and viroid-like RNAs

Summary

Viroids have not been systematically sought in high-throughput sequencing data, which has been a major limitation in the discovery of new viroids and viroid-like RNAs. To address this, I developed a pipeline, `vdsearch`, that is capable of identifying viroid-like RNAs in both transcriptomes and metatranscriptomes after assembly. Relying on either direct sequence similarity between putative cccRNAs and known viroids or the presence of self-cleaving ribozymes, the pipeline is capable of identifying viroid-like RNAs in both transcriptomes and metatranscriptomes. It was validated by application to a large plant transcriptomic data set, in which it was able to identify both known viroids and a novel viroid-like RNA.

3.1 Introduction

My primary methodological advance has been the development of an integrated, scalable computational pipeline for the discovery and analysis of viroids and viroid-like cccRNAs directly from assembled transcriptomes and metatranscriptomes (Fig-

ure 4.1). The pipeline, which I have named `vdsearch`, starts with the reference-free and *de novo* identification of cccRNAs or RCR intermediates, capitalizing on the fact that assemblies of both complete circular monomers and multimeric linear intermediates contain head-to-tail repeats.[1] The identified sequences are then cleaved *in silico* to unit length and deduplicated, taking circularity into account. Starting from the set of detected cccRNAs, the pipeline performs both alignment-free and alignment-based searches. The primary approach for the identification of viroid-like agents among the cccRNAs is the prediction of self-cleaving ribozymes using RNA sequence and secondary structure covariance models.[2] Assuming that the diversity of ribozymes in viroid-like RNAs could be greater than so far uncovered, a database of known self-cleaving ribozyme models was curated from Rfam,[3] of which only a minority were detected in viroids and viroid-like RNAs. This model database was supplemented with the pospiviroid RY motif[4] to enable detection of potential pospiviroids, which lack ribozymes. The pipeline also performs direct sequence similarity searches against reference databases such as ViroidDB.[5]

Ribozyme-containing cccRNA sequences were classified as symmetric or asymmetric depending on whether they contained predicted ribozymes in both or only one RNA polarity, respectively, reflecting the RCR mode these cccRNAs are likely to undergo. However, it cannot be ruled out that some apparently asymmetric cccRNAs actually contain a second ribozyme distinct from the currently known ones.

---

[1]Yangmei Qin et al., "Reference-Free and de Novo Identification of Circular RNAs," preprint, April 2020, doi:10.1101/2020.04.21.050617.

[2]Nawrocki and Eddy, "Infernal 1.1."

[3]Kalvari et al., "Rfam 14."

[4]M. Gozmanova, "Characterization of the RNA Motif Responsible for the Specific Interaction of Potato Spindle Tuber Viroid RNA (PSTVd) and the Tomato Protein Virp1," *Nucleic Acids Research* 31, no. 19 (October 2003): 5534–5543, doi:10.1093/nar/gkg777.

[5]Lee et al., "ViroidDB."

3.2 Prior work

The *de novo* discovery of viroid-like RNA agents *in silico* was first demonstrated by Qingfa Wu et al.[6], who analysed RNA sequencing data from a peach tree infected with peach latent mosaic viroid (PLMVd) and a grapevine infected with grapevine yellow speckle viroid (GYSVd) and hop stunt viroid (HSVd). Their method did not involve searching for homologous viroid sequences within the data but rather introduced an algorithm that progressively filtered the small RNAs (sRNAs) created by RNA interference present in the input dataset to include only sRNAs which overlap on both ends. This method allows circular sequences to be identified without needing to resort to homology-based searches. Further, this algorithm is capable of detecting viroid-derived sRNAs derived from both monomeric circular or multimeric linear intermediates. The authors were able to correctly assembly PLMVd in the peach sample as well as GYSVd and HSVd in the grapevine sample. Surprisingly, when they analysed the grapevine data, they also found a new viroid-like circular RNA with no significant sequence homology to known viroids that they named grapevine hammerhead viroid-like RNA (GHVd). Though GHVd's transmissibility–and therefore status as a *bona fide* viroid–has yet to be demonstrated *in vivo*, later work by Zhixiang Zhang et al.[7] using the same method identified two more viroid-like RNAs, one of which, Grapevine latent viroid (GLVd), displayed infectivity in grapevine seedlings. Later, the other viroid-like RNA was shown by Pedro Serra et al.[8] to be a true viroid capable of autonomous replication and was designated Apple hammerhead viroid (AHVd). More recently, Rachelle Bester, Sophia S. Malan, and Hans J Maree[9] discovered a circular RNA

[6] "Homology-Independent Discovery of Replicating Pathogenic Circular RNAs by Deep Sequencing and a New Computational Algorithm," *Proceedings of the National Academy of Sciences* 109, no. 10 (March 2012): 3938–3943, doi:10.1073/pnas.1117815109.

[7] "Discovery of Replicating Circular RNAs by RNA-Seq and Computational Algorithms," ed. Biao Ding, *PLOS Pathogens* 10, no. 12 (December 2014): e1004553, doi:10.1371/journal.ppat.1004553.

[8] "Apple Hammerhead Viroid-Like RNA Is a Bona Fide Viroid: Autonomous Replication and Structural Features Support Its Inclusion as a New Member in the Genus Pelamoviroid," *Virus Research* 249 (April 2018): 8–15, doi:10.1016/j.virusres.2018.03.001.

[9] "A Plum Marbling Conundrum: Identification of a New Viroid Associated with Marbling and Corky Flesh in Japanese Plums," *Phytopathology* 110, no. 8 (August 2020): 1476–1482, doi:10.1094/PHYTO-12-19-0474-R.

species by sequencing ten plum trees with symptomatic infections of unknown etiology and five healthy control trees. They named the viroid-like RNA plum viroid I (PVd-I), though its ability to autonomously replicate has yet to be shown. Notably, the discovery of PVd-I was done first through sequencing, not via RT-PCR followed by sequencing. A complete review of next-generation sequencing for viroid research with a full list of viroid-like RNA agents identified by NGS was written recently by Marina Barba and Ahmed Hadidi[10].

3.3 Pipeline implementation

```
~/r/vdsearch
BenjaminLee@mbp ~/r/vdsearch (master)> vdsearch --help

Usage: vdsearch [OPTIONS] COMMAND [ARGS]...

Workflows and utilities for finding and analyzing viroid-like circular RNAs.

─ Options ─
--version                  Show the version and exit.
                           [env var: None]
--verbose                  Show verbose output.
                           [env var: None]
--install-completion       Install completion for the current shell.
                           [env var: None]
--show-completion          Show completion for the current shell, to copy it or customize the installation.
                           [env var: None]
--help                     Show this message and exit.

─ Easy Workflows (recommended) ─
easy-search   Search for viroid-like sequences.

─ circRNA detection and handling (advanced) ─
find-circs     Search for circular sequences.
canonicalize   Compute a rotationally canonical representation of each sequence in a FASTA file.
dedup          Removes duplicate canonicalized sequences.

─ Ribozyme detection (advanced) ─
infernal          Run Infernal cmsearch or cmscan for provided covariance matrices.
rnamotif          Run RNAmotif search and format for provided motif descriptions.
ribozyme-filter   Using ribozyme search results, find viroid-like sequences.

─ Analysis Steps (advanced) ─
fold          Predict secondary structures of circRNAs.
orfs          Extract the ORFs from a FASTA file containing circRNAs.
cluster       Cluster circRNAs or ORFs.
ava2cluster   Convert an MMseqs all-vs-all search output file to a cluster file.
summarize     Generate a summary table of the analysis.

─ Dataset Management ─
download-cms        Download the latest covariance matrices for self-cleaving ribozymes.
download-viroiddb   Download the latest ViroidDB dataset.
purge               Purge the covariance matrix and sequence cache of all data.

For more information, please visit https://benjamindlee.com
```

Figure 3.1: The `vdsearch` command-line interface

The pipeline is built primarily in the Python programming language with performance-sensitive components implemented in lower-level programming

[10] "Application of Next-Generation Sequencing Technologies to Viroids," in *Viroids and Satellites* (Elsevier, 2017), 401–412, doi:10.1016/B978-0-12-801498-1.00038-3.

languages. At its core, `vdsearch` is a command-line tool (Figure 3.1) with a modular design. Each step of the pipeline is implemented as a separate subcommand, allowing for easy integration with other tools and workflows. A single "super" command inspired by `mmseqs2`,[11] `easy-search`, is provided to run the entire pipeline end-to-end.

Internally, the tools is built using the `typer` library for the command-line interface generation. Python functions for each module are automatically converted into command-line subcommands. This allows for the pipeline to be easily extended with new functionality and for the command-line interface parsing and documentation to be automatically generated from the code, greatly simplifying development. Aesthetic and functional improvements to the command-line interface are made by the `rich` library, which provides colorful and informative output. Tabular data parsing and manipulation is performed using the `pandas` library, which is widely used in the bioinformatics and data science community for its ease of use and performance. Bioinformatic file formats such as FASTA are parsed and manipulated using the `scikit-bio` library, which provides a wide range of functionality for working with biological data.

3.3.1 Circularity detection

In the `vdsearch` pipeline, cccRNAs are detected using a modified and improved version of the reference-free and *de novo* Cirit algorithm[12] implemented in the Nim programming language.[13] This method relies upon assembly errors for circular sequences resulting in terminal repeats. During assembly, reads spanning the backsplice junction are erroneously assembled into a linear contig with head-to-tail repeats. This is a detectable signal that can be used to identify both circular sequences and, with some additional modifications, multimeric linear intermediates.

[11] Steinegger and Söding, "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets."

[12] Yuan Gao, Jinfeng Wang, and Fangqing Zhao, "CIRI: An Efficient and Unbiased Algorithm for de Novo Circular RNA Identification," *Genome Biology* 16, no. 1 (December 2015): 4, doi:10.1186/s13059-014-0571-3; Qin et al., "Reference-Free and de Novo Identification of Circular RNAs."

[13] Andreas Rumpf, *Mastering Nim: A Complete Guide to the Programming Language*, 2022.

Detecting cccRNAs via this method requires searching forward for the last several bases of the sequence (the seed) and, if a match is found, comparing backwards to the start of the sequence. If the start and end of the sequence "overlap", this repetitive region is then trimmed off.

For example, consider this sequence:

```
>example
GTGGCAGTGTCTTTTAACCTGCATTGTAGGGCCTAGTGTCATATAACCAACACGAACTGGTCTATC
GTGGCAGTGTCTTTTAACCT
```

It is clear that the last 20 bases are identical to the first 20 bases. The algorithm would detect this by first searching for the last few bases (in the default implementation, 10). It would find a match starting at position 10:

```
>example
          **********
GTGGCAGTGTCTTTTAACCTGCATTGTAGGGCCTAGTGTCATATAACCAACACGAACTGGTCTATC
GTGGCAGTGTCTTTTAACCT
```

Them it would compare backwards from the start of the sequence and find that the last 20 bases are identical to the first 20 bases:

```
>example
********************
GTGGCAGTGTCTTTTAACCTGCATTGTAGGGCCTAGTGTCATATAACCAACACGAACTGGTCTATC
GTGGCAGTGTCTTTTAACCT
```

The algorithm would then trim off the repetitive region and report the sequence as a cccRNA:

```
>example
GTGGCAGTGTCTTTTAACCTGCATTGTAGGGCCTAGTGTCATATAACCAACACGAACTGGTCTATC
```

However, the basic implementation of Qin et al.[14] is unable to monomerise multimeric transcripts resulting from rolling circle replication among known viroid-like agents due to its single pass design. A simple multimer would only be trimmed once. For example,

[14] "Reference-Free and de Novo Identification of Circular RNAs."

```
>example
GTGGCAGTGTCTTTTAACCTGCATTGTAGGGCCTAGTGTCATATAACCAACACGAACTGGTCTATC
GTGGCAGTGTCTTTTAACCTGCATTGTAGGGCCTAGTGTCATATAACCAACACGAACTGGTCTATC
GTGGCAGTGTCTTTTAACCTGCATTGTAGGGCCTAGTGTCATATAACCAACACGAACTGGTCTATC
GTGGCAGTGTCTTTTAACCT
```

would have the last 20 bases trimmed off as above to erroneously yield:

```
>example
GTGGCAGTGTCTTTTAACCTGCATTGTAGGGCCTAGTGTCATATAACCAACACGAACTGGTCTATC
GTGGCAGTGTCTTTTAACCTGCATTGTAGGGCCTAGTGTCATATAACCAACACGAACTGGTCTATC
GTGGCAGTGTCTTTTAACCTGCATTGTAGGGCCTAGTGTCATATAACCAACACGAACTGGTCTATC
```

Another limitation of the software implemented by Qin et al.[15] is that it requires exact sequence identity within the repetitive regions. For viroids and viroid-like RNAs, given their replication with low-fidelity hijacked DdRP (see Section 1.1.3), mutations within the repeated region (which could be an entire monomer) would prevent trimming to unit length. The re-implementation in `vdsearch` solves these problems by reiteratively attempting to monomerise putative cccRNAs while allowing for a configurable minimum identity within repeats. For viroid detection, the default requires minimum of 95% identity with no insertions or deletions within the overlapping region. In addition, the ratio of the length of the contig and computed monomers is reported. Sequences with monomer lengths below a threshold of 100 nt are automatically excluded as unlikely to be viroid-like.

Because this is the first step of the filtering pipeline, performance is critical. The implementation of Qin et al.[16] is written in Java and is not particularly performant. The software also requires a large amount of memory, which is surprising given the sequence-by-sequence nature of the data processing. Therefore a re-implementation was warranted Python, a dynamic language without a just-in-time compiler, could not plausibly be used for this step due to its poor performance on string manipulation. A lower-level language such as C or Rust would be more appropriate, but

[15]

[16]"Reference-Free and de Novo Identification of Circular RNAs."

the Nim programming language was chosen for its balance of performance and ease of use. Nim is a statically-typed language with a syntax similar to Python with a focus on performance. Indeed, the translation of the original re-implementation of the Cirit algorithm was straightforward and resulted in a dramatic performance improvement. To implement communication between the Python and Nim code, the Nim code is compiled as a shared library and called from Python using the `nimporter` package. This package is used to import the Nim module as if it were a Python module, allowing for seamless integration with the rest of the pipeline. However, as described below in Section 3.4.3, this approach has limitations in terms of packaging and distribution.

The limitations of and improvements to this implementation are discussed later in Chapter 5, in which an improved version is introduced.

3.3.2 cccRNA deduplication

Standard approaches to sequence deduplication are insufficient for cccRNAs. Most modern approaches, such as the extremely performant `seqkit rmdup` command,[17] rely on hashing for memory efficiency. Such approaches are effective for linear sequences but circular sequences pose a challenge due to the arbitrariness of their start position.

Consider for example the sequences `TAAAA` and `TTATT`. If one were to take the reverse complement of `TAAAA` (*i.e.,* `ATTTT`) and rotate it to the right by two bases, one would have `TTATT`. No standard deduplication method would consider these sequences identical due to their differing polarities and, more importantly, their differing start positions. However, if these sequences were to be viroid-like cccRNAs, they would in fact be equivalent after considering both sequence polarity and rotation.

To enable deduplication of putative cccRNAs, a canonical representation must be defined. For the purpose of `vdsearch` (and later `circkit` in Chapter 5), the

[17] Wei Shen et al., "SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation," ed. Quan Zou, *PLOS ONE* 11, no. 10 (October 2016): e0163962, doi:10.1371/journal.pone.0163962.

canonical representation is defined as the lexicographically smaller (*i.e.*, alphabetically earlier) of the lexicographically minimal rotations of the sequence and its reverse complement. This approach is drawn from k-mer counting methods[18] and is particularly useful because of its simplicity. Indeed, the Nim implementation of this method is only eight lines long:

```
proc minimalCanonicalRotation*(x: Dna): Dna =                             ①
  var rotatedCanonical: Dna                                               ②
  result = x                                                              ③
  var rc = x.reverseComplement                                            ④
  for i in 0..x.high:                                                     ⑤
    rotatedCanonical = min($rotate($x, i), $rotate($rc, i)).toDna         ⑥
    if rotatedCanonical < result:                                         ⑦
      result = rotatedCanonical                                           ⑧
```

① The `Dna` type is a custom type representing a DNA sequence but implemented internally as precisely identical to a `string`.

② The `rotatedCanonical` variable will store the lexicographically minimal rotation of the sequence and its reverse complement while iterating through the sequence

③ The `result` is initialized assuming the assuming the sequence is already rotationally canonical. This variable is automatically returned by the procedure.

④ The reverse complement function is provided by a library I created for this project `bioseq` that is not yet independently distributed.

⑤ The loop iterates through all rotations of the sequence and its reverse complement.

⑥ The lexicographically minimal rotation is computed for both polarities using `string` methods (hence the `$` for conversion) and converted back to a `Dna` object.

[18] Páll Melsted and Jonathan K Pritchard, "Efficient Counting of k-Mers in DNA Sequences Using a Bloom Filter," *BMC Bioinformatics* 12, no. 1 (December 2011): 333, doi:10.1186/1471-2105-12-333; Guillaume Marçais and Carl Kingsford, "A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of k-Mers," *Bioinformatics* 27, no. 6 (March 2011): 764–770, doi:10.1093/bioinformatics/btr011.

⑦ The new rotation is compared to the current result.

⑧ If the new rotation is lexicographically smaller, it is stored as the new result

However, this implementation is not particularly efficient. It has a time complexity of $O(n^2)$, where n is the length of the sequence. This is because the base-by-base comparison of the rotations is $O(n)$ and the number of rotations is also $O(n)$. This is not a problem for the typical length of viroids, but it is a problem for the length of the cccRNAs identified in Chapter 4.

Luckily, it is able to be further optimized. The lexicographically minimal rotation of a sequence can be computed in linear time and constant memory using the Duval algorithm.[19] This algorithmic improvement is implemented in the `circkit` library in the Rust programming language, which is discussed in Chapter 5.

3.3.3 Ribozyme-based filtering

The primary method by which putative cccRNAs are identified as viroid-like is the presence of self-cleaving ribozymes. To identify sequences likely to replicate via ribozyme-catalysed RCR, cccRNAs are searched for the presence of known self-cleaving ribozymes using Infernal.[20] To use pre-built covariance models for the ribozymes, the pipeline automatically downloads a curated set of models from Rfam[21] and uses them to search for ribozymes in the cccRNAs. In addition, the pipeline also searches for the pospiviroid RY motif, which is not a ribozyme, using an Rfam covariance model.

In each polarity, ribozymes are identified above Rfam's curated gathering cutoff or with E-values < 0.1 [22]. Some sequences do not contain a single strongly significant ribozyme but may contain two moderately significant ribozymes. As such, sequences with ribozymes in both polarities that met these criteria are considered viroid-like.

[19] Jean Pierre Duval, "Factorizing Words over an Ordered Alphabet," *Journal of Algorithms* 4, no. 4 (December 1983): 363–381, doi:10.1016/0196-6774(83)90017-2.
[20] Nawrocki and Eddy, "Infernal 1.1."
[21] Kalvari et al., "Rfam 14."
[22] The other curated thresholds, the trusted cutoff and noise cutoff may also be used.

As an alternative inclusion criterion, sequences are considered as viroid-like if they had with one ribozyme with an E-value < 0.01 or a score above the gathering cutoff. For each polarity, only the most significant (by E-value) ribozyme are considered. To identify more divergent ribozymes that were not detected using Infernal, sequences containing one significant (E-value < 0.01) are searched using RNAmotif.[23] The `vdsearch` program automatically downloads RNAmotif models for the HHR motif as part of the installation process. If cccRNAs have a match to the HHR motif and another ribozyme match via Infernal, they are also considered viroid-like.

3.3.4 Direct sequence search

To perform direct searches against ViroidDB (see Chapter 2), MMseqs2 `easy-search` tool (version 13.45111)[24] is used with the highest available sensitivity (`-s 7.5`). This search allows for the identification of viroid-like RNAs that do not contain self-cleaving ribozymes, such as pospiviroids. For each sequence, only the most significant match as determined by bit score is considered. MMseqs2 was chosen over BLAST because it performs better at scale and can be used in the clustering pipeline, described below. The pipeline automatically allocates a temporary directory for MMseqs2 to use, although this can be overridden by the user if their specific situation requires (*e.g.*, node-local disks on a compute cluster or an external SSD on a computer with a hard disk).

3.3.5 RNA secondary structure prediction

The pipeline predicts the secondary structures of all viroid-like cccRNAs for both polarities using the ViennaRNA package.[25] For each predicted structure, the pipeline computes the percentage of bases paired and the number of hairpins

[23] T. J. Macke, "RNAMotif, an RNA Secondary Structure Definition and Search Algorithm," *Nucleic Acids Research* 29, no. 22 (November 2001): 4724–4735, doi:10.1093/nar/29.22.4724.

[24] Steinegger and Söding, "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets."

[25] Lorenz et al., "ViennaRNA Package 2.0."

present. By default, `vdsearch` uses a temperature of 25º C and the circular prediction mode.

3.3.6 Clustering

In addition to simply identifying viroid-like RNAs, the pipeline also clusters the sequences based on their pairwise sequence identity. While several types of clustering were developed, including both alignment-based and alignment-free methods, the alignment-based method was chosen for its superior performance and ease of use. The primary success criterion for the development of the clustering method was the ability to cluster known viroid sequences at the species level. Because of the circular nature of these sequences, existing methods for clustering linear sequences were not appropriate without modification. Earlier attempts to cluster viroids produced far too many clusters (see Table 2.2) despite attempts to compensate for circularity.

To produce the alignment-based clustering, the pipeline performs an all-versus-all search using MMseqs2.[26] For this method, each sequence is concatenated to itself to compensate for potential variation in the sequence relative to otherwise-similar sequences due to their circular nature.

The clustering method is based on the pairwise average nucleotide identity (ANI) between sequences. First, MMseqs2 (version 13.45111, with command-line parameters `easy-search -s 7.5 --min-seq-id 0.40 --search-type 3 -e 0.001 -k 5 --max-seqs 1000000`) is executed. The pairwise ANI between sequences is computed by examining the alignment identity of the best hit for each pair. However, to accurately compute the ANI, the effect of concatenation must be accounted for. For two self-concatenated sequences, the sequence length is computed by taking the length of the smaller sequence and dividing by two. The computed alignment length the length is capped at the length of the now-monomerized smaller sequence. The ANI is then defined as the percent identity within the aligned region times the alignment length divided by the

[26]Steinegger and Söding, "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets."

smaller sequence monomer length. Similarly, the alignment fraction is defined as as the smaller of the doubled query coverage, doubled target coverage, or one.

To cluster the viroid-like cccRNAs based on their pairwise ANI, an undirected graph is constructed by connecting pairs of sequences where the alignment covers at least 25% of the shorter sequence with 40% identity within the alignment. The connections between the sequences are then weighted by the ANI and the Leiden algorithm (as implemented in the igraph Python library, version 0.9.10) is employed to delineate communities of similar sequences.[27] The clustering granularity is optimized by iterating over the resolution parameter space until the difference between average intra-cluster ANI and the target ANI began to increase. This process may be repeated for different target ANI values to produce clusters at different taxonomic levels.

One drawback of this method is that it is comparatively slow. While the all-versus-all search is parallelized by MMseqs2 and already quite optimized, the clustering step is not. Future work will focus on improving the performance of the clustering step, potentially by reimplementing it in a lower-level language or by using a different clustering algorithm.

3.3.7 ORF prediction

The `orfipy` library[28] is used to find ORFs present within the sequences. This library is highly optimized to perform the computationally demanding task of ORF prediction in the Cython programming language, a close relative of Python that is compiled to C [29]. However, the `orfipy` tool is not designed to handle circular sequences. To compensate for this, the sequences are concatenated to themselves

[27] V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: Guaranteeing Well-Connected Communities," *Scientific Reports* 9, no. 1 (December 2019): 5233, doi:10.1038/s41598-019-41695-z.

[28] Urminder Singh and Eve Syrkin Wurtele, "Orfipy: A Fast and Flexible Tool for Extracting ORFs," ed. Inanc Birol, *Bioinformatics* 37, no. 18 (September 2021): 3019–3020, doi:10.1093/bioinformatics/btab090.

[29] Nim is a similar language to Cython but differs in that it is its own distinct programming language that, while syntactically inspired by Python, is not directly related.

and the ORFs are searched for in the concatenated sequence. This ensures that ORFs spanning the origin were detected.

As an example, consider this sequence:

```
>ORF example
ATGAAATAG
```

If this sequence happened to be circular, the cccRNA deduplication step (see Section 3.3.2) would have rotated the sequence to its canonical representation:

```
>ORF example (canonical)
AAATAGATG
```

However, running `orfipy` on this sequence would not detect the ORF since there is no longer a start codon followed by a stop codon. The self-concatenation step ensures that the ORF is detected:

```
>ORF example (canonicalized, then concatenated)
AAATAGATGAAATAGATG
      ^^^^^^^^^
```

Now the ORF is detected as expected. However, this self-concatenation step is not without its drawbacks. For example, it yields two ORFs for the non-canonicalized sequence `ATGAAATAG`. More importantly, it is unable to detect ORFs similar to the one encoded by the satRYMV genome.[30] Improvements to the ORF prediction step that do not rely on self-concatenation are discussed in Chapter 5.

By default, the pipeline only considers ORFs longer than 100 amino acids and using the standard genetic code.

3.4 Key features

3.4.1 Simplicity and ease of use

The pipeline is designed to be easy to use and understand. Cognizant of the limited bioinformatic experience of many potential users, who range from plant pathologists to virologists, the pipeline is designed to be run with a single command that takes

[30] AbouHaidar et al., "Novel Coding, Translation, and Gene Expression of a Replicating Covalently Closed Circular RNA of 220 Nt."

as input a single file. It also automatically skips steps that have already been completed by checking for the presence of the expected output files, as would be expected from a Makefile-type system.

Should the pipeline fail, it provides helpful error messages and suggestions for how to fix the problem. It also automatically prevents data loss. If the pipeline fails during the execution of a command that outputs a newline-delimited file (*e.g.*, an Infernal or BLAST search), a future run of the pipeline might use the partial results to skip the failed step and continue from where it left off. This would result in the next step of the pipeline using partial results as input, which could lead to misleading results. To prevent this, it creates a hidden file, `.vdsearch`, at the start of each pipeline execution in the working directory that contains the results of each step of the pipeline. If the pipeline succeeds, this file is deleted. If the pipeline fails, the `.vdsearch` file is not deleted, and future runs of the pipeline will be able to detect the previous failure and warn the user to check for partial results.

3.4.2 Scalability

An early version of the pipeline was built using the Snakemake workflow management system.[31] However, the pipeline was later reimplemented as a command-line tool to improve the range of computational resources it could utilize. The Snakemake pipeline was heavily optimized for use on the NIH Biowulf high-performance computing cluster, but the reimplemented pipeline is designed to be run on a single machine with multiple cores. This eases the development and testing of the pipeline on local machines and also allows for users to run the pipeline on their own personal computers for small-scale analyses. In addition to scaling down, the pipeline can also scale up to take advantage of high-performance computing clusters. The `easy-search` subcommand take a `--threads` option that allows the user to specify the number of threads to use, and the pipeline will automatically parallelise commands that support it.

[31] J. Koster and S. Rahmann, "Snakemake—a Scalable Bioinformatics Workflow Engine," *Bioinformatics* 28, no. 19 (October 2012): 2520–2522, doi:10.1093/bioinformatics/bts480.

3.4.3 Portability

The pipeline is implemented using Python 3 and is designed to be run on any Unix-like operating system. Unfortunately, the pipeline as used in Chapter 4 is not readily packaged for distribution via package repositories such as PyPI or Conda, but it is available for download from the GitHub repository. The reason for this is the dependency on Nim, which is not widely used and requires a special build process in which a package, `nimporter` is available during dependency resolution and build time but not at runtime. As a temporary workaround, the pipeline provides a Dockerfile and container image available on Docker Hub that can be used to build a Docker image containing the pipeline and all of its dependencies. At run time, the pipeline verifies that all dependencies are available and provides helpful error messages if they are not.

Efforts to improve portability are discussed in Chapter 5, in which the circular sequence analysis and manipulation modules of `vdsearch` are reimplemented in the Rust programming language and made available as a standalone library and command-line tool.

3.5 Validation

The `vdsearch` pipeline was validated by demonstrating its ability to recover known viroid-like RNAs in transcriptomes and metatranscriptomes. For the transcriptomic validation, 1,000 Plant transcriptome (1KP) data set[32] was processed and searched. This data set was chosen due to the known presence of all type of viroid-like cccRNAs except ribozyviruses in plant transcriptomes. Before applying the search pipeline to 1KP, the raw reads were filtered for quality using the

[32] One Thousand Plant Transcriptomes Initiative, "One Thousand Plant Transcriptomes and the Phylogenomics of Green Plants," *Nature* 574, no. 7780 (October 2019): 679–685, doi:10.1038/s41586-019-1693-2.

`fastp` command-line tool[33] and assembled them using the `rnaSPAdes` package[34] using default parameters. Assembling the raw reads of these 1,344 transcriptomes resulted in 103,139,086 contigs, of which 163,970 were predicted to be circular. Of these putative cccRNAs, 42 were identified as viroid-like via ribozyme search (15 sequences), sequence search against ViroidDB (33 sequences), or both (6 sequences).

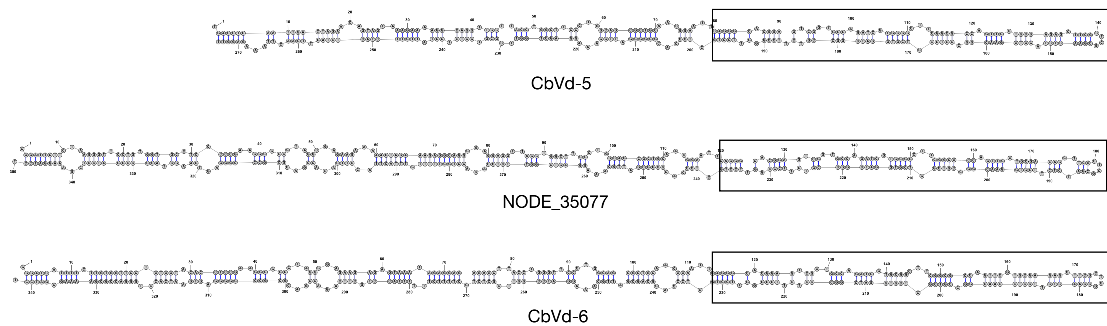


Figure 3.2: A CbVd-like sequence detected using `vdsearch`. CbVd-5 and CbVd-6 are shown for comparison, with the conserved right arm boxed.

All reads in the 1KP dataset were mapped against ViroidDB, which confirmed the presence of viroid and satellite RNA (at least one read) in 19 samples. In these samples, 16 distinct viroids and satRNAs were identified.

To verify the efficacy of the detection method, a direct search was performed of all contigs against ViroidDB, which identified 12 contigs that matched a known viroid sequence with at least 50% target coverage. The detection pipeline found four of these potentially complete contigs. Of the rejected contigs, four were much larger than typical viroids (>1000 nt) and contained major ambiguous regions.

[33] Shifu Chen et al., "Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor," *Bioinformatics* 34, no. 17 (September 2018): i884–i890, doi:10.1093/bioinformatics/bty560.

[34] Elena Bushmanova et al., "rnaSPAdes: A de Novo Transcriptome Assembler and Its Application to RNA-Seq Data," *GigaScience* 8, no. 9 (September 2019): giz100, doi:10.1093/gigascience/giz100.

The other four were low-coverage fragments that could not be verified as circular being smaller than unit length. Iresine viroid 1, Citrus exocortis viroid, and a Coleus blumei viroid (CbVd) were successfully retrieved. While the former two were nearly identical to the corresponding reference sequences, the CbVd-like sequence was not. At 350 nt, this sequence differed in length from all known coleviroids and in the terminal conserved region, which was identical to that of Dahlia latent viroid, suggesting an origin of this viroid by recombination, as reported for other CbVd species.[35] At 85% identity, this CbVd-like sequence falls below the species membership threshold for coleviroids.[36] The relatively low number of viroids identified in plant transcriptomes is likely due to selection of healthy plants for RNA isolation and transcriptome analysis.

This dataset was suboptimal for validation due to the low number of viroids present and the lack of a gold standard for *de novo* viroid detection in transcriptomes. In addition to validation on real data, simulated data would also be useful for validation. To identify the level of false negatives, I would simulate reads from known viroids and viroid-like RNAs and assemble them using `rnaSPAdes`.[37] A challenge in this approach is simulating the circular nature of viroids, which is not supported by most read simulators. However, there are tools available that can simulate circular reads such as `readSimulator`[38] that could potentially be used for this purpose.

To assess for false positives, the pipeline could be run on a set of negative control data, although the expanded host range now known for viroids and viroid-like RNAs makes using real data difficult (see Chapter 6). A simulated human transcriptome would be a good negative control. To do this simulation, I would suggest using

[35] Wan-Ying Hou et al., "Coleus Blumei Viroid 6: A New Tentative Member of the Genus Coleviroid Derived from Natural Genome Shuffling," *Archives of Virology* 154, no. 6 (June 2009): 993–997, doi:10.1007/s00705-009-0388-7; Xianzhou Nie and Rudra P. Singh, "Coleus Blumei Viroids," in *Viroids and Satellites* (Elsevier, 2017), 289–295, doi:10.1016/B978-0-12-801498-1.00027-9.

[36] Chiumenti et al., "Reassessing Species Demarcation Criteria in Viroid Taxonomy by Pairwise Identity Matrices."

[37] Bushmanova et al., "rnaSPAdes."

[38] Yu Wan, "readSimulator," December 2019, https://github.com/wanyuac/readSimulator.

a tool such as ART[39] to simulate reads from a human transcriptome and then assemble them using `rnaSPAdes`.[40] The pipeline could then be run on the resulting contigs to assess for false positives.

To validate the clustering method, it was tested on ViroidDB. Previously, 458 clusters had been identified at the average nucleotide identity (ANI) 90% level in ViroidDB using a method that was not circular-aware (Chapter 2). Using the improved method, 50 clusters were identified in ViroidDB, generally corresponding to individual species.

3.6 Conclusion

The `vdsearch` pipeline is a powerful tool for the discovery and analysis of viroid-like RNAs in high-throughput sequencing data. It is capable of identifying viroid-like RNAs in both transcriptomes and metatranscriptomes directly from assembled contigs. These contigs can be identified as putatively circular in either their monomeric or multimeric forms. From these cccRNAs, the pipeline can identify viroid-like RNAs using a combination of ribozyme search and direct sequence search, followed by structural analysis and clustering.

In additional to its utility, the pipeline is designed to be easy to use and understand, scalable, and portable. It is implemented in Python and designed to be run on any Unix-like operating system ranging from personal computers to high-performance computing clusters. This should increase the accessibility of viroid research to a wider audience and enable the discovery of new viroid-like agents in a wide range of organisms.

Validation of the pipeline demonstrated its ability to recover known viroid-like RNAs in transcriptomes and metatranscriptomes, as well as to identify a novel viroid-like RNA in a plant transcriptome. While an advancement in the field, this finding does not yet represent a new viroid species, as it has not been shown to

[39] Weichun Huang et al., "ART: A Next-Generation Sequencing Read Simulator," *Bioinformatics* 28, no. 4 (February 2012): 593–594, doi:10.1093/bioinformatics/btr708.
[40] Bushmanova et al., "rnaSPAdes."

be infectious or to replicate autonomously. However, the novelty of the putative coleviroid is limited by the fact that it is in part a recombinant of known viroids.

To find a greater number of viroid-like RNAs, the pipeline could be applied to a wider range of samples. The 1KP samples are derived from isolated plants rather than environmental samples, which are likely to contain greater diversity. Just such an analysis is described in Chapter 4, in which the pipeline is applied to a large metatranscriptomic data set from around the world.

4

A vast world of viroid-like circular RNAs

Summary

Viroids and viroid-like covalently closed circular (ccc) RNAs are minimal replicators that typically encode no proteins and hijack cellular enzymes for replication. The extent and diversity of viroid-like agents are poorly understood. A computational pipeline was developed to identify viroid-like cccRNAs and applied to 5,131 metatranscriptomes and 1,344 plant transcriptomes. The search yielded 11,378 viroid-like cccRNAs spanning 4,409 species-level clusters, a five-fold increase compared to the previously identified viroid-like elements. Within this diverse collection, numerous putative viroids, satellite RNAs, retrozymes, and ribozy-like viruses were discovered. Diverse ribozyme combinations and unusual ribozymes within the cccRNAs were identified. Self-cleaving ribozymes were identified in ambiviruses, some mito-like viruses and capsid-encoding satellite virus-like cccRNAs. The broad presence of viroid-like cccRNAs in diverse transcriptomes and ecosystems implies that their host range is far broader than currently known, and matches to CRISPR spacers suggest that some cccRNAs replicate in prokaryotes.

4.1 Introduction

Viroids and viroid-like cccRNAs are minimal replicators, or ultimate parasites, that lack genes and effectively consist only of RNA structures required for replication. As reviewed in Section 1.7, this extreme simplicity of viroids triggered speculation on their direct descendance from primordial RNA replicators.[1] However, the apparent narrow host range of viroids, which at the beginning of this thesis had been reported only in plants, was poorly compatible with this evolutionary scenario. Instead, given the similarities between retrozymes (introduced in Section 1.4) and avsunviroids, it had been suggested that avsunviroids descended from retrozymes.[2]

Given the ultimate structural simplicity of viroids and related cccRNAs and the universality of the DNA-dependent RNA polymerases involved in their replication across life forms, the apparently narrow spread and limited diversity of parasitic cccRNAs appeared puzzling. Furthermore, this apparent paucity of viroid-like agents was in stark contrast to the burgeoning diversity of RNA viruses, many thousands of which including numerous, novel groups have been discovered by metatranscriptome analyses.[3] At the beginning of this thesis, there were at least three orders of magnitude more known RNA viruses than there are viroids and viroid-like cccRNAs.

An exhaustive search for cccRNAs was performed in a collection of 5,131 diverse metatranscriptomes that had been recently employed for massive RNA virus discov-

[1] Diener, "Viroids"; Ricardo Flores et al., "A Scenario for the Emergence of Protoviroids in the RNA World and for Their Further Evolution into Viroids and Viroid-Like RNAs by Modular Recombinations and Mutations," *Virus Evolution* 8, no. 1 (January 2022): veab107, doi:10.1093/ve/veab107.

[2] De la Peña and Cervera, "Circular RNAs with Hammerhead Ribozymes Encoded in Eukaryotic Genomes"; Benjamin D. Lee and Eugene V. Koonin, "Viroids and Viroid-like Circular RNAs: Do They Descend from Primordial Replicators?" *Life* 12, no. 1 (January 2022): 103, doi:10.3390/life12010103.

[3] Wolf et al., "Doubling of the Known Set of RNA Viruses by Metagenomic Analysis of an Aquatic Virome"; Robert C. Edgar, Jeff Taylor, Victor Lin, et al., "Petabase-Scale Sequence Alignment Catalyses Viral Discovery," *Nature* 602, no. 7895 (February 2022): 142–147, doi:10.1038/s41586-021-04332-2; Ahmed A. Zayed et al., "Cryptic and Abundant Marine Viruses at the Evolutionary Origins of Earth's RNA Virome," *Science* 376, no. 6589 (April 2022): 156–162, doi:10.1126/science.abm5847; Neri et al., "Expansion of the Global RNA Virome Reveals Diverse Clades of Bacteriophages."

ery.[4] This search yielded more than 10,000 viroid-like cccRNAs that represented an about five-fold increase of the known diversity of viroid-like agents. Further analysis of these cccRNAs led to the identification of 4,121 clusters of putative viroids, satRNAs, retrozymes and ribozy-like viruses.

4.2 Methods

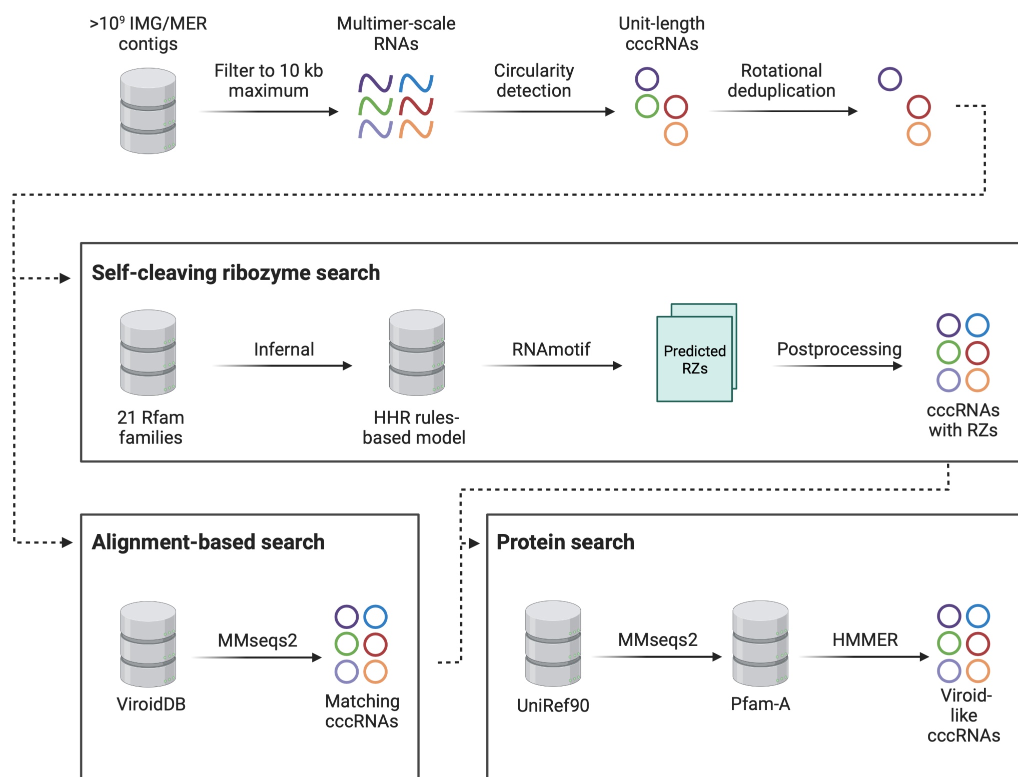


Figure 4.1: Viroid-like cccRNA detection pipeline

4.2.1 Data acquisition

The search for cccRNAs described here was performed on a collection of 5,131 assembled metatranscriptomes sourced from the IMG/MER[5] database and 1,341

[4] Neri et al., “Expansion of the Global RNA Virome Reveals Diverse Clades of Bacteriophages.”

[5] I-Min A Chen et al., “The IMG/M Data Management and Analysis System v.6.0: New Tools and Advanced Capabilities,” *Nucleic Acids Research* 49, no. D1 (January 2021): D751–D763, doi:10.1093/nar/gkaa939.

plant transcriptomes[6] from the One Thousand Plant Transcriptomes (1KP) project. This dataset is the same as the dataset used by Neri et al.[7] for a large-scale expansion of the RNA virome. The metatranscriptomes were selected to represent a wide range of ecosystems, including soil, freshwater, marine, and engineered environments. Among these, 38% were from marine environments, 21% from soil, and 13% from freshwater. Additional sources included engineered environments (8%), animal microbiomes (6%), and host-associated plant environments, specifically 4% from plant rhizospheres and 3% from plant phyllospheres. Other minor sources included plant litter (4%), general "other" categories (4%), and deep subsurface samples, which contributed less than 1% of the total. For the metatranscriptomes for which metadata were available, the vast majority (83%) were sequenced using rRNA depletion as opposed to poly(A) enrichment (7%) or total RNA sequencing (9%). Furthermore, these metatranscriptomes were chosen to represent a wide range of geographic locations, with samples from all continents. Additionally, these metatranscriptomes often contained paired metagenomic data, which was used to identify potential hosts of the identified viroid-like agents. Most metatranscriptomes were prokaryote-dominated (68%), but a significant fraction were eukaryote-dominated (20%) or mixed (11%). In addition to the IMG/MER data, the 2021-09-07 release of ViroidDB and the set of HPR sequences identified by Weinberg et al.[8] were also included. All data were downloaded, decompressed, and then checked for identical cryptographic checksums using `md5sum`.

4.2.2 Viroid-like cccRNA detection

Viroid-like cccRNAs were detected using a scalable computational pipeline developed specifically for this search. Described in full detail in Chapter 3, the pipeline identifies putative cccRNAs, removes overlapping regions, and identifies rotationally identical sequences. The pipeline then searches for self-cleaving ribozymes and

[6] One Thousand Plant Transcriptomes Initiative, "One Thousand Plant Transcriptomes and the Phylogenomics of Green Plants."

[7] "Expansion of the Global RNA Virome Reveals Diverse Clades of Bacteriophages."

[8] "Identification of over 200-Fold More Hairpin Ribozymes Than Previously Known in Diverse Circular RNAs."

matches to known viroid-like sequences and marks sequences with these hallmarks as "viroid-like." A summary of the pipeline and the data upon which it was run is shown in Figure 4.1.

4.2.3 Clustering

Viroid-like cccRNAs detected using the pipeline were clustered by sequence similarity. The methods developed for clustering circular sequences are described in Section 3.3.6. For this clustering, the sequences of ViroidDB (described in Chapter 2), the HPR dataset of Weinberg et al.[9], and the sequences identified in this chapter were combined. This combined clustering enables analysis of the extent of the novelty of the sequences identified here.

4.2.4 Protein searches

All viroid-like cccRNAs were searched for matches to known proteins. The primary search method used was performing translated searches (BLASTX-style) against the UniRef90 protein database[10] using the MMseqs2 package.[11] For each cccRNA, only the best match by E-value was considered.[12]

As a second approach, the ORFs from all cccRNAs, viroid-like or not, were also searched using the HMMER package.[13] Both the full Pfam-A profile database as well as a curated subset (the profiles for RdRP clan combined with the HDVAg profile) were searched using the `hmmsearch` and `hmmscan` commands, respectively.

[9]"Identification of over 200-Fold More Hairpin Ribozymes Than Previously Known in Diverse Circular RNAs."

[10]B. E. Suzek et al., "UniRef Clusters: A Comprehensive and Scalable Alternative for Improving Sequence Similarity Searches," *Bioinformatics* 31, no. 6 (March 2015): 926–932, doi:10.1093/bioinformatics/btu739.

[11]Steinegger and Söding, "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets."

[12]Keeping only the best match made analysis significantly simpler and no viroid-like cccRNA was known to encode more than one protein. However, with nucleotide-nucleotide searches, it was not implausible that the same sequence could match two different viroids. Indeed, just such a case happened in Section 3.5, in which a hybrid viroid was found. In retrospect, a hierarchical data storage format would have been preferable.

[13]Sean R. Eddy, "Accelerated Profile HMM Searches," ed. William R. Pearson, *PLoS Computational Biology* 7, no. 10 (October 2011): e1002195, doi:10.1371/journal.pcbi.1002195.

4.2.5 HDV antigen analysis

Sequences were clustered using the CLANS package with BLASTP option (BLOSUM62 matrix, E-value cutoff of 1×10^{-3}).[14] Sequence similarity among reference HDVAg from GenBank and those from metatranscriptomic datasets was analysed with the Sequence Demarcation Tool.[15] For phylogenetic analysis, HDVAg-like sequences were aligned using PROMALS3D.[16] Due to the short length of the sequences, the alignment was not further processed. Maximum likelihood phylogenetic analysis was performed using IQ-TREE.[17] The best fitting model was selected by IQ-TREE and was VT+F+R4. The tree was visualized with iTOL.[18]

4.2.6 Read mapping

Bowtie2 was used to perform read mapping from the entire viroid-like cccRNA to the 1KP transcriptomes data set in parallel. Bowtie was configured to use its most sensitive setting (`--very-sensitive`) and ignore unaligned reads, as only the reads that mapped to the cccRNAs were of interest.

4.2.7 CRISPR spacer analysis

Viroid-like sequences were compared to predicted CRISPR spacers from prokaryotic (meta)genomes to identify potential cases of spacer acquisition from, and possible defense against, viroids by prokaryotes. The full set of 22,109 viroid and viroid-like sequences, including all reference sequences and sequences identified in this

[14]Tancred Frickey and Andrei Lupas, "CLANS: A Java Application for Visualizing Protein Families Based on Pairwise Similarity," *Bioinformatics* 20, no. 18 (December 2004): 3702–3704, doi:10.1093/bioinformatics/bth444.

[15]Brejnev Muhizi Muhire, Arvind Varsani, and Darren Patrick Martin, "SDT: A Virus Classification Tool Based on Pairwise Sequence Alignment and Identity Calculation," ed. Jens H. Kuhn, *PLoS ONE* 9, no. 9 (September 2014): e108277, doi:10.1371/journal.pone.0108277.

[16]Jimin Pei and Nick V. Grishin, "PROMALS3D: Multiple Protein Sequence Alignment Enhanced with Evolutionary and Three-Dimensional Structural Information," in *Multiple Sequence Alignment Methods*, ed. David J Russell, vol. 1079 (Totowa, NJ: Humana Press, 2014), 263–271, doi:10.1007/978-1-62703-646-7_17.

[17]Lam-Tung Nguyen et al., "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies," *Molecular Biology and Evolution* 32, no. 1 (January 2015): 268–274, doi:10.1093/molbev/msu300.

[18]Ivica Letunic and Peer Bork, "Interactive Tree Of Life (iTOL) V5: An Online Tool for Phylogenetic Tree Display and Annotation," *Nucleic Acids Research* 49, no. W1 (July 2021): W293–W296, doi:10.1093/nar/gkab301.

chapter, was compared to 1,961,109 CRISPR spacers predicted from whole genomes of bacteria and archaea (vJune2022) and 61,658,467 CRISPR spacers predicted from metagenomes in the IMG database[19] using BLASTN v2.9.0 with options `-dust no -word_size 7`.

To initially link viroid-like sequences to CRISPR spacers, only hits with 0 or 1 mismatch over the entire spacer were considered. To minimise the number of false-positive hits due to low-complexity and/or repeat sequences, CRISPR spacers were excluded from this analysis if any of the following were true:

(i) They were encoded in a predicted CRISPR array including 2 spacers or less

(ii) Less than 66% of the predicted repeats were 100% identical to each other

(iii) They were more than 20 bp

(iv) They included a low-complexity or repeat sequence as detected by dust-masker (v1.0.0)[20] (options `-window 20 -level 10`) or a direct repeat of at least 4 bp detected with etandem[21] (options `-minrepeat 4 -maxrepeat 15 -threshold 2`)

To find additional spacer matches, all members of the clusters with a spacer match were searched against the IMG public metagenomic spacer data (dated 2022-06-18) set using IMG's workspace BLAST with a minimum E-value of 1×10^{-5}. The repeats matching loci were extracted using MinCED[22] and searched against nt using BLASTN v2.13.0.[23]

[19] Chen et al., "The IMG/M Data Management and Analysis System v.6.0."

[20] Aleksandr Morgulis et al., "A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences," *Journal of Computational Biology* 13, no. 5 (June 2006): 1028–1040, doi:10.1089/cmb.2006.13.1028.

[21] P. Rice, I. Longden, and A. Bleasby, "EMBOSS: The European Molecular Biology Open Software Suite," *Trends in Genetics* 16, no. 6 (June 2000): 276–277, doi:10.1016/s0168-9525(00)02024-2.

[22] Charles Bland et al., "CRISPR Recognition Tool (CRT): A Tool for Automatic Detection of Clustered Regularly Interspaced Palindromic Repeats," *BMC Bioinformatics* 8, no. 1 (December 2007): 209, doi:10.1186/1471-2105-8-209.

[23] S. F. Altschul et al., "Basic Local Alignment Search Tool," *Journal of Molecular Biology* 215, no. 3 (October 1990): 403–410, doi:10.1016/S0022-2836(05)80360-2; Christiam Camacho et al., "BLAST+: Architecture and Applications," *BMC Bioinformatics* 10, no. 1 (December 2009): 421, doi:10.1186/1471-2105-10-421.

4.2.8 Software resources

All the software resources used in this chapter are freely available and listed in Table 4.1.

Table 4.1: Software used for viroid-like cccRNA detection and analysis

Software	Author
vdsearch pipeline	Chapter 3
CD-HIT v4.8.1	Weizhong Li and Adam Godzik[24]
HMMER v3.3.2	Eddy[25]
Infernal v1.1.4	Nawrocki and Eddy[26]
python-igraph v0.9.10	Csárdi Gábor and Tamás Nepusz[27]
scikit-bio v0.5.6	J Gregory Caporaso et al.[28]; Rob Knight et al.[29]
MMseqs2 v13.45111	Steinegger and Söding[30]
R v4.2.0	R Foundation for Statistical Computing
Python v3.8.3	Python Software Foundation
Nim v1.6.2	Rumpf[31]
ViennaRNA v2.5.0	Lorenz et al.[32]
SeqKit v2.1.0	Shen et al.[33]
Pandas v1.2.0	Wes McKinney[34]
ggplot2 v3.3.6	Hadley Wickham[35]
orfipy v0.0.4	Singh and Wurtele[36]
fastp v0.20.1	Chen et al.[37]

[24] "Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences," *Bioinformatics* 22, no. 13 (July 2006): 1658–1659, doi:10.1093/bioinformatics/btl158.

[25] "Accelerated Profile HMM Searches."

[26] "Infernal 1.1."

[27] "The Igraph Software Package for Complex Network Research," *InterJournal Complex Systems*, November 2005, 1695.

[28] "QIIME Allows Analysis of High-Throughput Community Sequencing Data," *Nature Methods* 7, no. 5 (May 2010): 335–336, doi:10.1038/nmeth.f.303.

[29] "PyCogent: A Toolkit for Making Sense from Sequence," *Genome Biology* 8, no. 8 (2007): R171, doi:10.1186/gb-2007-8-8-r171.

[30] "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets."

[31] *Mastering Nim.*

[32] "ViennaRNA Package 2.0."

[33] "SeqKit."

[34] "Data Structures for Statistical Computing in Python" (Python in Science Conference, Austin, Texas, 2010), 56–61, doi:10.25080/Majora-92bf1922-00a.

[35] *Ggplot2*, Use R! (Cham: Springer International Publishing, 2016), doi:10.1007/978-3-319-24277-4.

[36] "Orfipy."

[37] "Fastp."

Software	Author
Snakemake v6.10.0	Felix Mölder et al.[38]
circlize v0.4.15	Zuguang Gu et al.[39]
RNAmotif v3.1.1	Macke[40]
rnaSPAdes v3.14.1	Bushmanova et al.[41]
MinCED v0.4.2	Bland et al.[42]
IQ-TREE v1.6.12	Nguyen et al.[43]
CLANS	Frickey and Lupas[44]
Sequence Demarcation Tool (SDT) v1.2	Muhire, Varsani, and Martin[45]
PROMALS3D	Pei and Grishin[46]
iTOL v6.5.8	Letunic and Bork[47]
bowtie2 v2.4.2	Ben Langmead and Steven L Salzberg[48]
BLAST+ suite	Altschul et al.[49]; Camacho et al.[50]
EMBOSS etandem v6.0.0	Rice, Longden, and Bleasby[51]

4.3 Results

4.3.1 A five-fold expansion of the known diversity of viroid-like cccRNAs

After testing the pipeline on plant transcriptomes (described in Section 3.5), it was applied to a set of 5,131 diverse metatranscriptomes totalling 1.5 billion metatranscriptomic contigs (708 Gbp) after size filtration. 10,183,455 putative cccRNAs were identified with a median contig length of 269 nt. After removing overlapping regions and eliminating rotationally identical sequences, the median length of the

[38] "Sustainable Data Analysis with Snakemake," *F1000Research* 10 (January 2021): 33, doi:10.12688/f1000research.29032.1.
[39] "Circlize Implements and Enhances Circular Visualization in R," *Bioinformatics* 30, no. 19 (October 2014): 2811–2812, doi:10.1093/bioinformatics/btu393.
[40] "RNAMotif, an RNA Secondary Structure Definition and Search Algorithm."
[41] "rnaSPAdes."
[42] "CRISPR Recognition Tool (CRT)."
[43] "IQ-TREE."
[44] "CLANS."
[45] "SDT."
[46] "PROMALS3D."
[47] "Interactive Tree Of Life (iTOL) V5."
[48] "Fast Gapped-Read Alignment with Bowtie 2," *Nature Methods* 9, no. 4 (April 2012): 357–359, doi:10.1038/nmeth.1923.
[49] "Basic Local Alignment Search Tool."
[50] "BLAST+."
[51] "EMBOSS."

8,748,001 resultant monomers was 165 nt. Of these, 2,791,251 were within the known size range of viroids (200–400 nt). A total of 11,378 cccRNAs ranging in size from 100 nt to 5,867 nt were classified as viroid-like because they contained a confidently predicted self-cleaving ribozyme in at least one RNA polarity. No metatranscriptomic cccRNAs matched the pospiviroid RY motif.

Among these viroid-like cccRNAs, 10,181 were detected by alignment-free methods (*i.e.*, they showed no detectable sequence similarity to known viroids). The remaining 1,197 sequences shared significant nucleotide sequence similarity with known viroid-like RNAs spanning the entire gamut from viroids to satRNAs to retrozymes. 907 sequences were identified as viroid-like by both the ribozyme detection and alignment-based search approaches. Among the 10,181 ribozyme-containing viroid-like cccRNAs unrelated to known viroids, 3,434 were symmetric, (*i.e.*, they contained predicted ribozymes in both polarities). Of the 5,131 metatranscriptomes searched, 1,841 contained at least one viroid-like cccRNA.

Of the sequences aligning to viroid-like agents, the majority only contained short (<40 nt) alignable regions, generally localized to the ribozyme motifs. However, 33 sequences yielded long (>100 nt) alignments. These cccRNAs aligned to Tobacco ringspot virus satRNA (satTRSV), Lucerne transient streak virus satRNA (satLTSV), citrus dwarfing viroid (CDVd), and two retrozymes. For satTRSV and satLTSV, the range of identity between the recovered cccRNAs and the reference sequence ranged 80%–98% and 81%–99%, respectively. The match to CDVd was 80% identical to the nearest reference sequence. In all cases, the cccRNAs were similar in length and structure to the reference sequences.

The viroid-like cccRNAs identified in this chapter were clustered to estimate the increase in diversity compared to previously known viroid-like RNAs (Figure 4.2). Aligning cccRNAs poses a challenge due to the variation in the rotation of the sequences. Two identical cccRNAs could appear to have only half the bases aligning if rotated completely out of phase. Therefore, special care was taken

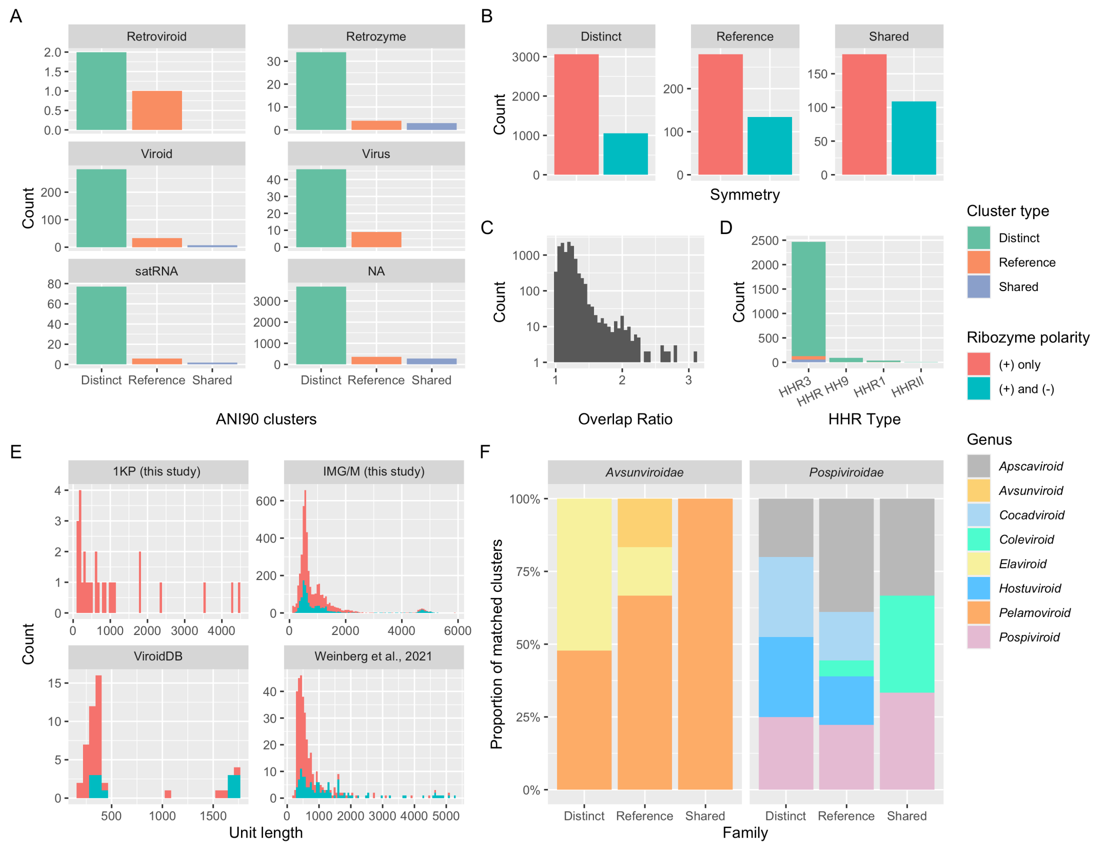


Figure 4.2: Viroid-like cccRNAs identified in metatranscriptomes. A. Number of ANI90 clusters with most significant matches to given viroid-like cccRNA agent types which are either "distinct" (derived exclusively from transcriptome and metatranscriptome analysis in this work), "reference" (no members distinct from previously identified ones), or "shared" (containing at least one of both types of sequence). B. Comparative distribution of inferred ribozyme architectures by cluster type. C. Plot of overlap ratios in cccRNAs, defined as the assembled length divided by the monomer length, from IMG and 1KP. D. Counts of HHR types in representative clusters. E. Length distributions of cluster representatives in the present analysis (transcriptomes and metatranscriptomes), ViroidDB, and a previous study Weinberg et al.[52] F. Relative abundance of clusters matching different genera within each viroid family by cluster type

to compensate for the circularity of the sequences during the postprocessing of the pairwise nucleotide search results (see Section 3.3.6).

In the combined metatranscriptomic, transcriptomic, and reference datasets, 4,823 ANI90 clusters were identified, of which 4,121 did not include any sequences from the reference datasets and thus were considered distinct. This corresponds to a 5.9-fold increase in viroid-like RNA diversity. Of the remaining 702 clusters containing at least one known sequence, 288 (41%) had distinct members in addition to

the reference sequences. Notably, 39 distinct clusters were represented in plant transcriptomes, of which 8 were symmetric.

The relative abundance of HHR types in the cccRNAs varied significantly from what would be expected given the sequence and species count. Within Rfam, HHR1 swamps HHR3 by two orders of magnitude by sequence count (190,679 vs 538 sequences). However, among the cccRNA cluster representatives, HHR3 was found in 94% and was two orders of magnitude more common than HHR1 (1,952 vs 32). Given the dominance of HHR3 in known viroids,[53] this overabundance of HHR3 is suggestive of the presence of numerous viroid-like cccRNAs.

4.3.2 Putative novel viroid-like cccRNAs

The 5 largest novel ANI90 clusters (denoted 1 to 5, in the descending order of the cluster size) derived from the metatranscriptomic data are described to exemplify the type of findings obtained in this chapter. All these clusters included members with symmetric, matched ribozymes. The cccRNAs in four of these clusters contained matched HHR3s, whereas those in the fifth cluster contained twister-P1 ribozymes.

Table 4.2: Summary of the five largest clusters of novel viroid-like cccRNAs

Cluster	Cluster size	Mean length	Ribozymes
1	149	562 nt	HHR3/HHR3
2	68	494 nt	Twister-P1/Twister-P1
3	61	605 nt	HHR3/HHR3
4	61	615 nt	HHR3/HHR3
5	55	454 nt	HHR3/HHR3

Cluster 1

The largest, cluster 1, consisted of 149 sequences with a mean length of 562 nt (±9.0 nt). During circularity detection, an average of 18% of the monomer was removed, although one member of the cluster yielded a contig with a 60% (341 nt) overlap.

[53]Flores, Minoia, et al., "Viroid Replication"; Navarro, Rubino, and Di Serio, "Small Circular Satellite RNAs."

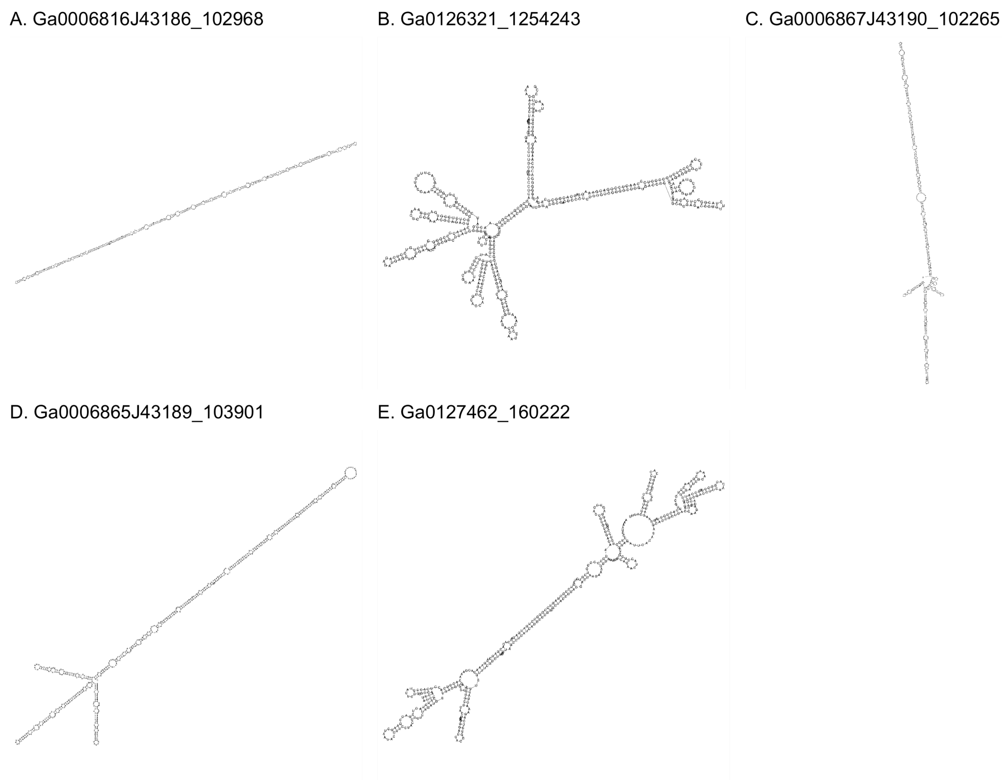


Figure 4.3: Predicted secondary structures of representatives of the five largest clusters of distinct viroid-like cccRNAs. Structures were predicted using ViennaRNA's RNAfold program configured to operate on circular sequences. Sequence data and metadata are available in Supplementary Table S1.

The cccRNAs in this cluster are predicted to adopt a rod-shaped conformation with 74% of the bases (on average) paired in both polarities. Most members of this cluster (n=137) contain symmetric HHR3s, whereas for the remaining 12 members, only one ribozyme was predicted, suggesting the presence of a divergent HHR3. Among the members of this cluster, 38 sequences yielded a short (26–37 nt) alignment to the HHR3 of Eggplant latent viroid or Grapevine hammerhead viroid-like RNA. However, the cccRNAs comprising this cluster are substantially longer than the respective viroids (334 and 375 nt, respectively). The majority of the cluster members (n=133) were found in terrestrial metatranscriptomes from 11 distinct locations, 4 members were identified in 3 distinct freshwater locations, and one member was found in a spruce rhizosphere sample.

Cluster 2

The cccRNAs in cluster 2 (68 members) contained twister-P1 ribozymes in both polarities. All but 4 of these cccRNAs are 494 nt in length and form a branched structure with a mean of 70% base complementarity. These cccRNAs were found in 10 locations in terrestrial ecosystems, such as soil and plant litter, with nearly half (n=31) identified in switchgrass phyllosphere samples. Cluster members were repeatedly found during sampling of the switchgrass phyllosphere over the course of a year at a sample site in Michigan, United States.

Cluster 3

Like cluster 1, cluster 3 (61 members) consisted of comparatively large (605 nt), rod-shaped cccRNA with symmetric HHR3s. However, unlike the largest cluster, no member shows a detectable direct nucleotide match to any ViroidDB sequence, ribozyme or otherwise. In the majority (n=40) of the members symmetric ribozymes were not detected, but remaining ones were symmetric, again, suggesting the presence of divergent HHRs.

On average, 28% of the monomer's length was cleaved during circularity detection although one member was sequenced at 2.78 unit length (almost a complete head-to-tail trimer). The largest two members of the cluster (1,693 and 1,640 nt originally, 1,224 nt after cleavage) were not correctly monomerized by the circularity detection procedure due to mismatches in the seed sequence (see Methods). Manual monomerisation of these sequences showed they both were 612 nt, resulting in overlap ratios of 2.76 and 2.67, respectively. Alignment results in 99.0% and 99.5% identity between monomers within each dimer, approximately the error rate of RNA polymerase II when using an RNA template.[54]

[54] Jian Wu and David M. Bisaro, "Biased Pol II Fidelity Contributes to Conservation of Functional Domains in the Potato Spindle Tuber Viroid Genome," *PLoS Pathogens* 16, no. 12 (December 2020): e1009144, doi:10.1371/journal.ppat.1009144; Gago et al., "Extremely High Mutation Rate of a Hammerhead Viroid"; López-Carrasco et al., "Different Rates of Spontaneous Mutation of Chloroplastic and Nuclear Viroids as Determined by High-Fidelity Ultra-Deep Sequencing."

Between 73% and 83% of the bases in these cccRNAs were paired in the (+) polarity and between 74% and 82% of bases were paired in the (-) polarity. Almost all members of this cluster were identified in soil samples from 6 locations around the world (including Colombia, Czech Republic, Germany, and United States), and two were found in creek freshwater samples.

Cluster 4

Cluster 4 (*n*=61) is similar to clusters 1 and 3 in that its members consist of 615 nt and are predicted to form rods containing HHR3s in both polarities. The secondary structure results in a high mean self-complementarity of 75% of bases. However, these cccRNAs contain no regions that were able to be aligned to the other large clusters or to any ViroidDB sequence. As with many other HHR3 rods, the majority (*n*=39) contain only one HHR3 above the significance threshold. All but two members of the cluster were derived from eight soil locations, and the remaining ones were found in the spruce rhizosphere.

Cluster 5

Among the largest clusters examined, cluster 5 (55 members) is the only one that consists of cccRNAs with predicted quasi-rod shaped structures, with 64% bases paired on average. At 454 nt mean length (±8.9 nt), these cccRNAs are the smallest among the 5 largest clusters.

In this case, all members contained two HHR3s, but no nucleotide matches to ViroidDB were detected. Most members were found in terrestrial samples including soil (*n*=38) and plant litter and peat (*n*=2). As in the case of cluster 2, 14 members of this cluster were also found over the span of a year in the phyllosphere of switchgrass, and one member was found in a freshwater sample. Altogether, nine distinct locations contained members of this cluster.

In summary, analysis of these 5 largest clusters showed that they consisted of cccRNAs endowed with all the hallmarks of viroids including symmetric ribozymes (HHR3s, with one notable exception), extensive branched or (quasi) rod-shaped

secondary structure, and evidence of multimeric intermediates. Furthermore, the members of these clusters were independently identified in diverse samples, primarily, those from soil, indicating they are widespread and are likely to be infectious agents.

4.3.3 Geographic and ecological distribution of viroid-like cccRNAs

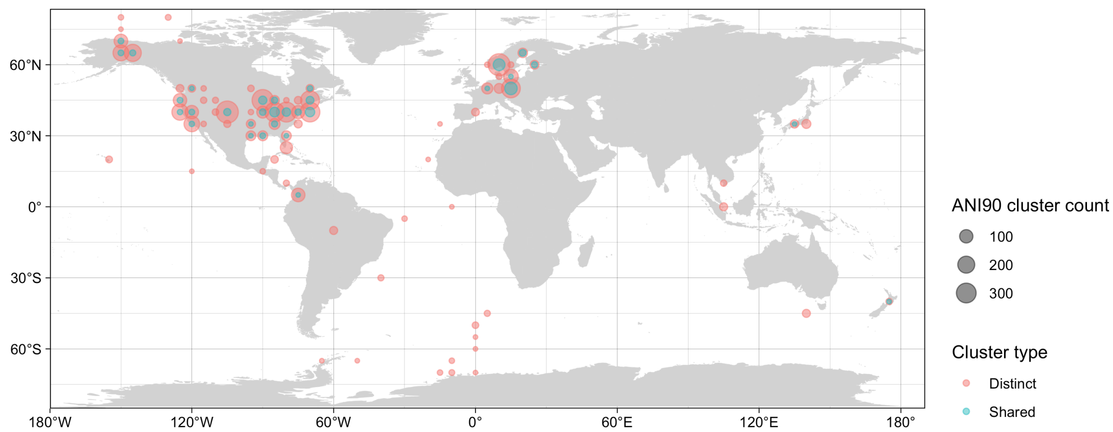


Figure 4.4: Map of sample locations from which viroid-like cccRNAs were detected. The size of each circle corresponds to the number of clusters identified in each location (grouped to the nearest five degrees of latitude and longitude) while the color represents the fraction of distinct clusters (blue shows reference clusters and red, distinct clusters).

The global distribution of the cccRNA clusters was examined. Distinct clusters were found throughout the world (Figure 4.4) and in all types of ecosystems (Figure 4.5).

Table 4.3: Summary of viroid-like cccRNA detection in different ecosystems

Ecosystem	Positive Samples	Negative Samples	Total Samples	Percentage Positive
Aquatic; Marine	323	1,645	1,968	16.4%
Aquatic; Freshwater	282	386	668	42.2%
Aquatic; Other	39	143	182	21.4%
Terrestrial; Soil	668	439	1,107	60.3%
Terrestrial; Plant litter and peat	162	22	184	88.0%
Terrestrial; Deep subsurface	0	7	7	0.0%
Host-associated; Plant Rhizosphere	128	54	182	70.3%
Host-associated; Plant Phyllosphere	64	77	141	45.4%

Ecosystem	Positive Samples	Negative Samples	Total Samples	Percentage Positive
Host-associated; Animal Microbiome	48	276	324	14.8%
Engineered	123	264	387	31.8%
Total	**1,837**	**3,313**	**5,150**	**35.7**

Across the diverse ecosystems analysed, viroid-like cccRNAs were detected with varying frequencies, as summarised in Table 4.3. In plant-associated terrestrial ecosystems, a high prevalence was observed: in plant litter and peat, 162 out of 184 samples were positive for viroid-like cccRNAs, while in the plant rhizosphere, 128 out of 182 samples contained these sequences. Soil samples similarly showed a significant number of positives, with 668 out of 1,107 samples testing positive. Soil samples were the primary source of both distinct and shared clusters, reflecting the apparent greater sequence diversity in soils. In aquatic ecosystems, detection rates varied; freshwater environments had 282 positives out of 668 samples, whereas marine environments had 323 positives out of 1,968 samples. Host-associated animal microbiome samples showed lower detection rates, with 48 positives out of 324 samples, and notably, no viroid-like cccRNAs were detected in the terrestrial deep subsurface samples (0 out of 7). These findings indicate that viroid-like cccRNAs are most prevalent in plant-associated terrestrial ecosystems, less common in aquatic environments, and rare or undetected in animal-associated and deep subsurface samples.

The viroid-like cccRNAs displayed non-uniform ribozyme distribution among ecosystems (Figure 4.6). Mismatched HPR/HHR ribozymes were especially prevalent among samples from plant rhizospheres, whereas matched HHRs were notably more abundant in engineered ecosystems, such as bioreactors, than in soil environments.

Among the 10 most geographically dispersed, distinct clusters, 8 included symmetric ribozymes, of which 6 were matched HHR3s. The other two symmetric clusters contain HPR-meta1/twister-P1 ribozymes and the two asymmetric clusters contain

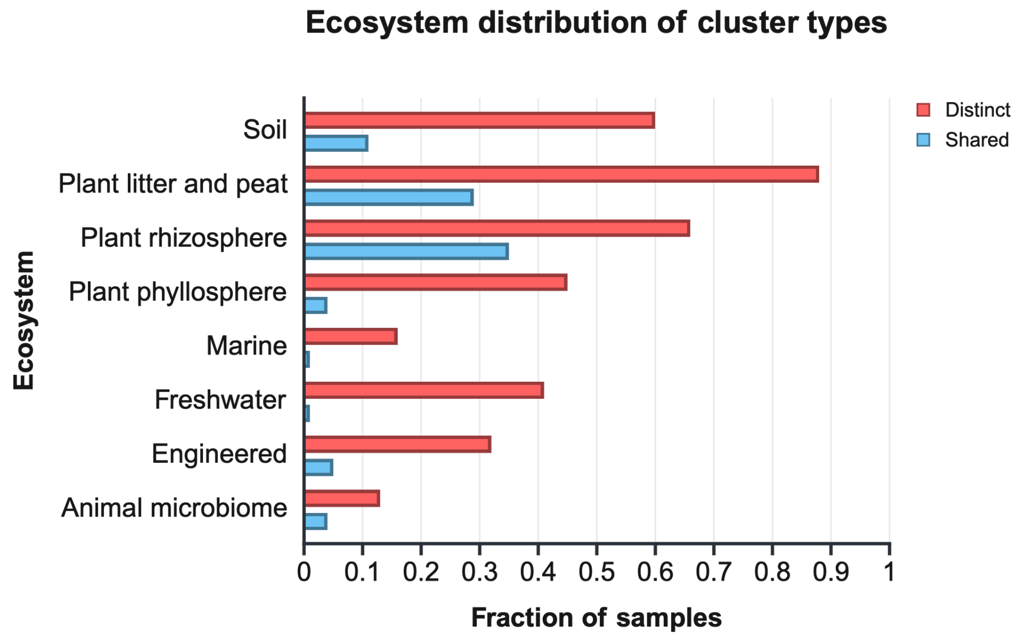


Figure 4.5: Proportion of samples in each ecosystem type that contained one or more distinct viroid-like cccRNA clusters (red) and reference clusters (blue). Created with BioRender.com.

HHR3s. These widely dispersed clusters ranged in length from 372 nt to 1,039 nt and were predicted to adopt either a rod-like shape (the 6 HHR3 clusters) or branched conformations.

The coverage depth of the viroid-like cccRNAs relative to the other contigs within the same metatranscriptomes was examined. The absolute coverage depth is not directly informative due to the variation between sequencing methods used, but the ranked coverage depth percentile can serve as a proxy for the abundance of a sequence within the sample. In the 747 samples for which coverage information was available, viroid-like cccRNAs ranged from the most covered to least covered contigs. Notably, 150 viroid-like cccRNAs were within the top percentile. Among these highly covered contigs are known viroid-like cccRNAs, such as satLTSV.

Identifying potential hosts of the viroid-like agents within these ecosystems remains a challenge. Based on the IMG annotation pipeline, the majority of the analysed

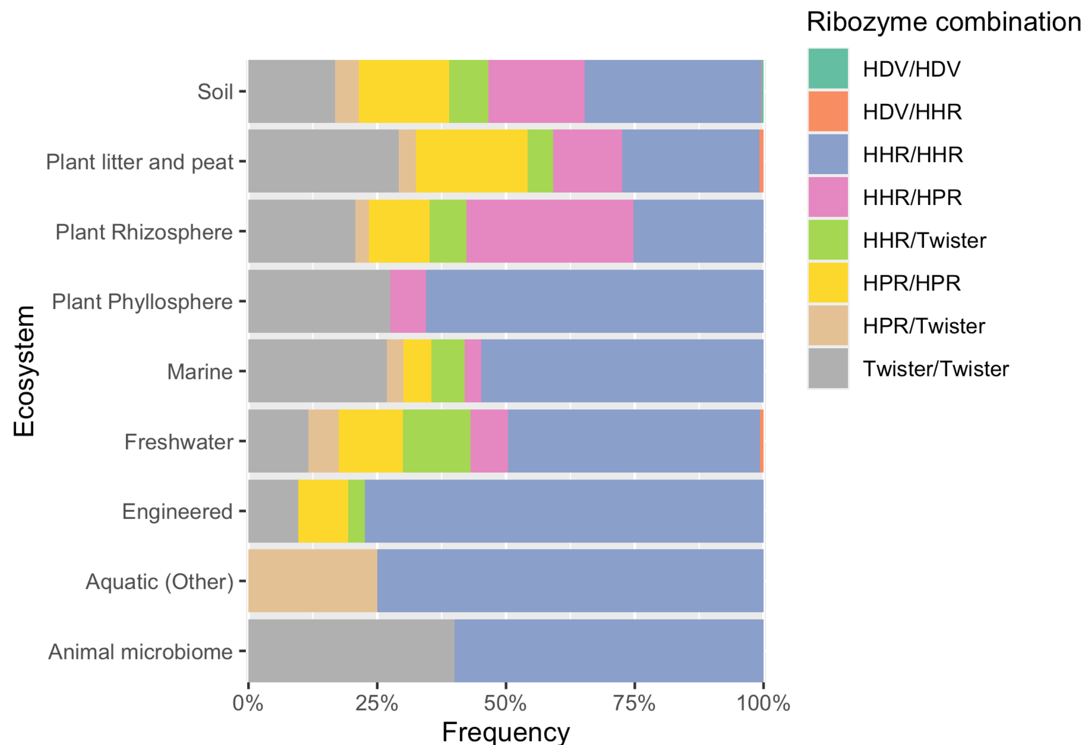


Figure 4.6: Relative frequencies of ribozyme combinations within symmetric cccRNA clusters in each ecosystem type.

metatranscriptomes were dominated by prokaryotic sequences,[55] but still contained at least 1% of contigs affiliated to eukaryotes. Nonetheless, 187 metatranscriptomes in which viroid-like cccRNAs were detected contained $< 0.1\%$ of eukaryotic contigs, suggesting that these elements replicate in either rare and undetected eukaryotes or in some of the much more abundant prokaryotic hosts. Notable among these datasets were the hot spring metatranscriptomes in which CRISPR spacers targeting viroid-like cccRNAs were identified (see Section 4.3.7). The apparent lack of eukaryotic RNA in these samples strengthen the hypothesis of prokaryotic hosts. Additionally, 104 symmetric clusters were found in marine samples which are far beyond the habitation range of the known hosts of viroids and satRNAs. These findings combined with the clusters from prokaryote-dominated samples suggests that the ecological and host ranges of viroid-like agents are far broader than currently appreciated.

[55] Neri et al., "Expansion of the Global RNA Virome Reveals Diverse Clades of Bacteriophages."

4.3.4 Virus-like elements blurring lines between riboviruses, ribozyviruses and viroids

Among the cccRNAs containing symmetric HHR3s, rod-shaped sequences up to 4,705 nt were identified. These sequences were far outside the length range of viroids and ribozyviruses. It was hypothesized that these cccRNAs could be previously uncharacterized ribozy-like viruses. To perform a comprehensive search for potential ribozy-like viruses, all open reading frames (ORFs) longer than 75 codons present in cccRNAs were translated, and the resulting sequences of putative proteins were clustered by amino acid sequence similarity and compared to protein sequence databases. Almost all reliable matches were to virus proteins.

Ambiviruses

One protein cluster showed significant sequence similarity to the predicted RdRPs of a distinct group of ssRNA viruses, ambiviruses, that were recently discovered in fungal isolates and transcriptomes.[56] Ambiviruses have RNA genomes of approximately 4 kb which encompass bidirectional ORFs, one of which encodes a predicted RdRP. To date, ambiviruses have not been reported to be circular.[57]

In the IMG metatranscriptomes, 163 ANI90 clusters (274 cccRNAs total) were found to encode ambivirus-like RdRP (E-values between 1.3×10^{-229} and 8.7×10^{-4}). Notably, these clusters of cccRNAs were also predicted to contain HHR3, HPR-meta1, and CPEB3 ribozymes, including symmetric sequences with different ribozymes in the two RNA polarities. All these sequences were predicted to adopt a rod-like structure in which the two ORFs encoding, respectively, the RdRP and

[56] Suvi Sutela et al., "The Virome from a Collection of Endomycorrhizal Fungi Reveals New Viral Taxa with Unprecedented Genome Organization," *Virus Evolution* 6, no. 2 (July 2020): veaa076, doi:10.1093/ve/veaa076; M. Forgia et al., "Virome Characterization of Cryphonectria Parasitica Isolates from Azerbaijan Unveiled a New Mymonavirus and a Putative New RNA Virus Unrelated to Described Viral Sequences," *Virology* 553 (January 2021): 51–61, doi:10.1016/j.virol.2020.10.008; Riikka Linnakoski et al., "Armillaria Root Rot Fungi Host Single-Stranded RNA Viruses," *Scientific Reports* 11, no. 1 (December 2021): 7336, doi:10.1038/s41598-021-86343-7.

[57] Forgia et al., "Virome Characterization of Cryphonectria Parasitica Isolates from Azerbaijan Unveiled a New Mymonavirus and a Putative New RNA Virus Unrelated to Described Viral Sequences."

an uncharacterized protein are arranged along the rod in the opposite strands (Figure 4.7). These sequences showed varying degrees of terminal overlap, with a median trimmed repeat length of 123 nt. Three representative sequences were recovered with >2000 nt overlaps, of which one was an almost-complete dimer, suggestive of RCR. Three of the ambi-like clusters were detected at very low levels in 10 plant transcriptomes via read mapping. Soil was the predominant ecosystem in which these cccRNAs were found.

The detection pipeline was the run on the 30 ambivirus and ambivirus-like sequences from GenBank. It found significant ribozyme matches in 15 of these sequences, of which 13 contained two predicted ribozymes. Of the remaining 15 sequences, 11 showed ribozyme matches in the expected locations that failed to pass the significance threshold. As in the IMG data, the HHR3 and HPR-meta1 ribozymes are present in both matched and mismatched combinations. Similarly, three of the published genomes (MT354566.2, MN793994.1, and MT354567.1) contain terminal overlaps of 160-250 nt, suggestive of circularity. Furthermore, all known ambivirus and ambivirus-like sequences were predicted to adopt a rod-shaped conformation. Taken together, these observations strongly suggest that ambiviruses comprise a distinct group of ribozy-like viruses that encode a RdRP homologous to the RdRPs of riboviruses.

Mitoviruses

Three cccRNA clusters with significant mitovirus RdRP matches were detected, including two with symmetric ribozymes. The symmetric singletons are 3,283 and 3,058 nt in size and contain matched twister-P1 ribozymes and an HHR3/twister-P1 combination, respectively. The HHR3 aligns to ELVd with 96% identity. A third cccRNA cluster with three members encoding putative mitovirus-like RdRP, of 3,363 nt, contains a similar match to ELVd (including the HHR conserved core) that was not identified as an HHR by either detection method. This cccRNA lacks the HHR core in the opposite polarity but shows weak similarity (E-value = 0.19) to twister-P1. All cccRNAs were detected with >100 nt overlaps and

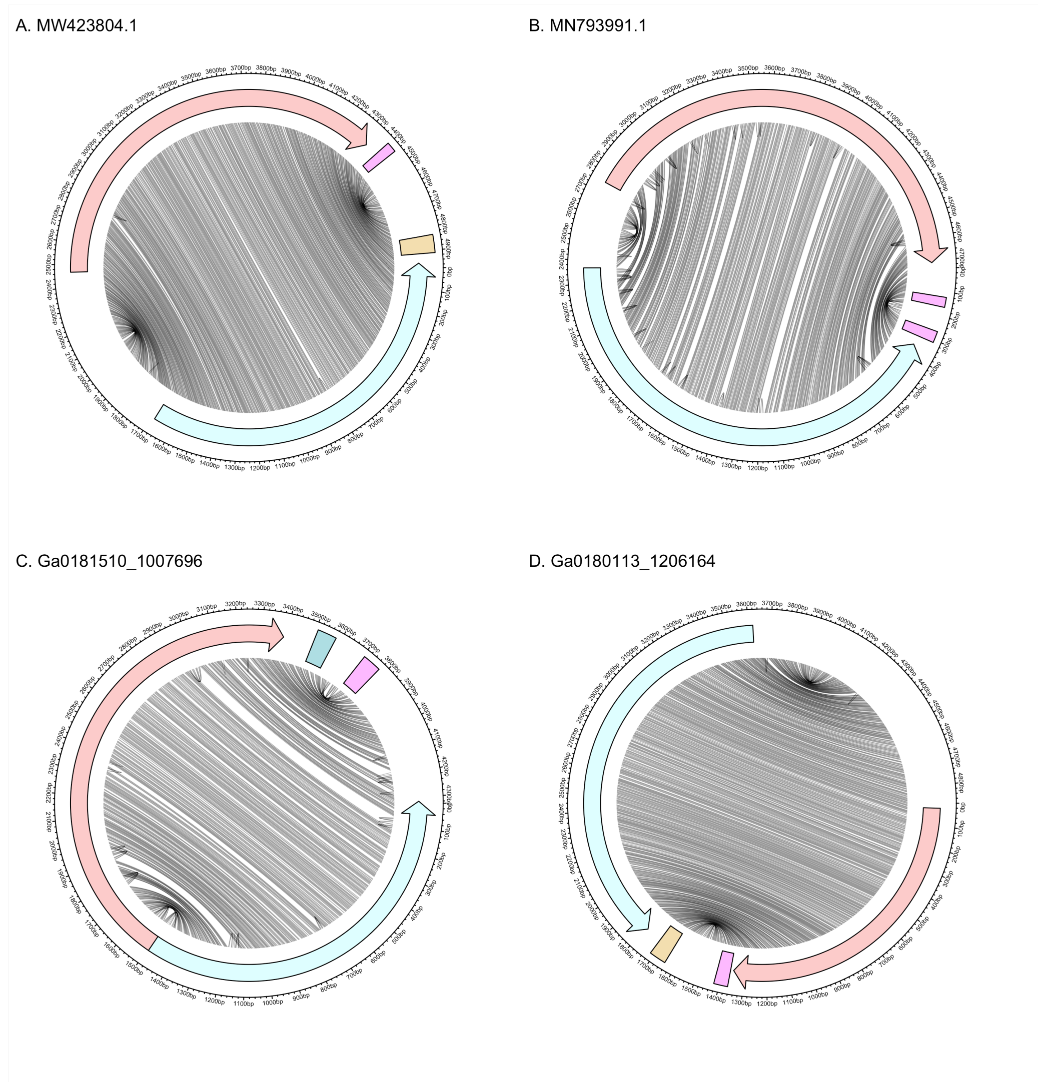


Figure 4.7: Genomic and secondary structure of *Armillaria borealis* ambi-like virus 1 (A), *Tulasnella* ambivirus 1 (B), and ambi-like sequences discovered here (C, D). Red and blue denote (+) and (-) polarities, respectively. Lines connect bases in the genome that are paired in the predicted secondary structure. Arrows represent ORFs and rectangles represent self-cleaving ribozymes. The (+) and (-) ribozymes are HHR3 and HPR-meta1 in (A), HHR3 and HHR3 in (B), CPEB3 and HHR3 in (C), and HHR3 and HPR-meta1 in (D). In all cases, the ribozymes are located outside the ORFs at the end of the rod.

are predicted to adopt a branched conformation with between 63% and 66% of bases paired in both polarities. These three genomes have a low (36–40%) GC content, a hallmark of mitoviruses.[58] Searching the predicted RdRPs against the protein sequence databases yielded the most significant matches for the three cccRNAs to Grapevine-associated mitovirus 13, Grapevine-associated mitovirus 14, and *Fusarium asiaticum* mitovirus 8. Upon closer inspection of Grapevine-associated mitoviruses 11 and 13, it was revealed they also contained HHR3s in both polarities and a twister-P1/HHR3 combination, respectively. Ribozyme searches of all available *Lenarviricota* (taxid 2732407) and unclassified *Riboviria* (taxid 2585030) sequences did not yield other matches besides the ambiviruses and these few mitoviruses.

Putative satellite riboviruses

Apart from the RdRps, 135 sequences were identified comprising 53 ANI90 clusters with significant similarity to capsid proteins of single-stranded (ss) DNA viruses, in particular, circular Rep-encoding single-stranded (CRESS) viruses.[59] The sequences in 50 of these clusters contained predicted ribozymes, and 13 contained HHR3s in both polarities. Two clusters contained paired HHR3 and twister-P1 ribozymes, whereas two other clusters contained symmetric HP-meta1 ribozymes. 26 clusters contained a single HHR3, three a twister-P1, and four a HPR-meta1. 21 clusters, including all three without ribozyme profile matches, produced a nucleotide alignment to a known viroid's ribozyme, ranging in length from 25 to 50 nt at 83–96% identity. The cccRNAs in these clusters varied between 1,092 and 1,632 nt in length, with a mean of 1,317 nt and GC content with the mean of 44%. Four cccRNAs were sequenced as complete head-to-tail dimers. The secondary structures of these cccRNAs included, on average, 66% paired bases. Given the

[58] Armelle Marais et al., "Determination of the Complete Genomic Sequence of Neofusicoccum Luteum Mitovirus 1 (NLMV1), a Novel Mitovirus Associated with a Phytopathogenic Botryosphaeriaceae," *Archives of Virology* 162, no. 8 (August 2017): 2477–2480, doi:10.1007/s00705-017-3338-9.

[59] Mart Krupovic et al., "Cressdnaviricota: A Virus Phylum Unifying Seven Families of Rep-Encoding Viruses with Single-Stranded, Circular DNA Genomes," *Journal of Virology* 94, no. 12 (June 2020): e00582–20, doi:10.1128/JVI.00582-20.

strong evidence of circularity, extensive self-complementarity resulting in predicted branched structure and confident prediction of ribozymes, these cccRNAs most likely represent a novel group of ribozy-like satellite viruses.

4.3.5 Diverse deltavirus-like viruses

In addition to finding sequences that resemble ribozyviruses conceptually (*i.e* protein-coding viroid-like cccRNA but encoded unrelated proteins), actual relatives of deltaviruses were identified. Clusters of ORFs from the identified cccRNAs were compared to the sequences of the HDV antigen (HDVAg) and its homologs from other ribozyviruses. A total of 12 ORF clusters were identified above the HDVAg Pfam profile's gathering threshold; additional 21 representative ORFs were significant at the E-value $< 1 \times 10^{-3}$ level, and 34 at E-value $< 1 \times 10^{-2}$. Of these clusters, only one showed a significant nucleotide alignment to a known HDV-like virus. The other clusters were found in a variety of environments ranging from soil to wastewater to coastal wetland sediment. Samples with matching cccRNAs were collected from as far north as Alaska to as far south as Florida. All 69 members of the clusters encompassing ORFs with HDVAg profile matches (E-value $< 1 \times 10^{-2}$) were predicted to adopt a rod shape in both polarities. The genome size of ribozyviruses in ViroidDB ranges from 1,547 to 1,735 nt. However, among the HDV-like clusters identified in metatranscriptomes, the size ranged from 1,019 nt to 1,757 nt, with a median length of 1,317 nt.

Clustering the HDV-like sequences in combination with the known ribozyviruses in ViroidDB produces no ANI80 clusters with both reference and distinct members. Each of the 26 HDV-like clusters falls below the species demarcation criterion for ribozyviruses (80% nucleotide identity).[60] Clustering the ORFs from both the detected HDV-like sequences and reference ribozyviruses with 60% minimum identity (the genus demarcation criterion) using CD-HIT resulted in 36 clusters,

[60] Jussi Hepojoki et al., "Create One New Realm (*Ribozyviria*) Including One New Family (*Kolmioviridae*) Including Genus *Deltavirus* and Seven New Genera for a Total of 15 Species," Proposal (International Committee on Taxonomy of Viruses, March 2020), http://doi.org/10.13140/RG.2.2.31235.43041.

of which 10 consisted entirely of reference sequences whereas 26 contained only sequences discovered here.

Clustering of the HDVAg sequences and their homologs from other animals and metatranscriptomes with a permissive threshold showed that all previously known HDVAg homologs formed a single tight cluster whereas the metatranscriptome sequences formed multiple smaller clusters and singletons distant from each other and from HDV (Figure 4.8). The conservation profile of the multiple alignment of the HDVAg homologs showed that the dimerisation region and one of the RNA-binding regions were prominently conserved whereas the second RNA-binding region was not (Figure 4.9 and Figure 4.10). The sequences of the distant HDVAg homologs from metatranscriptomes showed low sequence similarity to the previously known HDVAgs, far below the similarity among the latter, with the distributions of percent identities almost non-overlapping (Figure 4.11). Finally, in the phylogenetic tree of the HDVAg homologs, all previously known sequences formed one compact clade, whereas the homologs from metatranscriptomes identified here comprised several remaining clades, with a much greater phylogenetic depth (Figure 4.12).

The nucleotide sequences of these HDV-like cccRNAs formed 26 ANI90 clusters, none of which contained confidently predicted self-cleaving ribozymes. However, 13 of these clusters produced weak ribozyme matches (E-value $< 1 \times 10^{-1}$), and 8 of these were symmetric. Both HHR-like (n=36) and HDV-like (n=26) ribozymes were detected although no clusters contained ribozymes of both types. Of the HDV-like ribozymes detected, only five most closely matched the canonical HDV ribozyme. Ten putative ribozymes showed the strongest similarity to the HDV ribozyme (HDVR) found in the genome of *Faecalibacterium prausnitzii*,[63] seven were most similar to the HDV-like ribozyme found in the genome of *Anopheles gambiae*,[64] and four were most similar to the mammalian CPEB3 ribozyme.[65] The

[63] Chiu-Ho T. Webb et al., "Widespread Occurrence of Self-Cleaving Ribozymes," *Science* 326, no. 5955 (November 2009): 953, doi:10.1126/science.1178084.

[64] Webb et al.

[65] Durga M. Chadalavada, Elizabeth A. Gratton, and Philip C. Bevilacqua, "The Human HDV-like CPEB3 Ribozyme Is Intrinsically Fast-Reacting," *Biochemistry* 49, no. 25 (June 2010): 5321–5330, doi:10.1021/bi100434c; Kourosh Salehi-Ashtiani et al., "A Genomewide Search for

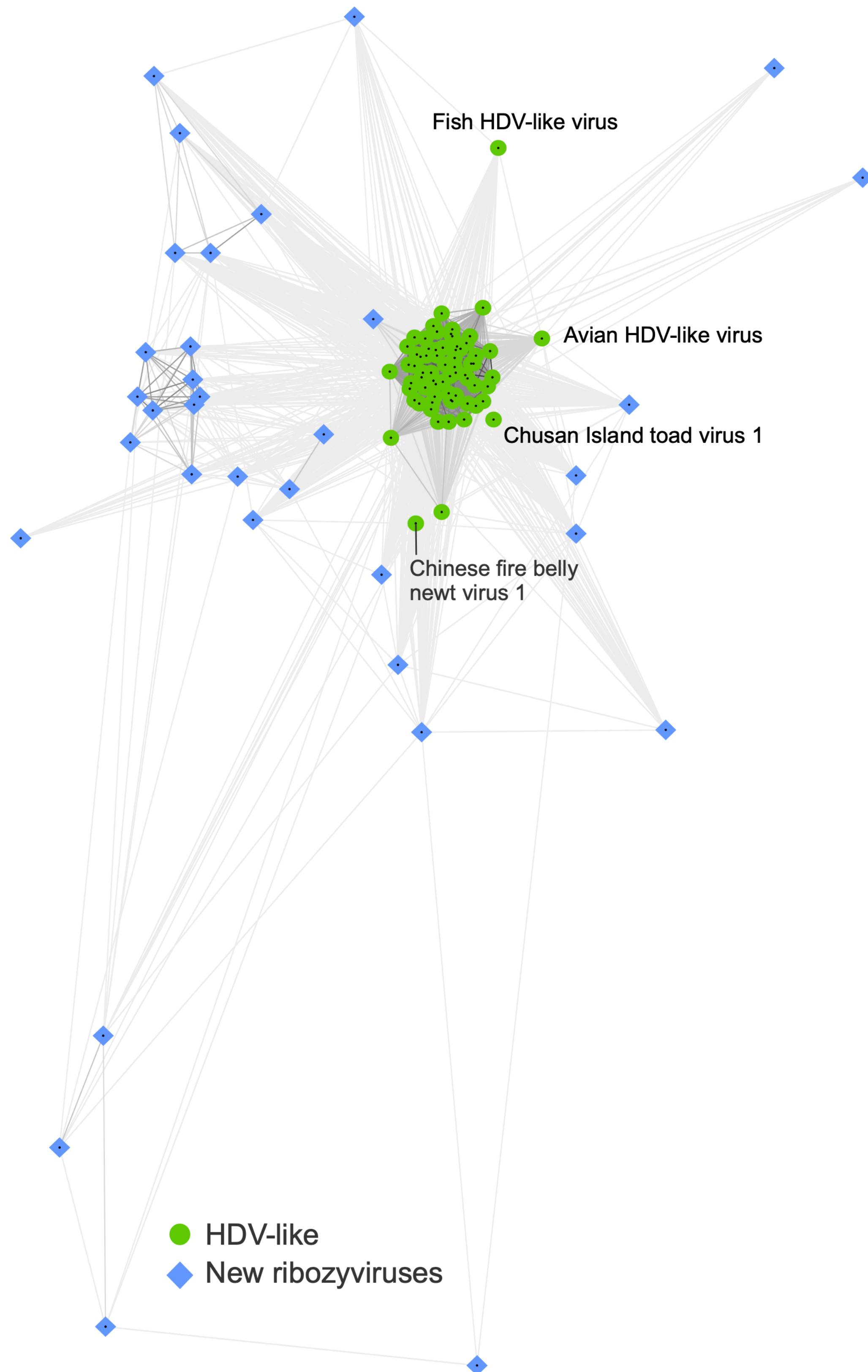


Figure 4.8: Clustering of the HDV antigen (Ag)-like protein homologs based on their sequence similarity. Lines connect nodes (sequences) with P-value $< 1 \times 10^{-5}$. Reference HDVAg-like sequences from GenBank are shown as green circles, whereas those detected in metatranscriptomic datasets as blue diamonds. Some of the divergent reference sequences are labelled.

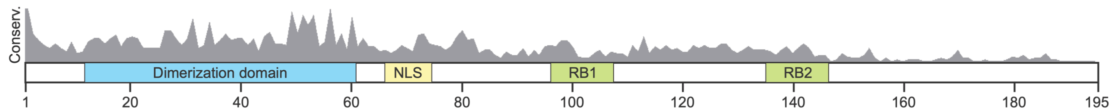


Figure 4.9: Schematic representation of the HDVAg with functionally important regions indicated with colored boxes. RB1 and RB2, RNA-binding sites 1 and 2, respectively; NLS, nuclear localisation signal. Gray histogram shows the sequence conservation (percent identity) of HDVAg-like sequences from metatranscriptomic datasets.

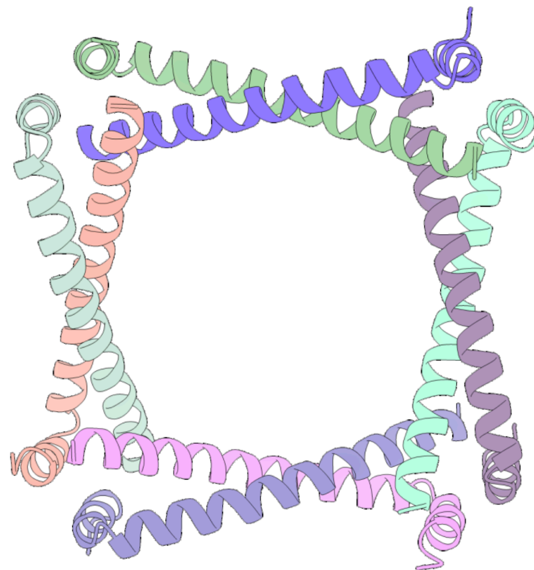


Figure 4.10: Octameric structure of the conserved dimerisation domain of HDVAg. PDB ID: 1A92.[61] Each protein molecule is shown with a different color.

causes of these similarities remain to be investigated; given the small size of the ribozyme, convergence cannot be ruled out.

The limited number of significant ribozyme matches among the HDV-like sequences posed an opportunity for detecting novel ribozymes or diverged variants of known ones. For example, an HDV-like cccRNA cluster (representative member 3300009579_Ga0115599_1049451) was identified with no predicted ribozymes. However, upon closer examination, sequences from this cluster were shown to contain the conserved HHR core in both polarities in the expected locations, a recently discovered ribozyme configuration.[66] Some clusters entirely lacked

Ribozymes Reveals an HDV-Like Sequence in the Human CPEB3 Gene," *Science* 313, no. 5794 (September 2006): 1788–1792, doi:10.1126/science.1129308.

[66]De la Peña et al., "Hepatitis Delta Virus-Like Circular RNAs from Diverse Metazoans Encode

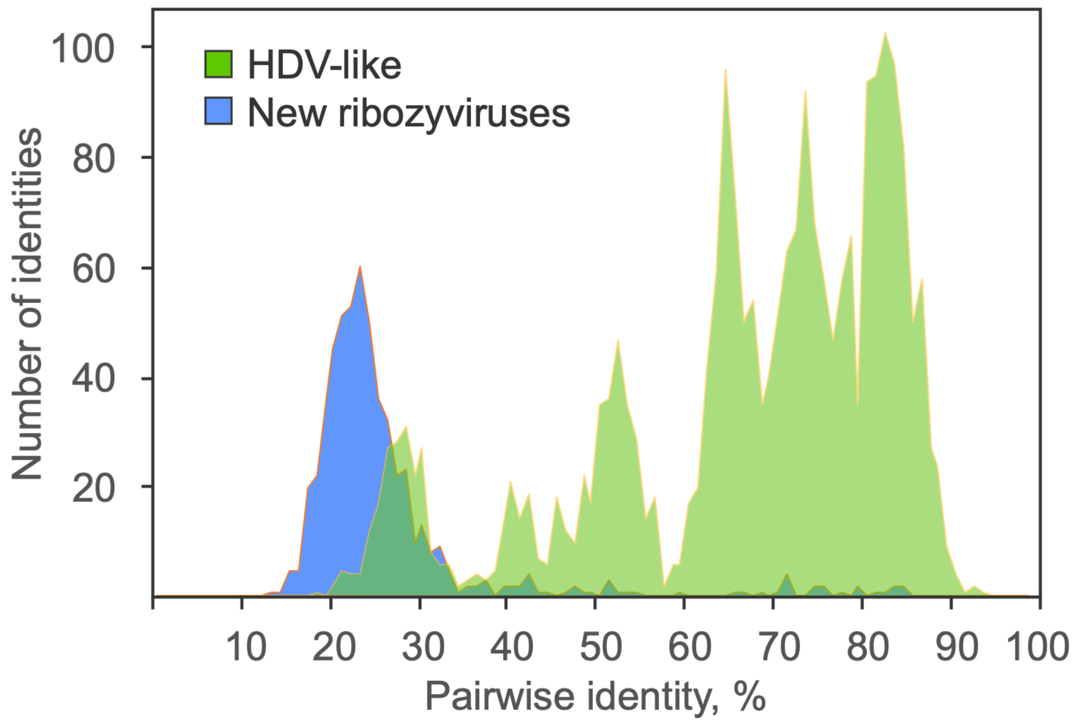


Figure 4.11: Comparison of the sequence conservation among reference HDVAg from GenBank (green) and those from metatranscriptomic datasets (blue).

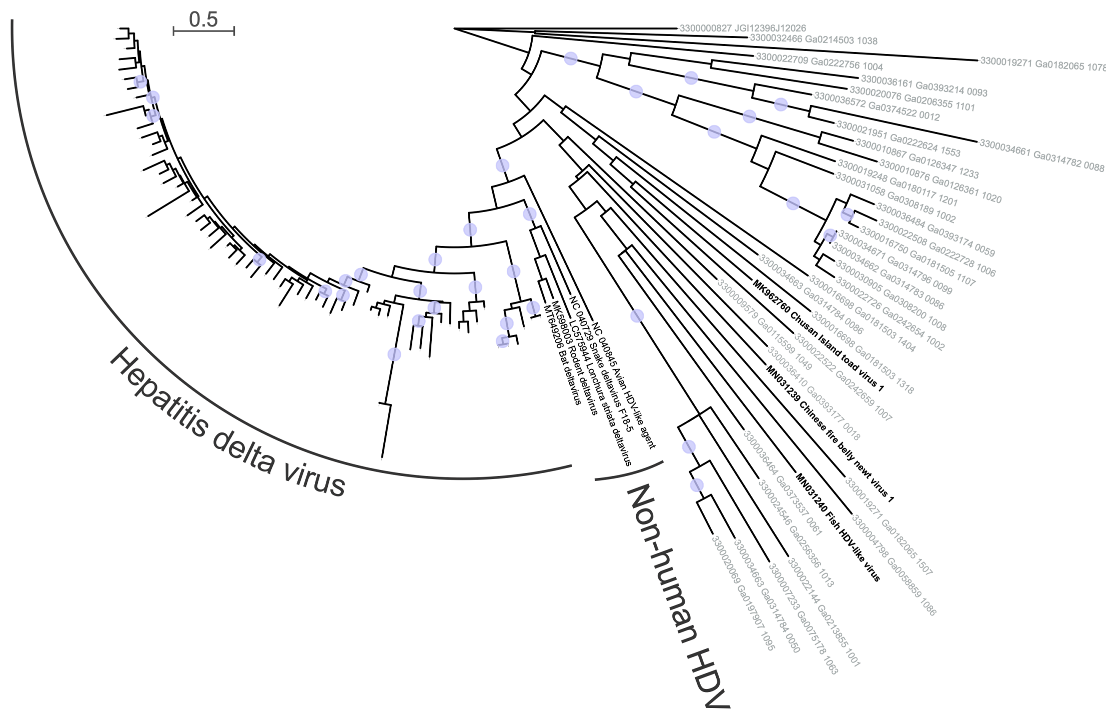


Figure 4.12: Maximum likelihood phylogeny of HDVAg-like sequences. The tree was constructed with IQ-TREE.[62] Circles at the nodes represent SH-aLRT support higher than 90%. The scale bar represents the number of substitutions per site.

the HHR core in either polarity, suggesting the use of alternative, yet unknown ribozymes.

4.3.6 Diverse ribozymes and ribozyme combinations

Almost all viroid-like RNAs described to date contain the same type of ribozyme in both polarities, with the exception of some satRNAs. Surprisingly, many viroid-like cccRNAs identified here were predicted to contain ribozyme combinations that have not been so far reported in replicating cccRNAs (Figure 4.13).

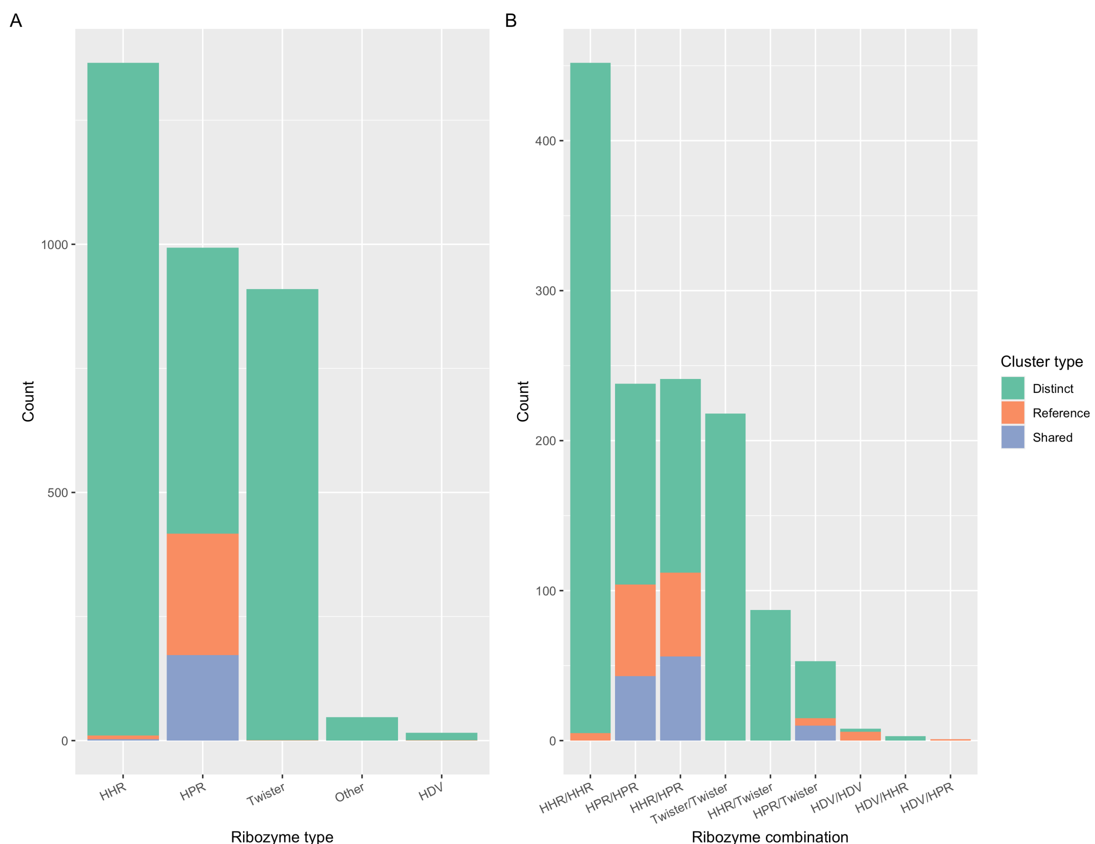


Figure 4.13: Ribozyme diversity in viroid-like cccRNAs. A. Distribution of ribozyme types in asymmetric clusters. B. Ribozyme co-occurrence within the symmetric viroid-like cccRNA cluster representatives derived from metatranscriptomes.

Twister ribozymes

Specifically, numerous cccRNAs were identified containing twister ribozymes, a recently described ribozyme motif that so far has only been found in combina-

Conserved Hammerhead Ribozymes.”

tion with the HP-meta1 ribozyme. Both symmetric (n=381) and asymmetric (n=930) variants are present in the metatranscriptomic cccRNA clusters. Most symmetric twister clusters contained matched twister ribozymes (218 clusters) in both polarities, an unexpected ribozyme combination. In 87 clusters including mitovirus-like and satellite-like cccRNAs, another novel combination of ribozymes was found with HHRs opposite twister ribozymes. The unusual twister ribozyme is widespread in plant transcriptomes, with 59% of the transcriptomes containing reads mapping to a twister-bearing cccRNA. Indeed, three asymmetric cccRNA clusters that contained a twister-P1 ribozyme were recovered from plants.

Symmetric cccRNAs with mismatched ribozymes

In addition to the twister combinations, several other unusual ribozyme combinations were found in symmetric cccRNAs. Previously, HHR3s have been found in conjunction with HDVAg,[67] but have not been reported to be paired with HDV ribozymes. Three clusters in which HHR3s were paired with HDV-type ribozymes, namely, CPEB3 and HDVR *F. prausnitzii* were identified. The CPEB3-HHR3 combination was found in an ambivirus-like sequence, and the two HDVR *F. prausnitzii*/HHR3 clusters were both predicted to adopt rod-shaped structures. One of these, a 1,052 nt singleton, did not match any sequences in ViroidDB, nt, or UniRef90, but in the other cluster (978 nt, two members), the HHR was closely similar to that of *Cryphonectria parasitica* ambivirus 1 (44/50 nt identical), whereas the HDVR *F. prausnitzii* motif (32/33 nt and 41/46 nt) aligned to two chromosomes of the *Vanessa atalanta* butterfly.

Asymmetric cccRNAs with hatchet and pistol ribozymes

Among the asymmetric cccRNAs, two additional types of self-cleaving ribozymes were identified. The hatchet ribozyme was found in 34 ANI90 clusters that ranged from 357 nt to 567 nt in length and came primarily from aquatic metatranscriptomes, in contrast to the general trend among the cccRNAs that derived from soil

[67] De la Peña et al., "Hepatitis Delta Virus-Like Circular RNAs from Diverse Metazoans Encode Conserved Hammerhead Ribozymes."

metatranscriptomes (Figure 4.6). For example, the most diverse of these clusters (440 nt) contained 22 members from aquatic (almost all freshwater) metatranscriptomes sampled from around the United States. For the hatchet clusters, the Rfam profile matches were the strongest among all detected ribozymes (median E-value 1.4×10^{-8}) and, unusually for viroid-like cccRNAs, the GC content was low (median 35%). The predicted structures of these sequences varied from branched to quasi-rod shaped, with a mean of 62% of the bases paired.

The pistol ribozyme was identified in asymmetric cccRNAs. Like the clusters containing the hatchet ribozyme, clusters with the pistol ribozyme were found primarily (9/13) in marine metatranscriptomes ranging from the Antarctic Ocean to the Baltic Sea. The clusters have a slightly lower median profile match E-value (E-value $= 6.4 \times 10^{-5}$) compared to the hatchet ribozyme but, unlike the hatchet ribozyme, have GC content ranging from 33% to 59% (median 49%) more characteristic of viroid-like RNAs. The predicted secondary structures of the pistol-containing cccRNAs were branched, often with several long hairpin structures.

4.3.7 CRISPR spacers matching cccRNAs

CRISPR spacer matches provide one of the most reliable means for assigning hosts to viruses and other mobile genetic elements in prokaryotes, and for differentiating prokaryote-infecting from eukaryote-infecting viruses.[68] This is because CRISPR spacers are acquired from the genomes of the invading elements and are used to target them for destruction. The recent search for riboviruses in the same set of metatranscriptomes that is analysed in this chapter identified multiple spacers matching RNA viruses, resulting in the assignment of several groups of viruses to bacterial hosts including several previously thought to infect eukaryotes.[69] To identify viroid-like agents that potentially might replicate in prokaryotes, the viroid-like cccRNA sequences identified in this chapter were searched against the IMG

[68] David Paez-Espino et al., "Uncovering Earth's Virome," *Nature* 536, no. 7617 (August 2016): 425–430, doi:10.1038/nature19094; Jacob H. Munson-McGee et al., "A Virus or More in (Nearly) Every Cell: Ubiquitous Networks of Virus–Host Interactions in Extreme Environments," *The ISME Journal* 12, no. 7 (July 2018): 1706–1714, doi:10.1038/s41396-018-0071-7.

[69] Neri et al., "Expansion of the Global RNA Virome Reveals Diverse Clades of Bacteriophages."

CRISPR spacer database,[70] and detected 89 spacers with significant matches to viroid-like cccRNAs from 9 clusters.

One spacer was an identical match of 37 nt to a member of a cluster of 16 cccRNAs with prominent viroid-like features. The cccRNAs of this cluster are 315 nt long, contain symmetric HHR3s, and are predicted to adopt a rod shape, with 73% of the bases paired. Unusual for viroids, the cccRNAs comprising this cluster were found in Mushroom Spring, a hot spring at Yellowstone National Park. The matching spacer was also detected in a 60º C hot spring, Great Boiling Spring, albeit more than 800 km away.[71] The repeats in this CRISPR locus were identical to those in the type III CRISPR locus of *Roseiflexus* sp. RS-1, an anoxygenic filamentous bacterium of the phylum *Chloroflexota* that was itself identified in Yellowstone hot springs.[72] Searching for spacers matching all 16 cluster members with more relaxed criteria (*i.e.*, more than 1 mismatch but with E-value $< 1 \times 10^{-5}$), a further 13 nearly identical matching spacers were identified from 8 Yellowstone hot springs samples collected between 2007 and 2017. One spacer (35/37 identities, E-value $= 3 \times 10^{-6}$) included a precise match to the HHR core motif. The repeats from this expanded set of loci all matched those from *Roseiflexus* sp. RS-1. Previously, multiple spacers were identified in *Roseiflexus* sp. RS-1 that matched a group of partiti-like riboviruses that were accordingly assigned to this bacterial host.[73] Apparently, the type III CRISPR system of *Roseiflexus* sp. that encompasses a reverse transcriptase actively incorporates spacers from multiple RNA replicons.

---

[70] Chen et al., "The IMG/M Data Management and Analysis System v.6.0."

[71] Scott C. Thomas et al., "Position-Specific Metabolic Probing and Metagenomics of Microbial Communities Reveal Conserved Central Carbon Metabolic Network Activities at High Temperatures," *Frontiers in Microbiology* 10 (July 2019): 1427, doi:10.3389/fmicb.2019.01427.

[72] Marcel T. J. van der Meer et al., "Cultivation and Genomic, Nutritional, and Lipid Biomarker Characterization of Roseiflexus Strains Closely Related to Predominant In Situ Populations Inhabiting Yellowstone Hot Spring Microbial Mats," *Journal of Bacteriology* 192, no. 12 (June 2010): 3033–3042, doi:10.1128/JB.01610-09; Michael T. Madigan et al., "Diversity of Anoxygenic Phototrophs in Contrasting Extreme Environments," in *Geothermal Biology and Geochemistry in Yellowstone National Park*, ed. William Parks Inskeep and Timothy R. McDermott (Bozeman, MT: Thermal Biology Institute, 2005), 203–219.

[73] Neri et al., "Expansion of the Global RNA Virome Reveals Diverse Clades of Bacteriophages."

The cluster with the most spacer matches—57 matches spanning 26 metagenomes, largely from sludge and bioreactor samples—includes a single cccRNA of 606 nt (recovered as 841 nt) with an asymmetric twister-P5 ribozyme and a predicted branched conformation with 63% of bases paired. Eight spacers, detected in 7 metagenomes, covered the predicted ribozyme region.

The second most frequently matched cluster, also a singleton, was associated with 13 spacers from 10 metagenomes, all from the same location in the Southern Indian Ocean. This cccRNA is 286 nt long (recovered with a 123 nt overlap) and contains a predicted HHRII in one polarity only. Like the most matched singletons, this sequence also is predicted to adopt a branched conformation, with 57% of bases paired. The spacers collectively covered 33% of the sequence but do not include the HHR region.

4.4 Conclusion

Viroids and viroid-like cccRNAs, such as satRNAs and ribozyviruses, are the smallest, simplest known replicators that hijack either a host DNA-dependent RNA polymerase or a virus RdRP for their replication. Given the universality of the cellular transcription machinery across all life forms and the enormous diversity and near ubiquity of RdRP-encoding riboviruses, the narrow diversity and host range of the known viroid-like agents appeared puzzling. It was suspected that viroid-like agents actually could be far more common than presently known and, with this motivation, a collection of more than 5,000 metatranscriptomes for viroid-like cccRNAs was searched.

Millions of putative cccRNAs were identified by searching for signatures of circularity or RCR, namely, the presence of head-to-tail repeats in assembled contigs. Because reads spanning the origin cannot be reconciled with a linear sequence, the assembler produces contigs with the same subsequence repeated at both the end and the beginning.[74] Alternatively, when linear replication intermediates

[74]Qin et al., "Reference-Free and de Novo Identification of Circular RNAs."

containing head-to-tail repeats are sequenced, the ends of the sequences are also repeated. After compensating for the low-fidelity of RNA polymerase II by allowing for up to 5% mismatches in the repeated regions, testing this method on assembled plant transcriptomes demonstrated that known viroids were reliably recovered in the absence of major assembly errors (see Section 3.5).

However, the extensive secondary structure of many viroid-like cccRNAs and the use of poly-A enrichment during RNA isolation prior to sequencing in the 1KP data[75] makes it likely that many viroid-like cccRNAs either were not sequenced at all or were grossly misassembled and thus could not be recognized as circular. Even under a conservative approach, where only predicted cccRNAs containing confidently identified ribozymes counted as "viroid-like", this search resulted in an approximately five-fold increase in the diversity of viroid-like agents. This is most likely a substantial underestimate of the true span of the viroid-like domain of the replicator space because among the millions of the predicted cccRNAs, in which no ribozymes were confidently identified, some, and possibly, many could be viroid-like agents containing unknown ribozymes or lacking ribozymes altogether like pospiviroids. Although perhaps only a coincidence, it is worth noting that a recent analysis of the same collection of metatranscriptomes also yielded an approximately five-fold increase in the diversity of riboviruses.[76] It is further notable that a substantial number of viroid-like cccRNAs are among the most abundant sequences in the respective metatranscriptomes, emphasizing the prominence of these agents in diverse ecosystems.

The majority of the detected viroid-like cccRNAs possessed characteristic features of viroids including the presence of HHR, often in both polarities, and predicted rod-like or extensive branched conformation. However, the search resulted not only in quantitative expansion of viroid diversity but also in qualitatively novel findings, in particular, unexpected ribozyme combinations, such as those including twister,

[75] Marc T. J. Johnson et al., "Evaluating Methods for Isolating Total RNA and Predicting the Success of Sequencing Phylogenetically Diverse Plant Transcriptomes," ed. Christopher Quince, *PLoS ONE* 7, no. 11 (November 2012): e50226, doi:10.1371/journal.pone.0050226.

[76] Neri et al., "Expansion of the Global RNA Virome Reveals Diverse Clades of Bacteriophages."

and ribozymes not previously found in viroid-like RNAs, such as hatchet and pistol. There is little doubt that additional ribozymes and ribozyme combinations in viroid-like cccRNAs remain to be discovered.

Another key finding is the discovery of diverse groups of ribozy-like viruses. Even if perhaps not unexpected, it is notable that the diversity of the ribozyvirus sequences discovered in metatranscriptomes far exceeds that of the previously known HDV relatives including the recently discovered non-mammalian ones. Moreover, many of the identified ribozyviruses lack the HDV ribozyme or even any known ribozymes, suggesting novel replication mechanisms. The host range of the discovered ribozyviruses remains to be explored but, probably, includes non-animal hosts (see discussion below).

In contrast, the demonstration that ambiviruses are actually viroid-like agents and the discovery of viroid-like mitoviruses and satellite viruses was surprising. These three groups of viroid-like agents resemble ribozyviruses in that these are relatively large, protein-coding viroid-like cccRNAs. However, unlike HDV and its relatives, these viroid-like agents are clearly linked to riboviruses through the RdRPs encoded by ambiviruses and mitoviruses, and capsid proteins encoded by satellite viruses. These findings show that combinations of viroid-like cccRNA and protein-coding genes emerged on multiple occasions during evolution. These ribozy-like viruses unrelated to deltaviruses likely evolved through recombination between typical riboviruses and viroids. The implications for virus taxonomy, in particular, whether such viruses should be classified into the existing realm *Ribozyviria* or into the respective divisions of the realm *Riboviria*,[77] or perhaps, into a separate realm, remain to be sorted out.

One of the most interesting but also most challenging problems is the host range of the expanded diversity of viroid-like agents. There is currently no direct computational approach for connecting viroid-like RNAs with specific hosts. Nevertheless, it

[77]Eugene V. Koonin et al., "Global Organization and Proposed Megataxonomy of the Virus World," *Microbiology and Molecular Biology Reviews* 84, no. 2 (May 2020): e00061–19, doi:10.1128/MMBR.00061-19.

appears exceedingly unlikely that all or even the majority of the viroid-like cccRNAs discovered in metatranscriptomes are parasites of plants. Indeed, orders of magnitude more viroid-like cccRNAs in metatranscriptomes were identified than in plant transcriptomes, and most of the analysed metatranscriptomes are dominated by bacteria followed by unicellular eukaryotes. Furthermore, ambiviruses were isolated from fungi,[78] and the demonstration that these are viroid-like agents expands the host range of the latter. A potential prokaryotic connection of viroid-like cccRNAs through CRISPR spacer matches is particularly notable. The detected spacer matches were not numerous but reliable, in particular, because multiple spacers matching the same cccRNA were identified in diverse metagenomes. At least, the typical viroid-like cccRNAs that matched spacers from the reverse-transcribing type III CRISPR system of *Roseiflexus* sp. appear to be strong candidates for novel bacterial parasites. These viroid-like cccRNAs that likely replicate in bacteria merit further, dedicated metatranscriptome and metagenome searches as well as experimental investigation. These findings echo the recent expansion of the bacterial RNA virome through the search of the same metatranscriptome collection and suggest that bacteria might support a much greater diversity of RNA replicators than previously suspected.[79]

This work targets low hanging fruits in the search for viroid-like agents, being limited to the cccRNAs that contain reliably identifiable, known ribozymes or align directly to known viroid-like agents. This conservative approach was adopted purposefully, in order to avoid potential artifacts resulting from erroneous identification of cccRNA, contamination with DNA-encoded sequences or other sources. A potential opportunity for the discovery of a far greater diversity of viroid-like agents and a challenge for further research is a comprehensive analysis of the massive set of predicted cccRNAs that lack known ribozymes. Computational

[78]Sutela et al., "The Virome from a Collection of Endomycorrhizal Fungi Reveals New Viral Taxa with Unprecedented Genome Organization"; Forgia et al., "Virome Characterization of Cryphonectria Parasitica Isolates from Azerbaijan Unveiled a New Mymonavirus and a Putative New RNA Virus Unrelated to Described Viral Sequences"; Linnakoski et al., "Armillaria Root Rot Fungi Host Single-Stranded RNA Viruses."

[79]Neri et al., "Expansion of the Global RNA Virome Reveals Diverse Clades of Bacteriophages."

methods for *de novo* prediction of ribozymes need to be developed to advance such analyses. Furthermore, there are numerous additional metatranscriptomes, in particular, those recently analysed by the Serratus[80] and Tara[81] teams, as well as numerous animal metatranscriptomes, that should be searched for viroid-like cccRNAs. To make a comprehensive search practicable, more efficient algorithms for circularity detection are required. A technical limitation is that some of the software used to assemble metatranscriptomes would cut circular molecules to unit length excluding terminal repeats. Although the method used for cccRNA detection took into account even short repeats, this feature of some assemblers could yield false negatives. Evidently, the computational approaches applied in this work only identify candidates for viroid-like agents. Experimental validation is needed and is especially important in the case of putative cccRNAs lacking known ribozymes.

[80] Edgar, Taylor, Lin, et al., "Petabase-Scale Sequence Alignment Catalyses Viral Discovery."

[81] Zayed et al., "Cryptic and Abundant Marine Viruses at the Evolutionary Origins of Earth's RNA Virome."

5

An advanced tool for circular sequence analysis

Summary

An efficient and flexible toolkit for circular sequence detection and manipulation, circKit, is introduced. While designed to make viroid-like RNA detection and characterisation faster and easier, the software is agnostic to nucleic acid type and size. It enables circular sequence detection, monomerisation, canonicalisation, deduplication, and ORF finding. Built in the Rust programming language, the tool also supports native multithreading, instruction-level parallelism, and data compression. This tool aims to be of wide applicability as the extent of the circular RNA world expands into the future.

5.1 Introduction

The `vdsearch` pipeline introduced in Chapter 3 is a powerful tool for identifying viroids and viroid-like sequences directly from assembled contigs. However, to scale the pipeline even further, consideration of the stages of the pipeline is necessary.

Specifically, the stages must be examined with respect to raw both computational performance and sensitivity.

The pipeline is effectively a series of filters, each of which is designed to remove sequences that are not viroid-like (see Figure 5.1 for a conceptual schematic). At the start, circular sequences are identified, and then the sequences are monomerized and deduplicated. Then, a variety of tools such as Infernal[1] and MMseqs2[2] are used to identify viroid-like sequences among these putative circular sequences. Finally, ORFs are predicted such that any encoded proteins can be examined.

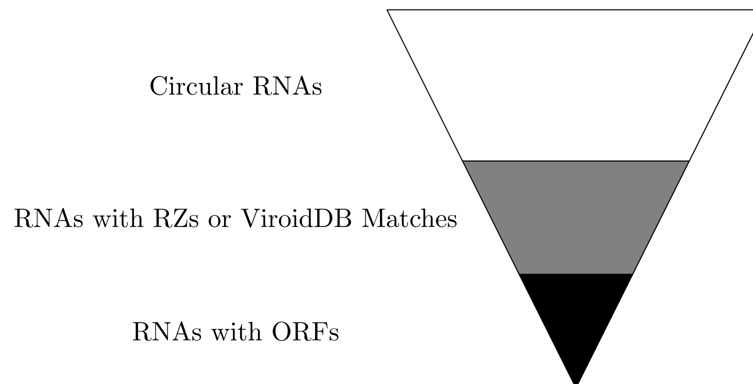


Figure 5.1: A conceptual schematic of viroid identification pipeline

The first layer, circular sequence identification, is run over a billion times in Chapter 4. Here, even a small improvement in performance can result in a large overall speedup and a small improvement in sensitivity can result in a large increase in the number of sequences that can be examined by the rest of the pipeline. The middle section of Figure 5.1, *i.e.*, sequence-sequence and sequence-profile search, uses tools that have been developed over decades and are highly optimized. The final layer, ORF prediction, is relatively simple but, as will be discussed in Section 5.3.3, contains subtleties that can be exploited for performance and sensitivity improvements.

The implementation of the circular sequence detection and monomerisation algorithm as used in Chapter 4 was written in Nim and was the result of a tradeoff

[1] Nawrocki and Eddy, "Infernal 1.1."

[2] Steinegger and Söding, "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets."

between performance and ease of implementation. Python was too slow to be practical, and so the Nim programming language was chosen for its ease of interoperability with Python and syntactical similarity. However, Nim's multithreading support is not mature (nor has it matured in the years since `vdsearch` was created), and thus the original implementation was unable to take advantage of multiple cores. While this proved sufficient for the IMG search in Chapter 4, larger searches would require significantly greater performance.

In addition, the Nim implementation was not as sensitive as it could be. The algorithm was designed to be fast and simple, and therefore left out a number of potential circular sequences that would require more complex and time-consuming algorithms to detect.

Finally, the integration with Python was not as seamless as was originally hoped. While Nim was able to be used directly from Python, this approach resulted in a number of challenges. Packaging the Nim code with the Python code made distribution considerably more difficult and resulted in numerous installation issues reported by users [3]. A better architecture would have the `vdsearch` package as a standalone Python package with the high-performance code as a separate package that can be installed as a dependency. Other tools that the pipeline uses are installed separately and interfaced with via the command line, and so this would be a more consistent approach.

Logically, this architecture makes sense as well. The algorithms are completely independent of the viroid use case and could be used for any research requiring circular sequence detection and manipulation. Distributing the algorithms as a standalone package would allow for a wider audience than the small (yet growing) viroid research community. Thus, the `circKit` package was born.

`circKit` is a Rust library and command-line tool that provides high-performance circular sequence manipulation algorithms. It aims, in order of priority, to be

[3] I would like to express my gratitude to those who reported issues, even though I was not especially quick to resolve issues, unfortunately.

correct, fast, and easy to use. The library is designed to be used as a crate[4] (*i.e.*, a software module integrated into other Rust software projects directly) and as a command-line tool.

5.2 Language choice: Rust

Rust is a systems programming language that provides memory safety without garbage collection. In concrete terms, this means that Rust programs are as fast as C and C++ programs, but without the risk of memory leaks or segfaults. Rust achieves this by using a sophisticated type system and a strict ownership model that ensures that memory is freed when it is no longer needed. Memory-unsafe operations, such as dereferencing a null pointer or writing to an array out of bounds, are caught at compile time. These memory safety features enable Rust to have robust support for memory-safe multithreading and parallelism, making it an ideal choice for high-performance bioinformatics applications.[5] Finally, Rust has a modern package management tool, `cargo`, that makes it easy to distribute and install Rust packages and their dependencies. A mature interface for Python, `pyo3`, allows Rust code to be used directly from Python, making it easy to integrate Rust code into existing Python projects [6].

Rust was a natural choice for the `circKit` project. The language's strong support for multithreading and parallelism would allow for the algorithms to take advantage of modern hardware, and the language's strong emphasis on correctness and testing would ensure that the algorithms are implemented correctly. Finally, the language's modern package management tool would make it easy to distribute the algorithms to a wide audience.

[4] A "crate" is Rust's term for a library or package distributed via the central registry, crates.io.

[5] Johannes Köster, "Rust-Bio: A Fast and Safe Bioinformatics Library," *Bioinformatics* 32, no. 3 (February 2016): 444–446, doi:10.1093/bioinformatics/btv573.

[6] Such was the promise of Nim too, so skepticism towards this claim is warranted. However, the PyO3 Github repository has over 10,000 stars at the time of this writing (March 2024), compared to 804 for Nimporter. More tellingly, the Nimporter PyPI package was downloaded 1,345 times in the past month compared to 978,777 for Maturin, the PyPI package for PyO3.

5.3 Algorithms implemented

The `circKit` package provides a number of high-performance algorithms for circular sequence manipulation. Each of these algorithms is separately implemented and tested. The `circKit` command-line interface is shown in Figure 5.2.

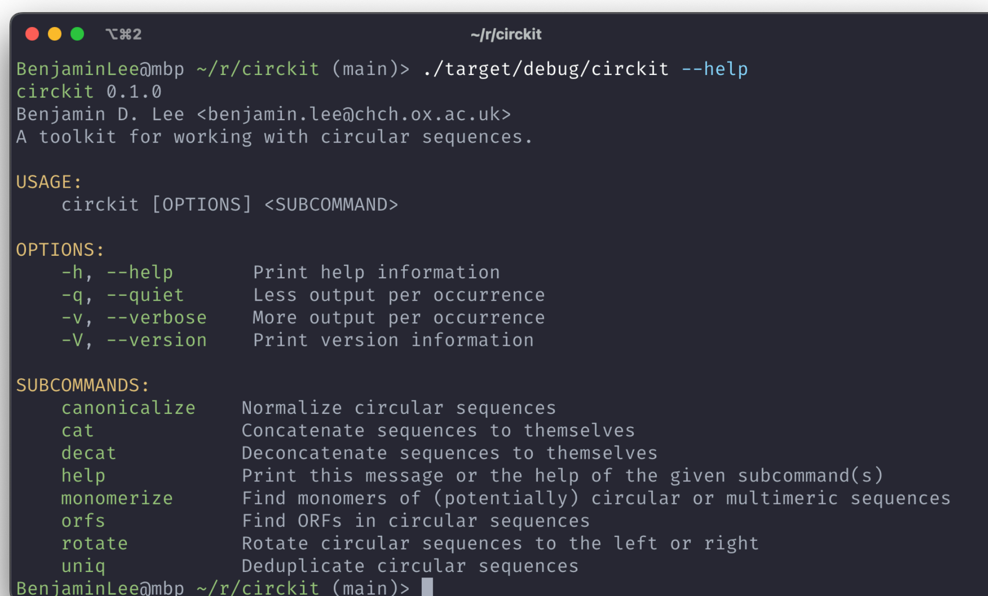


Figure 5.2: The `circKit` command-line interface

5.3.1 Circularity detection and monomerisation

The key motivation for the `circKit` package is the implementation of a high-performance circular sequence detection and monomerisation algorithm. In essence, the algorithm implemented here is the same as the one created by Gao, Wang, and Zhao[7] and used in Chapter 4, but with a number of improvements.

All potential seed matches considered

For performance reasons, the previous implementation only considered the first potential monomer for each sequence during each monomerisation iteration. This limitation results in the inability to monomerise sequences in which the starter seed

[7] “CIRI.”

is present again within the monomer. The implementation in `circKit` removes this limitation and is therefore able to recognise a greater number of circular sequences.

Consider the following sequence:

```
>3300000336_thermBogB3DRAFT_134832
ACACCAGTGTTCCACTAGGAGGAGAGCTGGGCAGACTAAA
CACCAGTGTTCCACTAGGAGGAGGAGAGCTGGGCAGACTA
AACACCAGTGTTCCACTAGGAGGAGAGCTGGGCAGACTAA
ACACCAGTGTTCCACTAGGAGGAGGAGAGCTGGGCAGAC
```

The previous implementation, as described in Chapter 3, would have been unable to monomerise this sequence. Using the final 10-mer as a seed, the first potential overlap would not be valid under a strict identity cutoff. Here, the last 36 nucleotides are compared to the first 36 nucleotides:

```
ACACCAGTGTTCCACTAGGAGGAGAGCTGGGCAGAC
CCAGTGTTCCACTAGGAGGAGGAGAGCTGGGCAGAC
 **    *   * *  ********************
```

However, the second potential overlap would certainly be valid:

```
ACACCAGTGTTCCACTAGGAGGAGAGCTGGGCAGACTAAA
ACACCAGTGTTCCACTAGGAGGAGAGCTGGGCAGACTAAA
****************************************

CACCAGTGTTCCACTAGGAGGAGGAGAGCTGGGCAGAC
CACCAGTGTTCCACTAGGAGGAGGAGAGCTGGGCAGAC
**************************************
```

After removing the terminal repeat, the monomerisation process begins again on this sequence:

```
>3300000336_thermBogB3DRAFT_134832
ACACCAGTGTTCCACTAGGAGGAGAGCTGGGCAGACTAAA
CACCAGTGTTCCACTAGGAGGAGGAGAGCTGGGCAGACTA
A
```

Performing the search again for the final 10-mer, the overlap is invalid:

```
ACACCAGTGTTCCACTAGGAGGAGAGCTGGGCAGACTAA
CCAGTGTTCCACTAGGAGGAGGAGAGCTGGGCAGACTAA
 **    *   * *  ***********************
```

Careful examination of the input sequence will reveal that it is in fact a 4-mer, not a 2-mer. The monomerisation algorithm only removes two of these due to a three nucleotide deletion in the first and third monomers relative to the second and third.

```
AGTGTTCCACT---AGGAGGAGAGCTGGGCAGACTAA
AGTGTTCCACTAGGAGGAGGAGAGCTGGGCAGACTAA
***********   ***********************
```

However, because the algorithm relies on the Hamming distance rather than the Levenshtein distance, insertions and deletions have outsized effect on the computed distance. This example demonstrates the improvements in the `circKit` implementation over the previous implementation as well as the intrinsic limitations of the Hamming distance for this application.

Both strands considered

The previous implementation only considered the forward strand of the sequence. The high mutation rate of viroids and viroid-like RNAs means that mutations in the seed used for potential overlap identification are not uncommon. Rather than use an approximate string matching algorithm such as a bounded Levenshtein distance, the `circKit` implementation simply considers both strands of the sequence. If one strand does not yield a valid overlap, the other strand is considered. Only when both strands fail to yield a valid overlap is the sequence considered to be linear (if no overlaps have been found) or a monomer (if at least one overlap has been found). This “sensitive” mode is not enabled by default, but can be enabled with the `--sensitive` flag.

Consider this sequence:

```
ATGCCCATGCGCCAGCGCAAATGCCCATGCGCCAGCGCAG
```

Using a seed length of 10, the terminal 10-mer is not found elsewhere in the sequence. Consider its reverse complement, with line breaks added for clarity:

```
CTGCGCTGGCGCATGGGCAT
TTGCGCTGGCGCATGGGCAT
```

Now, the terminal 10-mer is found elsewhere in the sequence. Indeed, the sequence is an almost perfect tandem repeat. Using sensitive mode, the sequence is correctly identified as circular and monomerized.

Customizable overlap filtering

The new implementation allows for various overlap filtering options. An overlap can be filtered by absolute size, relative size, or both. In addition, it can be filtered for a minimum identity or a maximum number of mismatches. These parameters make the algorithm more flexible and allow for the detection of a wider range of circular sequences.

Single Instruction, Multiple Data (SIMD) acceleration

The new implementation uses SIMD acceleration to speed up the monomerisation process. SIMD is a type of parallel processing that allows a single CPU instruction to perform the same operation on multiple data points simultaneously. Modern CPUs have SIMD instructions that can operate on 128, 256, or 512 bits of data at a time, depending on the CPU architecture. The `circKit` implementation uses the Rust-Bio[8] crate to take advantage of these SIMD instructions, resulting in a significant speedup in the seed search phase of monomerisation.

5.3.2 canonicalisation and deduplication

The `circKit` package also provides an algorithm for canonicalizing and deduplicating sequences. This algorithm, introduced in Section 3.3.2, is able to create a single canonical representation of a sequence and, with a hash table, deduplicate sequences.

[8] Köster, "Rust-Bio."

The main difference between the `circKit` implementation and the previous implementation is that the `circKit` implementation uses a linear-time algorithm for canonicalisation. The previous implementation is $O(n^2)$, where n is the length of the sequence, and so is not practical for large sequences. The `circKit` implementation is $O(n)$ and is able to handle sequences of any length. It is able to achieve this improvement by implementing Duval's algorithm,[9] which is a linear-time algorithm for finding the lexicographically smallest rotation of a string.

With respect to deduplication, the primary improvement is the integration with the canonicalisation algorithm. The `vdsearch` implementation first canonicalizes the sequences, writes to disk, reads from disk to `seqkit` for deduplication, and then writes to disk again. The `circKit` implementation is able to canonicalise and deduplicate sequences in a single pass, which is more efficient and requires less disk space. It saves memory by using a hash table to store only the hashes of the canonicalized sequences, rather than the sequences themselves. While the possibility of hash collisions exists, the probability of a collision is low, and the algorithm specifically uses a 64-bit hash to reduce the probability of a collision.

5.3.3 ORF prediction

The `circKit` package also provides an algorithm for predicting open reading frames (ORFs) in a circular sequence. While a simple implementation using concatenation is described in Section 3.3.7, this implementation is capable of identifying *all* ORFs in a sequence.

Finding an ORF that is broken by the point of linearisation is relatively simple. Indeed, the concatenation approach will find all such ORFs. Where the ORF finding problem becomes more difficult is in finding ORFs containing overlapping reading frame, as is present in satRYMV[10] (discussed in Section 1.2).

[9] Duval, "Factorizing Words over an Ordered Alphabet."

[10] AbouHaidar et al., "Novel Coding, Translation, and Gene Expression of a Replicating Covalently Closed Circular RNA of 220 Nt."

Frame 0	3	12	36	98	126	132	150	159
Frame 1	7	19	49	70	124	148		
Frame 2	11	14	23	74	131	140	155	

Figure 5.3: A jagged stop codon index array. Note that frames are numbered 0, 1, and 2 to correspond to each index's remainder (*i.e., modulus*) when divided by 3.

The algorithm for finding circular ORFs starts by constructing a jagged array of stop codon indices. A jagged array is similar to a normal multidimensional array or matrix except that the lengths of each row are variable.[11] In each of the three arrays, numbered 0, 1, and 2, indices of the first positions of stop codons are pushed in the order that they appear (see Figure 5.3 for a visual depiction). Such an index can be constructed in $O(n)$ time. From this jagged stop codon index array, it is possible to find all possible ORFs, wrapped or otherwise by searching for start codons in any frame.

Consider the simple case where the sequence length modulus 3 is 0 (*i.e.*, the sequence length is divisible by 3). If an ORF were to wrap around the sequence, it would be in the same reading frame that it started. A self-overlapping ORF such as the one in satRYMV is not possible in this case.

In Figure 5.3, a stop codon index array is presented. For example, imagine a start codon were found starting at position 6. Since six is divisible by three, it is in frame 0. The next greater index in this frame is 12, so this putative ORF runs from the codon starting at position 6 to the codon starting at position 12. If a start codon were found at position 162 (also in frame 0), there is no stop codon at

[11] Internally, the jagged array is simply an array of pointers to three different length arrays.

Table 5.1: Next codon stop index search frame when wrapping around

(a) Sequence length modulus three equals one

Current Frame	Next Frame
0	2
1	0
2	1

(b) Sequence length modulus three equals two

Current Frame	Next Frame
0	1
1	2
2	0

a greater index. However, since the ORF may wrap around the sequence, it is also required to check stop codon indices in the same frame that occur *before* the index of the start codon. In the example of Figure 5.3, there is a stop codon at position 3, so the ORF would span the origin from positions 162 to 3. Start codons with positions modulus 3 equal to 1 and 2 would have their stop codons searched in the same manner in frames 1 and 2, respectively.

Where the ORF finding algorithm becomes more complex is when the sequence's length is not evenly divisible by three. After searching for greater stop codon indices within the same frame, the search must continue in the next frame (Table 5.1). Any stop codon in that frame would constitute a stop codon for the start codon. Should there not be a stop codon in the next frame, the frame after that must be checked for stop codons. If there is still not a stop codon, the same frame as the start codon is in must be checked for stop codons that may have occurred before the start codon. Thus, these checks have time complexity $O(1)$.

Different search patterns must be used depending on the length of the sequence. If the remainder of the sequence's length divided by three is one, the next frame to be searched is the frame *before* the current frame, also wrapping around (Table 5.1a). The opposite search pattern must be conducted if the remainder of the sequence's length divided by three is two (Table 5.1b). Both search patterns terminate when the same frame as the start codon is in is reached and stop codons with lower indices than the start codon's are searched [12].

[12]Despite having lower indices, these stop codons would still occur after the start codon due to wrapping.

Thus, all that is required to efficiently find circular ORFs is to construct the jagged stop codon index array ($O(n)$ time) and then search each start codon (of which there are $O(n)$) for the next valid stop codon index ($O(\log n)$ since binary search could be used on the sorted array). The worse-case time complexity is therefore $O(n \log n)$. Special care must be taken to also consider the two start codons spanning the end of the sequence.

This wrapping procedure demonstrates why simple concatenation is not able to find all circular ORFs. Even creating a trimer and using a traditional linear ORF finding tool would not be completely effective. As noted, an ORF's stop codon may be within the same frame as its start codon but occurring at a lower index. Only a tetramer would be fully effective. However, such methods have several disadvantages. Four times as much memory must be used per sequence. Additionally, special care must be taken to remove duplicate linear ORFs that may be found.

Such methods may be advisable if the were faster than the index-based approach. However, this implementation was four times faster per core at finding *regular* ORFs than the optimized Cython implementation of orfipy.[13] With multithreading, the implementation was eight times faster.

5.3.4 Concatenation and deconcatenation

Many bioinformatics tools assume the input sequences are linear. A common strategy for inputting circular sequences into these tools is to concatenate the sequence with itself. Indeed, just such a trick was used in Chapter 3 for clustering sequences and for performing ORF searches. `circKit` provides the ability to concatenate and deconcatenate sequences. While other tools can be used for concatenation (such as `seqkit`[14] using the `seqkit concat` command), `circKit` is able to perform this operation in reverse as well.

[13] Singh and Wurtele, "Orfipy."
[14] Shen et al., "SeqKit."

5.3.5 Sequence rotation

`circKit` also features the ability to rotate sequences. Like `SeqKit`,[15] users can rotate sequences by a fixed number of nucleotides in either direction. Unlike `SeqKit`, `circKit` automatically wraps the sequence, allowing users to rotate by any number of nucleotides. In addition, `circKit` can also rotate sequences by a given percentage of the sequence length.

5.4 Test suite

Unlike Nim, Rust has packages for property-based testing, which allows for more thorough testing of the algorithms. Property-based testing is a testing methodology in which the test cases are generated automatically based on the properties of the code. For example, a property-based test for canonicalisation would generate a random sequence and test that the canonicalisation of a canonicalized sequence is equal to itself (*i.e.*, that sequence canonicalisation is idempotent). Concatenation and deconcatenation are reciprocal operations, and so a property-based test would generate random sequences, concatenate them, and then deconcatenate the result to ensure that the original sequences are recovered.

`circKit`'s testing suite uses property-based testing to ensure that the algorithms and their implementations are correct. This is especially important for the canonicalisation algorithm, which is heavily optimized for performance at the expense of readability. Property-based testing makes it possible ensure that the optimisations do not affect the correctness of the algorithm by ensuring the output of the optimized algorithm is the same as the output of the un-optimized algorithm for all given inputs.

Many of these tests are "sanity checks" that help build confidence in the correctness of the algorithms. For example, a simple trimer should always be monomerized, so it is trivial to test:

[15] Shen et al.

```
proptest! {
    #[test]
    fn concatenated_monomerizes(input in "[ACGT]{12,100}") {          ①
        let concatenated = format!("{}{}{}", input, input, input); ②
        let m = Monomerizer::builder()
            .overlap_min_identity(0.95)
            .seed_len(10)
            .build()
            .unwrap();
        prop_assert_eq!(                                               ③
            m.clone().monomerize(concatenated.as_bytes()),
            input.as_bytes()
        );
    }
}
```

① The `proptest!` macro is used to define a property-based test. The `[ACGT]{12,100}` regular expression specifies that the input should be a random sequence of A, C, G, and T between 12 and 100 nucleotides long.

② The input is concatenated with itself twice to create a sequence that is known to be a tandem repeat.

③ The `prop_assert_eq!` macro is used to assert that the monomerisation of the concatenated sequence is equal to the original sequence.

Because of the comparative complexity of the canonicalisation algorithm, the property-based tests are especially important for this algorithm. Two additional independent algorithms (albeit with inferior performance characteristics) for finding the lexicographically smallest rotation of a string are used to compare the output of the `circKit` implementation to ensure that it is correct.

5.5 Compressed data I/O

Given the large amount of data that the `circKit` algorithms can process, it is important to be able to read and write compressed data. This was not possible with the Nim implementation, but is possible with the Rust implementation. Specifically, the `circKit` command-line tool can read and write compressed FASTA files using the `gzip`,[16] `bzip2`,[17] `xz`, and `zstd`[18] compression algorithms. To do this, it automatically detects the compression algorithm used (if parsing a file) or requested and (de)compresses the file as necessary. This is made possible by the `niffler` crate (https://github.com/luizirber/niffler), which provides a unified interface for reading and writing compressed files. Currently, compressed output is not supported for table output (*e.g.*, metadata tables from monomerisation and ORF prediction), but this is planned for a future release.

5.6 Conclusion

Future expansions to known viroid and viroid-like RNA sequences will require high-performance algorithms for circular sequence manipulation. Although built with this use case in mind, the `circKit` package is designed to be a general-purpose library for circular sequence manipulation. It is designed to be fast, correct, and easy to use, and is implemented in Rust to take advantage of modern hardware and to be easily integrated into existing bioinformatics pipelines.

The future of the `vdsearch` pipeline will be built on the `circKit` package, enabling the pipeline to scale to larger datasets and to be more sensitive to viroid-like sequences. Analysis of yet larger datasets will be possible, and the biological discoveries that result from these analyses will further our understanding of viroids and viroid-like RNAs.

---

[16] L. Peter Deutsch, "GZIP File Format Specification Version 4.3," Request for Comments (Internet Engineering Task Force, May 1996), doi:10.17487/RFC1952.

[17] Julian Seward, "Bzip2 and Libbzip2," December 2001, https://perso.crans.org/pklein/Info/emacs/manual.pdf.

[18] Yann Collet and Murray Kucherawy, "Zstandard Compression and the Application/Zstd Media Type," Request for Comments (Internet Engineering Task Force, October 2018), doi:10.17487/RFC8478.

6

Summary and implications

It is now clear that the subviral world is much more diverse than was known even a few years ago when this thesis was initiated. This expansion has been confirmed and extended by the work of others, who have discovered new viroid-like RNAs in a variety of hosts. Thus, a summary of the current state of the viroid-like RNA world is in order.

6.1 A new era of viroid-like RNA diversity

In addition to the classes of viroid-like RNA agents known at the start of this thesis (described in Chapter 1), others have now been discovered. This section will briefly describe them in an attempt to paint a fuller picture of the diversity of these agents.

A variety of delta-like viruses, termed epsilon and zeta viruses, have been discovered in another large-scale metatranscriptomic study.[1] These viruses are similar in that they contain ribozymes, encoded proteins, and form rod-like structures. Combined with the expansion of the ribozyviruses described in Section 4.3.5, the known

[1] Edgar, Taylor, Lin, et al., "Petabase-Scale Sequence Alignment Catalyses Viral Discovery."

greater host range,[2] and the confirmation of other ribozyme use,[3] it is safe to say that the delta-like viral world has been thoroughly shaken in the last five years.

A new class of viroid-like elements, termed *obelisks*, has recently been reported in human gut metatranscriptomes.[4] These elements are similar to viroids in that they are circular, single-stranded RNAs that have a predicted rod-like secondary structure. Not all were predicted to contain ribozymes, but when present, they were hammerhead ribozymes. These elements appear to replicate within the microbiome of the human gut and are likely to be the first of many viroid-like RNAs to be discovered in the human metatrascriptome. More research is needed to investigate these elements and their potential impact on human health. It can be stated with certainty that these agents dramatically expand the known diversity of viroid-like RNAs.

Ambiviruses, the rod-shaped viruses encoding bidirectional ORFs, have now had their viroid-like nature confirmed *in vivo*.[5] Furthermore, they have now been placed taxonomically in the realm *Riboviria* within their own new phylum, *Ambiviricota*.[6] Ambiviruses ("ambiviricots" more precisely) now have twenty species across four families. This placement will almost certainly need further changes as the relationship of this virus to the other members of the viroid-like RNA world are elucidated.

[2] Paraskevopoulou et al., "Mammalian Deltavirus Without Hepadnavirus Coinfection in the Neotropical Rodent Proechimys Semispinosus"; Hetzel et al., "Identification of a Novel Deltavirus in Boa Constrictors"; Szirovicza et al., "Snake Deltavirus Utilizes Envelope Proteins of Different Viruses To Generate Infectious Particles"; Chang et al., "Novel Hepatitis D-like Agents in Vertebrates and Invertebrates"; Wille et al., "A Divergent Hepatitis D-Like Agent in Birds"; Bergner et al., "Diversification of Mammalian Deltaviruses by Host Shifting."

[3] De la Peña et al., "Hepatitis Delta Virus-Like Circular RNAs from Diverse Metazoans Encode Conserved Hammerhead Ribozymes."

[4] Ivan N. Zheludev et al., "Viroid-Like Colonists of Human Microbiomes," *Cell* 187, no. 23 (November 2024): 6521–6536.e18, doi:10.1016/j.cell.2024.09.033.

[5] Marco Forgia et al., "Hybrids of RNA Viruses and Viroid-Like Elements Replicate in Fungi," *Nature Communications* 14, no. 1 (May 2023): 2591, doi:10.1038/s41467-023-38301-2.

[6] Massimo Turina et al., "Create One New Phylum, *Ambiviricota*, Including One New Class, One New Order, Four New Families, Four New Genera, and 20 New Species, in Kingdom *Orthornavirae* (Realm *Riboviria*)," Proposal (International Committee on Taxonomy of Viruses, April 2023), https://ictv.global/system/files/proposals/approved/Fungal_and_Protist_viruses/2023.007F.Ambiviricota_nphy.docx.

These discoveries have dramatically expanded the types of viroid-like RNAs that are now known, summarised in Table 6.1. The diversity of these agents is striking, and it is likely that even more types of viroid-like RNAs will soon be discovered.

Table 6.1: The major types of known viroid-like cccRNAs (updated)

Viroid-like RNA	Size	Host	Ribozymes	Coding Capacity
Viroids	246–450 nt	Plants	HHR when present	None
Satellite RNAs	220–457 nt	Plants	HHR or hairpin	None (except satRYMV)
Ribozyviruses	1547–1735 nt	Metazoans	HDVR or HHR	One conserved protein
Epsilon viruses	1000–1800 nt	Invertebrates	HHR	One conserved protein
Zeta viruses	324–789 nt	Unknown	HHR	Two conserved proteins
Retrozymes	174–1116 nt	Eukaryotic genomes	HHR	None
Obelisks	730–1340 nt	Human microbiota	HHR	Two conserved proteins
Retroviroids	275 nt	Carnations	HHR	None
Mitoviruses	2046–4367 nt	Fungi	HHR	RdRP
Ambiviruses	4579–4915 nt	Fungi	HHR	RdRP and another protein

6.2 New vocabulary is needed

Indeed, the expansion in diversity reaches to the point that the term "viroid-like RNAs" is somewhat of a stretch to describe the diversity of these agents. Some of these "viroid-like RNAs" are an order of magnitude larger than viroids, have a clear evolutionary relationship to viruses, and do not infect plants (*e.g.*, ambiviruses). Others are not infectious agents at all, but rather are retrozymes that are integrated into the genomes of their hosts, which also need not be plants.

It may be time to consider using a definition that encompasses the key aspects

of these agents, rather than the unwieldy term "viroid-like RNAs." The current definition of a viroid is:

> Viroids are defined operationally by the ICTV as a type of MGEs that are uncoated, small, circular, single-stranded RNAs that do not encode proteins and do not depend on viruses for transmission, and that replicate autonomously through an RNA-RNA rolling-circle mechanism mediated by host enzymes and, in some cases, by cis-acting hammerhead ribozymes; or MGEs that are derived from a viroid in the course of evolution. Any monophyletic group of MGEs that originates from a viroid ancestor should be classified as a group of viroids.[7]

Perhaps a new definition should be coined. I propose the term *rocirna*, which stands for "**ro**lling-**ci**rcle **RNA**." A new definition could be based on the viroid definition:

> A type of MGEs that are circular, single-stranded RNAs that replicate through an RNA-RNA rolling-circle mechanism.

Such a definition would unify the viroids, satellite RNAs, ribozyviruses, ambiviruses, mitoviruses, and other viroid-like RNAs under a single term. Focusing on the key characteristic of these agents (*i.e.*, rolling-circle replication) would allow for a more concise definition that encompasses the diversity of these agents. Whether this expanded scope is desirable is a matter of opinion, but what is clear is that the current taxonomy is not well-suited to the diversity of the subviral world.

6.3 New taxonomy is needed

One side effect of this massive expansion is that viroid-like RNA taxonomy has been placed in a challenging position. The current ICTV taxonomy rules require that evidence of independent replication be demonstrated before a new viroid species can be confirmed. Hence the myriad of viroid-like RNAs that have been discovered in the past decades cannot be classified as viroids.

[7] Peter J. Walker et al., "Changes to Virus Taxonomy and to the International Code of Virus Classification and Nomenclature Ratified by the International Committee on Taxonomy of Viruses (2021)," *Archives of Virology* 166, no. 9 (September 2021): 2633–2648, doi:10.1007/s00705-021-05156-1.

In contrast, the ICTV rules for virus classification are more flexible, and allow for the classification of viruses based on sequence alone.[8] This is a sensible approach, as it is difficult to demonstrate replication for viruses whose hosts are unknown. Taxonomy can be done based on phylogenetic analysis of the sequences, and the classification can be updated as more information becomes available.

I believe that the viroid taxonomy rules should be updated to allow for the classification of viroid-like RNAs based on sequence alone. Strong viroid-like candidates (*i.e.* with two ribozymes, secondary structure, evidence of circularity, and identification in multiple samples) should be classified as viroids, and the classification can be updated as more information becomes available.

6.4 Can a replicator be both a viroid and a virus?

Under a strict interpretation of the current International Code of Virus Classification and Nomenclature (ICVCN), it does appear that a replicator can indeed be both a viroid and a virus. Mobile genetic elements (MGEs) can meet the virus definition both traditionally (*i.e.*, as a regular virus) or via evidence of descent from a virus. Viroids are not defined as separate from viruses [9] but rather by operational criteria [10]. These operational criteria could be met by an MGE that is descended from a virus (*i.e.*, via the genome reduction pathway discussed in Section 1.7). Hence, a replicator could be both a viroid and a virus.

Whether such a replicator exists is an open question and whether this is a desirable outcome of the taxonomic rules is a matter of opinion [11]. There is precedent for similar challenges in the history ICTV taxonomy, as the deltaviruses meet both the definition of a satellite nucleic acid and a virus. Deltaviruses were classified as

[8] Simmonds et al., "Virus Taxonomy in the Age of Metagenomics."

[9] Satellite RNAs are mutually exclusive with viroids, however.

[10] Strictly interpreting the ICVCN, ribozymes other than the hammerhead are also excluded but this seems readily fixable.

[11] Such an agent should emphatically *not* be called a "virusoid" despite the temptation to combine the terms *virus* and *viroid*. That term is muddied enough.

satellite viruses,[12] though they were promoted to membership in an independent realm, *Ribozyviria*, in 2020.[13] Ultimately, the decision to classify the deltaviruses as viruses was sound, especially given the recent evidence of potential deltavirus independent replication.[14]

Should a replicator be both a viroid and a virus, I would argue that it should be classified as a virus for the purpose of expediency. Virus classification is more mature than viroid classification, and the viroid taxonomy rules are in need of an update.

6.5 Ribozymes as the key to the subviral world

The study of ribozymes is of special importance to the subviral world. Ribozymes are the unifying characteristic of almost all these selfish MGEs with the exception of the pospiviroids. As a conjecture, given the variety of ribozyme-based replicators now known, pospiviroids may have evolved from a ribozyme-based replicator that lost its ribozyme in favor of host-provided cleavage machinery.

Discovering and characterizing new ribozymes will likely be the key to fully charting the circular subviral world. This work has shown that the hammerhead ribozyme is not the only ribozyme that can be found in RNAs most similar to viroids: other ribozymes appear to be common. Similarly, previous work has also shown that the HDV ribozyme is not the only ribozyme that can be found in ribozyviruses.[15] It seems very plausible that there are other ribozyme motifs that have yet to be discovered in use within the subviral world.

[12] Mart Krupovic, Jens H. Kuhn, and Matthias G. Fischer, "A Classification System for Virophages and Satellite Viruses," *Archives of Virology* 161, no. 1 (January 2016): 233–247, doi:10.1007/s00705-015-2622-9.

[13] Hepojoki et al., "Create One New Realm (*Ribozyviria*) Including One New Family (*Kolmioviridae*) Including Genus *Deltavirus* and Seven New Genera for a Total of 15 Species."

[14] Paraskevopoulou et al., "Mammalian Deltavirus Without Hepadnavirus Coinfection in the Neotropical Rodent Proechimys Semispinosus."

[15] De la Peña et al., "Hepatitis Delta Virus-Like Circular RNAs from Diverse Metazoans Encode Conserved Hammerhead Ribozymes."

Luckily, the existence of mixed ribozymes (*i.e.*, with different ribozyme motifs in each polarity) means that the discovery of a new ribozyme in a viroid-like RNA could be relatively easy. One straightforward approach was developed by de la Peña et al.[16], in which homologous locations in ribozyviruses were screened for ribozymes. There were a large number of viroid-like RNAs with only one detected ribozyme, and these could be screened for the presence of other ribozymes in the same location. Barring the existence of clusters containing mixed ribozymes with known ribozyme locations, a general approach to predicting the location of ribozymes within viroid-like RNAs could rely on the secondary structure and location of the known ribozyme. This process could then be performed reiteratively, with the discovery of a new ribozyme leading to the discovery of new viroid-like RNAs with only one detected ribozyme, which in turn could lead to the discovery of new ribozymes.

Since the `vdsearch` pipeline is a filtration pipeline, an increase of even a few percentage points in the number of included sequences has a multiplicative effect on the number of sequences that are included in the final analysis. This means that the discovery of a new ribozyme could lead to the discovery of hundreds of new viroid-like RNAs. This is a powerful incentive to continue searching for new ribozymes, both within the subviral world and beyond.

6.6 The need to analyse more (meta)transcriptomes

A first analysis was done using the One Thousand Plant Transcriptomes Project.[17] It revealed a small number of viroid-like RNAs, but it was clear that the dataset was not large enough to fully capture the diversity of viroid-like RNAs. Scaling up the dataset to include a huge number of metatranscriptomes, each of which had a large amount of sequence diversity, was a logical next step.

[16]

[17] One Thousand Plant Transcriptomes Initiative, “One Thousand Plant Transcriptomes and the Phylogenomics of Green Plants.”

As a note, the very first method of viroid identification attempted in the course of this thesis work was a Nim reimplementation of PFOR algorithm of Wu et al.[18] and Zhang et al.[19] This method, which did work on small RNA datasets simply did not scale up to larger RNAseq datasets. While it cannot be stated with complete certainty that the algorithm itself cannot be made to work at greater scales, it was abandoned in favor of the Cirit,[20] rnaSPAdes,[21] and Infernal[22] approach that proved simpler. These methods *were* able to scale up to the large datasets that were used in this work.

However, they will not be able to scale up to the datasets of the future. Other projects[23] have used massive datasets with impressive results. These larger datasets would push the limits of what is capable with the existing pipeline. While greater parallelism could be achieved with more machines, this type of scaling poses its own difficulties. Fewer machines needed to perform an analysis means more accessible, inexpensive, and straightforward research for those seeking to explore transcriptomes.

Hence, the `circKit` project, described in Chapter 5, was initiated. This tool will allow the vdsearch pipeline to be run on a large number of transcriptomes and metatranscriptomes, which should further understanding of the subviral world. Just such analyses have been done recently to great effect.[24]

Future analyses will also need to include a larger number of metatranscriptomes. Methods for their analysis will need to be able to run on multiple cores, multiple nodes, and in the cloud. These methods will likely need to work on compressed

[18] "Homology-Independent Discovery of Replicating Pathogenic Circular RNAs by Deep Sequencing and a New Computational Algorithm."

[19] "Discovery of Replicating Circular RNAs by RNA-Seq and Computational Algorithms."

[20] Gao, Wang, and Zhao, "CIRI"; Qin et al., "Reference-Free and de Novo Identification of Circular RNAs."

[21] Bushmanova et al., "rnaSPAdes."

[22] Nawrocki and Eddy, "Infernal 1.1."

[23] Robert C. Edgar, Jeff Taylor, Tomer Altman, et al., "Petabase-Scale Sequence Alignment Catalyses Viral Discovery," preprint (Bioinformatics, August 2020), doi:10.1101/2020.08.07.241729; Zheludev et al., "Viroid-Like Colonists of Human Microbiomes."

[24] Zheludev et al., "Viroid-Like Colonists of Human Microbiomes"; Forgia et al., "Hybrids of RNA Viruses and Viroid-Like Elements Replicate in Fungi."

data or on a streaming basis, as the datasets will be too large to fit into memory. In short, viroid biology has entered the big data era, and the methods for its analysis will need to be developed accordingly.

6.7 Viroids as paragons of minimal replicator emergence

As discussed in Section 1.7, the possibility that viroids and other similar cccRNAs are direct descendants of primordial replicators that inhabited the hypothetical ancient RNA world appeared to be remote. While evidence of an expanded host range, both for *bona fide* viroid-like agents (described in Section 4.3.7) and other agents,[25] has provided new support for the RNA world origin hypothesis, it remains unlikely that the plant pathogens are direct descendants of the first replicators. However, the likely relatively recent origins of viroid-like RNAs cannot deprive them of the status of minimal replicators. It does not appear to be a chance coincidence that the Spiegelman monster, the minimal parasite artificially evolved under conditions when the sole selective factor was the speed of replication, is the same size as the smallest viroids and satRNAs.[26] This size, approximately 200 nt, is likely to be close to the ultimate low bound for a replicator.

Replicators are genetic elements that encode information that is necessary and sufficient for their replication but lack the full complement of genes required for providing energy and building blocks for replication.[27] For energy and building blocks, replicators depend on reproducers, biological entities whose reproduction

[25] Zheludev et al., "Viroid-Like Colonists of Human Microbiomes."

[26] Mills, Kramer, and Spiegelman, "Complete Nucleotide Sequence of a Replicating RNA Molecule."

[27] Richard Dawkins, *The Selfish Gene* (New York: Oxford University Press, 1976); Eörs Szathmáry and John Maynard Smith, "From Replicators to Reproducers: The First Major Transitions Leading to Life," *Journal of Theoretical Biology* 187, no. 4 (August 1997): 555–571, doi:10.1006/jtbi.1996.0389; Eörs Szathmáry, "The Evolution of Replicators," ed. B. Charlesworth and P. H. Harvey, *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 355, no. 1403 (November 2000): 1669–1676, doi:10.1098/rstb.2000.0730; Eugene V. Koonin and Petro Starokadomskyy, "Are Viruses Alive? The Replicator Paradigm Sheds Decisive Light on an Old but Misguided Question," *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 59 (October 2016): 125–134, doi:10.1016/j.shpsc.2016.02.016.

involves physical rather than only informational continuity as in the case of replicators.[28] The paradigmatic reproducers are, obviously, cells—*omnis cellula e cellula*, according to Virchow. More precisely, all cellular life forms can be conceived as symbioses of a reproducer and a replicator, the genome.[29] Cellular genomes are ultimate cooperative replicators, but generally, replicators span wide ranges of replicative autonomy and aggressiveness/cooperativity with respect to the host reproducer.[30]

Some replicators, such as many large viruses or self-synthesizing transposons, encode a (nearly) complete replication and sometimes transcription machinery, and thus possess a high degree of replicative autonomy. Others rely mostly or completely on the host replication and expression apparatus. Viroids are the extreme manifestation of the latter strategy—arguably, the ultimate parasites. Generally, the information content of replicators splits into "replicase expression signal" (RES), which can consist of one or more genes encoding replicative enzyme(s) and possibly various accessory proteins, and "replicase recognition signal" (RRS), noncoding elements that are necessary and sufficient to hijack the host replication (or expression) machinery.[31] Viroids have adopted the ultimate parasitic strategy, namely, selection for maximum efficiency of RRS accompanied by complete elimination of RES, under the reduction scenario or by never acquiring RES, under the *de novo* scenario. More precisely, this is the evolutionary strategy of pospiviroids, whereas avsunviroids add a notable twist by engaging viroid-embedded ribozymes in some stages of replication. The HHRs appear easy to evolve[32] and could be the

[28] Szathmáry and Maynard Smith, "From Replicators to Reproducers"; Koonin and Starokadomskyy, "Are Viruses Alive?"

[29] Shelley D. Copley, Eric Smith, and Harold J. Morowitz, "The Origin of the RNA World: Co-Evolution of Genes and Metabolism," *Bioorganic Chemistry* 35, no. 6 (December 2007): 430–443, doi:10.1016/j.bioorg.2007.08.001.

[30] Koonin and Starokadomskyy, "Are Viruses Alive?"; Matti Jalasvuori and Eugene V. Koonin, "Classification of Prokaryotic Genetic Replicators: Between Selfishness and Altruism," *Annals of the New York Academy of Sciences* 1341 (April 2015): 96–105, doi:10.1111/nyas.12696.

[31] Eugene V. Koonin, Yuri I. Wolf, and Mikhail I. Katsnelson, "Inevitability of the Emergence and Persistence of Genetic Parasites Caused by Evolutionary Instability of Parasite-Free States," *Biology Direct* 12, no. 1 (December 2017): 31, doi:10.1186/s13062-017-0202-5.

[32] Salehi-Ashtiani and Szostak, "In Vitro Evolution Suggests Multiple Origins for the Hammerhead Ribozyme."

simplest evolutionary step towards increased replicative autonomy.

Evolution towards ultimate parasitism occurs already at the stage of retrozymes as well as other non-autonomous transposons that hijack the replication and transposition machinery of autonomous transposons with closely similar RRS. However, retrozymes have only limited replication capacity restricted to copy-paste transposition within the host genome. Viroids likely escaped from the host genomes by evolving the ability to redirect the host DdRP towards viroid RNA replication. Identification of an intermediate on the proposed evolutionary path from retrozymes to viroid-like agents—namely, a retrozyme capable of replicating autonomously within a cell—would be a major step in demonstrating the emergence of viroid-like replicators as an ongoing process. There are some indications that retrozymes might possess that capacity, in particular because multimeric intermediates of both polarities have been detected and the high error rate of retrozymes suggests polymerisation by DdRP, but definitive experimental evidence is still lacking.[33]

Although viroids are unlikely to be relics of the pre-cellular stage of life evolution, they provide a unique window into the origin of parasitic replicators, which most likely occurred on multiple independent occasions, via retrozyme escape and, possibly, *de novo*. Similar events could have transpired during the pre-cellular stage of evolution where ultimate parasites similar to viroids with respect to size and reproduction cycle, but not necessarily secondary structure, could have evolved from autonomous replicators. The variety of viroid-like RNA agents summarized in Table 6.1 could still potentially share a common ancestor but it now looks increasingly likely that there are multiple independent origins of these replicators that may have emerged at dramatically different points in time. Indeed, this is likely to be an ongoing process.

[33] Cervera, Urbina, and de la Peña, "Retrozymes Are a Unique Family of Non-Autonomous Retrotransposons with Hammerhead Ribozymes That Propagate in Plants Through Circular RNAs"; Cervera and de la Peña, "Small circRNAs with Self-Cleaving Ribozymes Are Highly Expressed in Diverse Metazoan Transcriptomes."

6.8 Viroid-like RNAs and emergence of protein coding

If emergence of viroid-like replicators is an ongoing process, there seems to be no reason why some of these replicators would not evolve to encode proteins, and indeed, although neither viroids nor retrozymes encompass protein-coding genes, other members of the viroid brotherhood do. The best studied case, obviously, are the ribozyviruses that encode a single protein of unclear provenance, which is employed for a typical virus function, encapsidation of the genome.[34] An unrelated example is a satRNA of rice yellow mottle virus (satRYMV), the smallest known satRNA at 220 nt, that has been shown to encode proteins in overlapping open reading frames, covering most of the genome, and reiteratively translated, in parallel to the replication of satRYMV via reiterative transcription.[35] Translation of satRYMV proceeds through multiple rounds by switching frames and thus yielding proteins of supergenome size. Notably, satRYMV contains ribozymes in both polarities within the protein-coding region. Although too far reaching conclusions should not be drawn from this single case, satRYMV seems to present a remarkable example of *de novo* emergence of proteins. Search for other similar cases might shed light on the evolution of protein coding in new replicators and, possibly, even at processes that occurred in the primordial RNA world.

More generally, it appears likely that the diversity of protein-coding viroid-like agents is substantially underappreciated as suggested in particular by the recent discovery of an expanded range of ribozyviruses;[36] the viroid-like nature of ambiviruses and mitoviruses described in Section 4.3.4 and by Forgia et al.[37]; and

[34] Taylor, "Structure and Replication of Hepatitis Delta Virus RNA"; Sureau and Negro, "The Hepatitis Delta Virus."

[35] AbouHaidar et al., "Novel Coding, Translation, and Gene Expression of a Replicating Covalently Closed Circular RNA of 220 Nt."

[36] Wille et al., "A Divergent Hepatitis D-Like Agent in Birds"; Chang et al., "Novel Hepatitis D-like Agents in Vertebrates and Invertebrates"; Hetzel et al., "Identification of a Novel Deltavirus in Boa Constrictors"; Szirovicza et al., "Snake Deltavirus Utilizes Envelope Proteins of Different Viruses To Generate Infectious Particles"; de la Peña et al., "Hepatitis Delta Virus-Like Circular RNAs from Diverse Metazoans Encode Conserved Hammerhead Ribozymes."

[37] "Virome Characterization of Cryphonectria Parasitica Isolates from Azerbaijan Unveiled a New Mymonavirus and a Putative New RNA Virus Unrelated to Described Viral Sequences."

completely new protein-coding agents.[38] Furthermore, it seems likely that evolution of coding and non-coding viroid-like RNAs is a two-way street, with protein-coding genes both gained and lost on multiple occasions. Comprehensive screening of the rapidly growing metagenomic data for different types of such agents should reveal their actual diversity and evolutionary relationships.

6.9 An expanding host range

The findings of Wei et al.[39], that viroids can replicate in phytopathogenic fungi, have been strengthened by the discovery of ambiviruses, which are viroid-like RNAs that infect fungi (see Section 4.3.4). Although no direct evidence of *viroid* replication was found, it would otherwise be a remarkable coincidence that the same potential hosts of viroids (*i.e.*, *Cryphonectria parasitica*) are also hosts of the viroid-like ambiviruses. The discovery of the viroid-like nature of ambiviruses has certainly extended the host range of viroid-like RNAs to include fungi, lending credence to the idea that *bona fide* viroids are not limited to plants. Indeed, the coinciding host range of two ribozyme-replicating RNAs suggests a potential evolutionary relationship between viroids and ambiviruses. The absence of such a relationship would itself be an exciting discovery, as it would indicate that the ability to replicate via ribozyme-mediated RCR has evolved independently in viroids and ambiviruses.

An equally exciting prospect would be an expanded host range for ribozyviruses. The possibility that ribozyviruses could infect more than just metazoans is tantalizing. Just as with viroids, there does not appear to be a fundamental reason why ribozyviruses could not infect plants or fungi. The findings of distant ribozyviruses in metatranscriptomes largely absent animal material, discussed in Section 4.3.5 combined with the work of Edgar, Taylor, Lin, et al.[40], provides early evidence for

[38] Zheludev et al., "Viroid-Like Colonists of Human Microbiomes."
[39] "Symptomatic Plant Viroid Infections in Phytopathogenic Fungi."
[40] "Petabase-Scale Sequence Alignment Catalyses Viral Discovery."

this hypothesis. The discovery of a ribozyvirus that infects a plant or a fungus would be a major step forward in understanding the evolution of these agents.

It now seems clear that the host range of viroid-like RNAs includes bacteria, as evidenced by the discovery of obelisks in the human microbiome[41] and by means of CRISPR spacer analysis Section 4.3.7. Are archaea also hosts of viroid-like RNAs? Given the trend of expanding host range, it seems likely that viroid-like RNAs will be discovered in archaea. Can unencapsidated viroid-like RNAs infect animals? It certainly seems to be within the realm of possibility. Elucidating the host range of viroid-like RNAs will be critical to understanding their evolution and potential origins.

6.10 Evolution and possible origins of viroid-like agents

Given the small size of most viroid-like cccRNAs and the low level of sequence conservation despite the persistence of the secondary structure, it is difficult to assess the extent of common ancestry among these agents, and their origin on multiple, independent occasions cannot be ruled out.

What has become abundantly clear, however, is the tight evolutionary link between viroid-like agents and riboviruses that most likely involved multiple interconversions between these two types of agents. For some of the viroid-like viruses, such as mitoviruses and putative capsid-encoding satellite viruses, origin via recombination between a viroid and ribovirus appears to be the most likely evolutionary scenario. The same route of evolution appears plausible for ambiviruses although, in this case, the nature of the viral partner is unclear, suggesting that the founding recombination event occurred at an early stage in evolution.

The evolutionary scenarios for ribozyviruses and the Obelisks are less clear given that both encode small proteins without detectable homologs. In these case, *de*

[41] Zheludev et al., “Viroid-Like Colonists of Human Microbiomes.”

novo emergence of protein-coding genes in a viroid-like cccRNAs genomes appear to be a possibility.

The recent reconstruction of the virome of the Last Universal Cellular Ancestor (LUCA) left out ribozyviruses (and implicitly, viroids) on the account of their purported narrow host ranges.[42] The discovery of the broad diversity of viroid-like cccRNAs in numerous metatranscriptomes that are not associated with plants or animals suggest that this conclusion has to be reconsidered. Viroid-like agents including viruses most likely were a presence in the biosphere for billions of years. Thus, the relevance of viroid-like agents for the origin of life has to be reassessed as well. The early ideas of Diener on the descent of viroids from primordial replicators[43] might have been prescient, after all. Typical structures of viroid-like cccRNAs are not conducive to the evolution of complex ribozymes, such as the presumptive ribozyme RdRP.[44] By contrast, it seems plausible that viroid-like agents were among the earliest genetic parasites, fooling the primordial replication apparatus while employing simple ribozymes for some stages of their own replication.

6.11 A new era of viroid research

This work, as well as the work of others, have resulted in a dramatically different view of the subviral world than was known even a few years ago. If anything, there is now a greater need for research into viroids and viroid-like RNAs than ever before. The future is bright for viroid research, and I am excited to see what new discoveries will be made in the coming years. Only time will tell what new secrets the subviral world will reveal.

[42] Mart Krupovic, Valerian V. Dolja, and Eugene V. Koonin, "The LUCA and Its Complex Virome," *Nature Reviews. Microbiology* 18, no. 11 (November 2020): 661–670, doi:10.1038/s41579-020-0408-x.

[43] Diener, "Circular RNAs"; Theodor O Diener, "The Viroid: Biological Oddity or Evolutionary Fossil?" in *Advances in Virus Research*, vol. 57 (Elsevier, 2001), 137–184, doi:10.1016/S0065-3527(01)57003-7; Diener, "Viroids."

[44] David P. Horning and Gerald F. Joyce, "Amplification of RNA by an RNA Polymerase Ribozyme," *Proceedings of the National Academy of Sciences of the United States of America* 113, no. 35 (August 2016): 9786–9791, doi:10.1073/pnas.1610103113; Tjhung et al., "An RNA Polymerase Ribozyme That Synthesizes Its Own Ancestor."

References

Abbas, Zaigham, and Rafia Afzal. "Life Cycle and Pathogenesis of Hepatitis D Virus: A Review." *World Journal of Hepatology* 5, no. 12 (December 2013): 666–675. https://doi.org/10.4254/wjh.v5.i12.666.

Abeywickrama-Samarakoon, Natali, Jean-Claude Cortay, Camille Sureau, Susanne Müller, Dulce Alfaiate, Francesca Guerrieri, Apirat Chaikuad, et al. "Hepatitis Delta Virus Histone Mimicry Drives the Recruitment of Chromatin Remodelers for Viral RNA Replication." *Nature Communications* 11, no. 1 (January 2020): 419. https://doi.org/10.1038/s41467-020-14299-9.

AbouHaidar, Mounir Georges, Srividhya Venkataraman, Ashkan Golshani, Bolin Liu, and Tauqeer Ahmad. "Novel Coding, Translation, and Gene Expression of a Replicating Covalently Closed Circular RNA of 220 Nt." *Proceedings of the National Academy of Sciences* 111, no. 40 (October 2014): 14542–14547. https://doi.org/10.1073/pnas.1402814111.

Adkar-Purushothama, Charith Raj, Teruo Sano, and Jean-Pierre Perreault. "Hop Latent Viroid: A Hidden Threat to the Cannabis Industry." *Viruses* 15, no. 3 (March 2023): 681. https://doi.org/10.3390/v15030681.

Alazem, Mazen, and Na-Sheng Lin. "Large Satellite RNAs." In *Viroids and Satellites* (Elsevier, 2017), 639–648. https://doi.org/10.1016/B978-0-12-801498-1.00059-0.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215, no. 3 (October 1990): 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Asabe, Shinichi, Stefan F. Wieland, Pratip K. Chattopadhyay, Mario Roederer, Ronald E. Engle, Robert H. Purcell, and Francis V. Chisari. "The Size of the

Viral Inoculum Contributes to the Outcome of Hepatitis B Virus Infection." *Journal of Virology* 83, no. 19 (October 2009): 9652–9662. https://doi.org/10.1128/JVI.00867-09.

Ayad, Lorraine A. K., and Solon P. Pissis. "MARS: Improving Multiple Circular Sequence Alignment Using Refined Sequences." *BMC Genomics* 18, no. 1 (December 2017): 86. https://doi.org/10.1186/s12864-016-3477-5.

Badar, Uzma, Srividhya Venkataraman, Mounir AbouHaidar, and Kathleen Hefferon. "Molecular Interactions of Plant Viral Satellites." *Virus Genes* 57, no. 1 (February 2021): 1–22. https://doi.org/10.1007/s11262-020-01806-9.

Barba, Marina, and Ahmed Hadidi. "Application of Next-Generation Sequencing Technologies to Viroids." In *Viroids and Satellites* (Elsevier, 2017), 401–412. https://doi.org/10.1016/B978-0-12-801498-1.00038-3.

Barba, Marina, Munetaka Hosakawa, Qiao-Chun Wang, Anna Taglienti, and Zhibo Zhang. "Viroid Elimination by Thermotherapy, Cold Therapy, Tissue Culture, In Vitro Micrografting, or Cryotherapy." In *Viroids and Satellites* (Elsevier, 2017), 425–435. https://doi.org/10.1016/B978-0-12-801498-1.00040-1.

Barba, Marina, and Delano James. "Quarantine and Certification for Viroids and Viroid Diseases." In *Viroids and Satellites* (Elsevier, 2017), 415–424. https://doi.org/10.1016/B978-0-12-801498-1.00039-5.

Bektaş, A., K. M. Hardwick, K. Waterman, and J. Kristof. "Occurrence of Hop Latent Viroid in Cannabis Sativa with Symptoms of Cannabis Stunting Disease in California." *Plant Disease* 103, no. 10 (October 2019): 2699. https://doi.org/10.1094/PDIS-03-19-0459-PDN.

Bergner, Laura M., Richard J. Orton, Alice Broos, Carlos Tello, Daniel J. Becker, Jorge E. Carrera, Arvind H. Patel, Roman Biek, and Daniel G. Streicker. "Diversification of Mammalian Deltaviruses by Host Shifting." *Proceedings of the National Academy of Sciences* 118, no. 3 (January 2021): e2019907118. https://doi.org/10.1073/pnas.2019907118.

Bester, Rachelle, Sophia S. Malan, and Hans J Maree. "A Plum Marbling Conundrum: Identification of a New Viroid Associated with Marbling and Corky

Flesh in Japanese Plums." *Phytopathology* 110, no. 8 (August 2020): 1476–1482. https://doi.org/10.1094/PHYTO-12-19-0474-R.

Bland, Charles, Teresa L Ramsey, Fareedah Sabree, Micheal Lowe, Kyndall Brown, Nikos C Kyrpides, and Philip Hugenholtz. "CRISPR Recognition Tool (CRT): A Tool for Automatic Detection of Clustered Regularly Interspaced Palindromic Repeats." *BMC Bioinformatics* 8, no. 1 (December 2007): 209. https://doi.org/10.1186/1471-2105-8-209.

Bonfiglioli, Roderick G., Geoffrey I. McFadden, and Robert H. Symons. "In Situ Hybridization Localizes Avocado Sunblotch Viroid on Chloroplast Thylakoid Membranes and Coconut Cadang Cadang Viroid in the Nucleus." *The Plant Journal* 6, no. 1 (July 1994): 99–103. https://doi.org/10.1046/j.1365-313X.1994.6010099.x.

Boyer, Mickaël, Saïd Azza, Lina Barrassi, Thomas Klose, Angélique Campocasso, Isabelle Pagnier, Ghislain Fournous, et al. "Mimivirus Shows Dramatic Genome Reduction After Intraamoebal Culture." *Proceedings of the National Academy of Sciences of the United States of America* 108, no. 25 (June 2011): 10296–10301. https://doi.org/10.1073/pnas.1101118108.

Branch, Andrea D., Bonnie J. Benenfeld, and Hugh D. Robertson. "Evidence for a Single Rolling Circle in the Replication of Potato Spindle Tuber Viroid." *Proceedings of the National Academy of Sciences* 85, no. 23 (December 1988): 9128–9132. https://doi.org/10.1073/pnas.85.23.9128.

Breit, Timo M., Wim C. De Leeuw, Marina Van Olst, Wim A. Ensink, Selina Van Leeuwen, and Rob J. Dekker. "Genome Sequence of a New Carnation Small Viroid-Like RNA, CarSV-1." Edited by Jelle Matthijnssens. *Microbiology Resource Announcements* 12, no. 3 (March 2023): e01219–22. https://doi.org/10.1128/mra.01219-22.

Bushmanova, Elena, Dmitry Antipov, Alla Lapidus, and Andrey D Prjibelski. "rnaSPAdes: A de Novo Transcriptome Assembler and Its Application to RNA-Seq Data." *GigaScience* 8, no. 9 (September 2019): giz100. https://doi.org/10.1093/gigascience/giz100.

Bussière, F., J. Lehoux, D. A. Thompson, L. J. Skrzeczkowski, and J. Perreault. "Subcellular Localization and Rolling Circle Replication of Peach Latent Mosaic Viroid: Hallmarks of Group A Viroids." *Journal of Virology* 73, no. 8 (August 1999): 6353–6360. https://doi.org/10.1128/JVI.73.8.6353-6360.1999.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10, no. 1 (December 2009): 421. https://doi.org/10.1186/1471-2105-10-421.

Cao, Dan, Dirk Haussecker, Yong Huang, and Mark A. Kay. "Combined Proteomic–RNAi Screen for Host Factors Involved in Human Hepatitis Delta Virus Replication." *RNA* 15, no. 11 (November 2009): 1971–1979. https://doi.org/10.1261/rna.1782209.

Caporaso, J Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, et al. "QIIME Allows Analysis of High-Throughput Community Sequencing Data." *Nature Methods* 7, no. 5 (May 2010): 335–336. https://doi.org/10.1038/nmeth.f.303.

Catalán, Pablo, Santiago F. Elena, José A. Cuesta, and Susanna Manrubia. "Parsimonious Scenario for the Emergence of Viroid-Like Replicons De Novo." *Viruses* 11, no. 5 (May 2019): 425. https://doi.org/10.3390/v11050425.

Cervera, Amelia, and Marcos de la Peña. "Small circRNAs with Self-Cleaving Ribozymes Are Highly Expressed in Diverse Metazoan Transcriptomes." *Nucleic Acids Research* 48, no. 9 (May 2020): 5054–5064. https://doi.org/10.1093/nar/gkaa187.

Cervera, Amelia, and Marcos de la Peña. "Cloning and Detection of Genomic Retrozymes and Their circRNA Intermediates." In *Ribozymes: Methods and Protocols*, edited by Robert J. Scarborough and Anne Gatignol Methods in Molecular Biology (New York, NY: Springer US, 2021), 27–44. https://doi.org/10.1007/978-1-0716-0716-9_3.

Cervera, Amelia, Denisse Urbina, and Marcos de la Peña. “Retrozymes Are a Unique Family of Non-Autonomous Retrotransposons with Hammerhead Ribozymes That Propagate in Plants Through Circular RNAs.” *Genome Biology* 17, no. 1 (December 2016): 135. https://doi.org/10.1186/s13059-016-1002-4.

Chadalavada, Durga M., Elizabeth A. Gratton, and Philip C. Bevilacqua. “The Human HDV-like CPEB3 Ribozyme Is Intrinsically Fast-Reacting.” *Biochemistry* 49, no. 25 (June 2010): 5321–5330. https://doi.org/10.1021/bi100434c.

Chang, F L, P J Chen, S J Tu, C J Wang, and D S Chen. “The Large Form of Hepatitis Delta Antigen Is Crucial for Assembly of Hepatitis Delta Virus.” *Proceedings of the National Academy of Sciences of the United States of America* 88, no. 19 (October 1991): 8490–8494. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC52534/.

Chang, Jinhong, Xingcao Nie, Ho Eun Chang, Ziying Han, and John Taylor. “Transcription of Hepatitis Delta Virus RNA by RNA Polymerase II.” *Journal of Virology* 82, no. 3 (February 2008): 1118–1127. https://doi.org/10.1128/JVI.01758-07.

Chang, Jinhong, Xingcao Nie, Severin Gudima, and John Taylor. “Action of Inhibitors on Accumulation of Processed Hepatitis Delta Virus RNAs.” *Journal of Virology* 80, no. 7 (April 2006): 3205–3214. https://doi.org/10.1128/JVI.80.7.3205-3214.2006.

Chang, Jinhong, Patrick Provost, and John M. Taylor. “Resistance of Human Hepatitis Delta Virus RNAs to Dicer Activity.” *Journal of Virology* 77, no. 22 (November 2003): 11910–11917. https://doi.org/10.1128/JVI.77.22.11910-11917.2003.

Chang, Wei-Shan, John H.-O. Pettersson, Callum Le Lay, Mang Shi, Nathan Lo, Michelle Wille, John-Sebastian Eden, and Edward C. Holmes. “Novel Hepatitis D-like Agents in Vertebrates and Invertebrates.” *Virus Evolution* 5, no. 2 (July 2019): vez021. https://doi.org/10.1093/ve/vez021.

Chao, M, S Y Hsieh, and J Taylor. “Role of Two Forms of Hepatitis Delta Virus Antigen: Evidence for a Mechanism of Self-Limiting Genome Replication.”

Journal of Virology 64, no. 10 (October 1990): 5066–5069. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC247998/.

Chen, I-Min A, Ken Chu, Krishnaveni Palaniappan, Anna Ratner, Jinghua Huang, Marcel Huntemann, Patrick Hajek, et al. "The IMG/M Data Management and Analysis System v.6.0: New Tools and Advanced Capabilities." *Nucleic Acids Research* 49, no. D1 (January 2021): D751–D763. https://doi.org/10.1093/nar/gkaa939.

Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. "Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor." *Bioinformatics* 34, no. 17 (September 2018): i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

Chiumenti, Michela, Beatriz Navarro, Thierry Candresse, Ricardo Flores, and Francesco Di Serio. "Reassessing Species Demarcation Criteria in Viroid Taxonomy by Pairwise Identity Matrices." *Virus Evolution* 7, no. 1 (January 2021): veab001. https://doi.org/10.1093/ve/veab001.

Chou, Huei-Chi, Tsai-Yuan Hsieh, Gwo-Tarng Sheu, and Michael M. C. Lai. "Hepatitis Delta Antigen Mediates the Nuclear Import of Hepatitis Delta Virus RNA." *Journal of Virology* 72, no. 5 (May 1998): 3684–3690. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC109589/.

Cohen, Oded, Ozgur Batuman, Yoni Moskowits, Alexander Rozov, Elisha Gootwine, Munir Mawassi, and Moshe Bar-Joseph. "Goat Horns: Platforms for Viroid Transmission to Fruit Trees?" *Phytoparasitica* 33, no. 2 (April 2005): 141–148. https://doi.org/10.1007/BF03029972.

Collet, Yann, and Murray Kucherawy. "Zstandard Compression and the Application/Zstd Media Type." Request for Comments (Internet Engineering Task Force, October 2018). https://doi.org/10.17487/RFC8478.

Copley, Shelley D., Eric Smith, and Harold J. Morowitz. "The Origin of the RNA World: Co-Evolution of Genes and Metabolism." *Bioorganic Chemistry* 35, no. 6 (December 2007): 430–443. https://doi.org/10.1016/j.bioorg.2007.08.001.

Daròs, J. A., J. F. Marcos, C. Hernández, and R. Flores. "Replication of Avocado Sunblotch Viroid: Evidence for a Symmetric Pathway with Two Rolling Circles

and Hammerhead Ribozyme Processing." *Proceedings of the National Academy of Sciences of the United States of America* 91, no. 26 (December 1994): 12813–12817. https://doi.org/10.1073/pnas.91.26.12813.

Daròs, José-Antonio. "Eggplant Latent Viroid." In *Viroids and Satellites* (Elsevier, 2017), 339–344. https://doi.org/10.1016/B978-0-12-801498-1.00032-2.

Daròs, José-Antonio, Santiago F Elena, and Ricardo Flores. "Viroids: An Ariadne's Thread into the RNA Labyrinth." *EMBO Reports* 7, no. 6 (June 2006): 593–598. https://doi.org/10.1038/sj.embor.7400706.

Daròs, José-Antonio, and Ricardo Flores. "Identification of a Retroviroid-Like Element from Plants." *Proceedings of the National Academy of Sciences* 92, no. 15 (July 1995): 6856–6860. https://doi.org/10.1073/pnas.92.15.6856.

Davies, D. L., and M. F. Clark. "A Satellite-Like Nucleic Acid of Arabis Mosaic Virus Associated with Hop Nettlehead Disease." *Annals of Applied Biology* 103, no. 3 (December 1983): 439–448. https://doi.org/10.1111/j.1744-7348.1983.tb02781.x.

Davis, Trent J., Miguel I. Gómez, Scott J. Harper, and Megan Twomey. "The Economic Impact of Hop Stunt Viroid and Certified Clean Planting Materials." *HortScience* 56, no. 12 (December 2021): 1471–1475. https://doi.org/10.21273/HORTSCI15975-21.

Dawkins, Richard. *The Selfish Gene* (New York: Oxford University Press, 1976).

Delan-Forino, C., M.-C. Maurel, and C. Torchet. "Replication of Avocado Sunblotch Viroid in the Yeast Saccharomyces Cerevisiae." *Journal of Virology* 85, no. 7 (April 2011): 3229–3238. https://doi.org/10.1128/JVI.01320-10.

Deutsch, L. Peter. "GZIP File Format Specification Version 4.3." Request for Comments (Internet Engineering Task Force, May 1996). https://doi.org/10.17487/RFC1952.

Di Serio, Francesco, Shi-Fang Li, Vicente Pallás, Robert A. Owens, John W. Randles, Teruo Sano, Jacobus Th.J. Verhoeven, Georgios Vidalakis, and Ricardo Flores. "Viroid Taxonomy." In *Viroids and Satellites* (Elsevier, 2017), 135–146. https://doi.org/10.1016/B978-0-12-801498-1.00013-9.

Di Serio, Francesco, Robert A. Owens, Beatriz Navarro, Pedro Serra, Ángel Emilio Martínez De Alba, Sonia Delgado, Alberto Carbonell, and Selma Gago-Zachert. "Role of RNA Silencing in Plant-Viroid Interactions and in Viroid Pathogenesis." *Virus Research* 323 (January 2023): 198964. https://doi.org/10.1016/j.virusres.2022.198964.

Diener, T. O. "Are Viroids Escaped Introns?" *Proceedings of the National Academy of Sciences* 78, no. 8 (August 1981): 5014–5015. https://doi.org/10.1073/pnas.78.8.5014.

———. "Circular RNAs: Relics of Precellular Evolution?" *Proceedings of the National Academy of Sciences* 86, no. 23 (December 1989): 9370–9374. https://doi.org/10.1073/pnas.86.23.9370.

———. "Potato Spindle Tuber 'Virus'." *Virology* 45, no. 2 (August 1971): 411–428. https://doi.org/10.1016/0042-6822(71)90342-4.

———, ed. *The Viroids* (Boston, MA: Springer US, 1987). https://doi.org/10.1007/978-1-4613-1855-2.

Diener, T. O., M. P. McKinley, and S. B. Prusiner. "Viroids and Prions." *Proceedings of the National Academy of Sciences of the United States of America* 79, no. 17 (September 1982): 5220–5224. https://doi.org/10.1073/pnas.79.17.5220.

Diener, T. O., and W. B. Raymer. "Potato Spindle Tuber Virus: A Plant Virus with Properties of a Free Nucleic Acid." *Science* 158, no. 3799 (October 1967): 378–381. https://doi.org/10.1126/science.158.3799.378.

———. "Potato Spindle Tuber Virus: A Plant Virus with Properties of a Free Nucleic Acid." *Virology* 37, no. 3 (March 1969): 351–366. https://doi.org/10.1016/0042-6822(69)90219-0.

Diener, Theodor O. "The Viroid: Biological Oddity or Evolutionary Fossil?" In *Advances in Virus Research* (Elsevier, 2001), 57:137–184. https://doi.org/10.1016/S0065-3527(01)57003-7.

Diener, Theodor O. "Viroids: 'Living Fossils' of Primordial RNAs?" *Biology Direct* 11, no. 1 (December 2016): 15. https://doi.org/10.1186/s13062-016-0116-7.

Duval, Jean Pierre. “Factorizing Words over an Ordered Alphabet.” *Journal of Algorithms* 4, no. 4 (December 1983): 363–381. https://doi.org/10.1016/0196-6774(83)90017-2.

Eckert, Iris, Richard Friedrich, Christina E Weinberg, and Zasha Weinberg. “Discovery of Natural Non-Circular Permutations in Non-Coding RNAs.” *Nucleic Acids Research* 51, no. 6 (April 2023): 2850–2861. https://doi.org/10.1093/nar/gkad137.

Eddy, Sean R. “Accelerated Profile HMM Searches.” Edited by William R. Pearson. *PLoS Computational Biology* 7, no. 10 (October 2011): e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

Edgar, Robert C., Jeff Taylor, Tomer Altman, Pierre Barbera, Dmitry Meleshko, Victor Lin, Dan Lohr, et al. “Petabase-Scale Sequence Alignment Catalyses Viral Discovery.” Preprint (Bioinformatics, August 2020). https://doi.org/10.1101/2020.08.07.241729.

Edgar, Robert C., Jeff Taylor, Victor Lin, Tomer Altman, Pierre Barbera, Dmitry Meleshko, Dan Lohr, et al. “Petabase-Scale Sequence Alignment Catalyses Viral Discovery.” *Nature* 602, no. 7895 (February 2022): 142–147. https://doi.org/10.1038/s41586-021-04332-2.

Eigen, Manfred. “Selforganization of Matter and the Evolution of Biological Macromolecules.” *Die Naturwissenschaften* 58, no. 10 (October 1971): 465–523. https://doi.org/10.1007/BF00623322.

El-Dougdoug, Khaled A., Kadriye Çağlayan, Amine Elleuch, Hani Z. Al-Tuwariqi, Ebenezer A. Gyamera, and Ahmed Hadidi. “Geographical Distribution of Viroids in Africa and the Middle East.” In *Viroids and Satellites* (Elsevier, 2017), 485–496. https://doi.org/10.1016/B978-0-12-801498-1.00045-0.

Elena, Santiago F., Joaquín Dopazo, Marcos de la Peña, Ricardo Flores, Theodor O. Diener, and Andrés Moya. “Phylogenetic Analysis of Viroid and Viroid-Like Satellite RNAs from Plants: A Reassessment.” *Journal of Molecular Evolution* 53, no. 2 (August 2001): 155–159. https://doi.org/10.1007/s002390010203.

Faggioli, Francesco, Nuria Duran-Vila, Mina Tsagris, and Vicente Pallás. "Geographical Distribution of Viroids in Europe." In *Viroids and Satellites* (Elsevier, 2017), 473–484. https://doi.org/10.1016/B978-0-12-801498-1.00044-9.

Farci, Patrizia, and Grazia Niro. "Clinical Features of Hepatitis D." *Seminars in Liver Disease* 32, no. 3 (August 2012): 228–236. https://doi.org/10.1055/s-0032-1323628.

Farci, Patrizia, Grazia Anna Niro, Fausto Zamboni, and Giacomo Diaz. "Hepatitis D Virus and Hepatocellular Carcinoma." *Viruses* 13, no. 5, 5 (May 2021): 830. https://doi.org/10.3390/v13050830.

Fedor, M. J. "Tertiary Structure Stabilization Promotes Hairpin Ribozyme Ligation." *Biochemistry* 38, no. 34 (August 1999): 11040–11050. https://doi.org/10.1021/bi991069q.

Fedor, Martha J. "Structure and Function of the Hairpin Ribozyme." *Journal of Molecular Biology* 297, no. 2 (March 2000): 269–291. https://doi.org/10.1006/jmbi.2000.3560.

Fernandes, Francisco, Luísa Pereira, and Ana T Freitas. "CSA: An Efficient Algorithm to Improve Circular DNA Multiple Alignment." *BMC Bioinformatics* 10, no. 1 (2009): 230. https://doi.org/10.1186/1471-2105-10-230.

Ferré-D'Amaré, Adrian R., and William G. Scott. "Small Self-Cleaving Ribozymes." *Cold Spring Harbor Perspectives in Biology* 2, no. 10 (October 2010): a003574. https://doi.org/10.1101/cshperspect.a003574.

Ferré-D'Amaré, Adrian R., Kaihong Zhou, and Jennifer A. Doudna. "Crystal Structure of a Hepatitis Delta Virus Ribozyme." *Nature* 395, no. 6702 (October 1998): 567–574. https://doi.org/10.1038/26912.

Flores, Ricardo, Francesco Di Serio, Beatriz Navarro, and Robert A. Owens. "Viroid Pathogenesis." In *Viroids and Satellites* (Elsevier, 2017), 93–103. https://doi.org/10.1016/B978-0-12-801498-1.00009-7.

Flores, Ricardo, Selma Gago-Zachert, Pedro Serra, Marcos De la Peña, and Beatriz Navarro. "Chrysanthemum Chlorotic Mottle Viroid." In *Viroids and Satellites*

(Elsevier, 2017), 331–338. https://doi.org/10.1016/B978-0-12-801498-1.00031-0.

Flores, Ricardo, María-Eugenia Gas, Diego Molina-Serrano, María-Ángeles Nohales, Alberto Carbonell, Selma Gago, Marcos De la Peña, and José-Antonio Daròs. "Viroid Replication: Rolling-Circles, Enzymes and Ribozymes." *Viruses* 1, no. 2 (September 2009): 317–334. https://doi.org/10.3390/v1020317.

Flores, Ricardo, Douglas Grubb, Amine Elleuch, María-Ángeles Nohales, Sonia Delgado, and Selma Gago. "Rolling-Circle Replication of Viroids, Viroid-Like Satellite RNAs and Hepatitis Delta Virus: Variations on a Theme." *RNA Biology* 8, no. 2 (March 2011): 200–206. https://doi.org/10.4161/rna.8.2.14238.

Flores, Ricardo, Sofia Minoia, Amparo López-Carrasco, Sonia Delgado, Ángel-Emilio Martínez de Alba, and Kriton Kalantidis. "Viroid Replication." In *Viroids and Satellites* (Elsevier, 2017), 71–81. https://doi.org/10.1016/B978-0-12-801498-1.00007-3.

Flores, Ricardo, Beatriz Navarro, Sonia Delgado, Carmen Hernández, Wen-Xing Xu, Marina Barba, Ahmed Hadidi, and Francesco Di Serio. "Peach Latent Mosaic Viroid in Infected Peach." In *Viroids and Satellites* (Elsevier, 2017), 307–316. https://doi.org/10.1016/B978-0-12-801498-1.00029-2.

Flores, Ricardo, Beatriz Navarro, Pedro Serra, and Francesco Di Serio. "A Scenario for the Emergence of Protoviroids in the RNA World and for Their Further Evolution into Viroids and Viroid-Like RNAs by Modular Recombinations and Mutations." *Virus Evolution* 8, no. 1 (January 2022): veab107. https://doi.org/10.1093/ve/veab107.

Flores, R., S. Minoia, A. Carbonell, A. Gisel, S. Delgado, A. López-Carrasco, B. Navarro, and F. Di Serio. "Viroids, the Simplest RNA Replicons: How They Manipulate Their Hosts for Being Propagated and How Their Hosts React for Containing the Infection." *Virus Research* 209 (November 2015): 136–145. https://doi.org/10.1016/j.virusres.2015.02.027.

Forgia, Marco, Beatriz Navarro, Stefania Daghino, Amelia Cervera, Andreas Gisel, Silvia Perotto, Dilzara N. Aghayeva, et al. "Hybrids of RNA Viruses and Viroid-Like Elements Replicate in Fungi." *Nature Communications* 14, no. 1 (May 2023): 2591. https://doi.org/10.1038/s41467-023-38301-2.

Forgia, M., E. Isgandarli, D. N. Aghayeva, I. Huseynova, and M. Turina. "Virome Characterization of Cryphonectria Parasitica Isolates from Azerbaijan Unveiled a New Mymonavirus and a Putative New RNA Virus Unrelated to Described Viral Sequences." *Virology* 553 (January 2021): 51–61. https://doi.org/10.1016/j.virol.2020.10.008.

Frickey, Tancred, and Andrei Lupas. "CLANS: A Java Application for Visualizing Protein Families Based on Pairwise Similarity." *Bioinformatics* 20, no. 18 (December 2004): 3702–3704. https://doi.org/10.1093/bioinformatics/bth444.

Fu, T B, and J Taylor. "The RNAs of Hepatitis Delta Virus Are Copied by RNA Polymerase II in Nuclear Homogenates." *Journal of Virology* 67, no. 12 (December 1993): 6965–6972. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC238155/.

Gábor, Csárdi, and Tamás Nepusz. "The Igraph Software Package for Complex Network Research." *InterJournal Complex Systems*, November 2005, 1695.

Gago, Selma, Santiago F. Elena, Ricardo Flores, and Rafael Sanjuán. "Extremely High Mutation Rate of a Hammerhead Viroid." *Science* 323, no. 5919 (March 2009): 1308–1308. https://doi.org/10.1126/science.1169202.

Gao, Dongying, Jinfeng Chen, Mingsheng Chen, Blake C. Meyers, and Scott Jackson. "A Highly Conserved, Small LTR Retrotransposon That Preferentially Targets Genes in Grass Genomes." Edited by Mark A. Batzer. *PLoS ONE* 7, no. 2 (February 2012): e32010. https://doi.org/10.1371/journal.pone.0032010.

Gao, Yuan, Jinfeng Wang, and Fangqing Zhao. "CIRI: An Efficient and Unbiased Algorithm for de Novo Circular RNA Identification." *Genome Biology* 16, no. 1 (December 2015): 4. https://doi.org/10.1186/s13059-014-0571-3.

Geering, Andrew D. W. “Geographical Distribution of Viroids in Oceania.” In *Viroids and Satellites* (Elsevier, 2017), 497–506. https://doi.org/10.1016/B978-0-12-801498-1.00046-2.

Giguère, Tamara, Charith Raj Adkar-Purushothama, François Bolduc, and Jean-Pierre Perreault. “Elucidation of the Structures of All Members of the Avsunviroidae Family.” *Molecular Plant Pathology* 15, no. 8 (October 2014): 767–779. https://doi.org/10.1111/mpp.12130.

Giguère, Tamara, Charith Raj Adkar-Purushothama, and Jean-Pierre Perreault. “Comprehensive Secondary Structure Elucidation of Four Genera of the Family Pospiviroidae.” *PLOS One* 9, no. 6 (2014): e98655. https://doi.org/10.1371/journal.pone.0098655.

Glouzon, Jean-Pierre Sehi, François Bolduc, Shengrui Wang, Rafael J. Najmanovich, and Jean-Pierre Perreault. “Deep-Sequencing of the Peach Latent Mosaic Viroid Reveals New Aspects of Population Heterogeneity.” Edited by Yury E. Khudyakov. *PLoS ONE* 9, no. 1 (January 2014): e87297. https://doi.org/10.1371/journal.pone.0087297.

Gnanasekaran, Prabu, and Supriya Chakraborty. “Biology of Viral Satellites and Their Role in Pathogenesis.” *Current Opinion in Virology* 33 (December 2018): 96–105. https://doi.org/10.1016/j.coviro.2018.08.002.

Gozmanova, M. “Characterization of the RNA Motif Responsible for the Specific Interaction of Potato Spindle Tuber Viroid RNA (PSTVd) and the Tomato Protein Virp1.” *Nucleic Acids Research* 31, no. 19 (October 2003): 5534–5543. https://doi.org/10.1093/nar/gkg777.

Grabowski, Paula J., Arthur J. Zaug, and Thomas R. Cech. “The Intervening Sequence of the Ribosomal RNA Precursor Is Converted to a Circular RNA in Isolated Nuclei of Tetrahymena.” *Cell* 23, no. 2 (February 1981): 467–476. https://doi.org/10.1016/0092-8674(81)90142-2.

Greco-Stewart, Valerie, and Martin Pelchat. “Interaction of Host Cellular Proteins with Components of the Hepatitis Delta Virus.” *Viruses* 2, no. 1 (January 2010): 189–212. https://doi.org/10.3390/v2010189.

Griffin, Brittany L., Sergey Chasovskikh, Anatoly Dritschilo, and John L. Casey. "Hepatitis Delta Antigen Requires a Flexible Quasi-Double-Stranded RNA Structure To Bind and Condense Hepatitis Delta Virus RNA in a Ribonucleoprotein Complex." *Journal of Virology* 88, no. 13 (July 2014): 7402–7411. https://doi.org/10.1128/JVI.00443-14.

Gu, Zuguang, Lei Gu, Roland Eils, Matthias Schlesner, and Benedikt Brors. "Circlize Implements and Enhances Circular Visualization in R." *Bioinformatics* 30, no. 19 (October 2014): 2811–2812. https://doi.org/10.1093/bioinformatics/btu393.

Gudima, Severin, Kate Dingle, Ting-Ting Wu, Gloria Moraleda, and John Taylor. "Characterization of the 5′ Ends for Polyadenylated RNAs Synthesized During the Replication of Hepatitis Delta Virus." *Journal of Virology* 73, no. 8 (August 1999): 6533. https://doi.org/10.1128/jvi.73.8.6533-6539.1999.

Hadidi, Ahmed, Liying Sun, and John W. Randles. "Modes of Viroid Transmission." *Cells* 11, no. 4 (February 2022): 719. https://doi.org/10.3390/cells11040719.

Hadidi, Ahmed, Georgios Vidalakis, and Teruo Sano. "Economic Significance of Fruit Tree and Grapevine Viroids." In *Viroids and Satellites* (Elsevier, 2017), 15–25. https://doi.org/10.1016/B978-0-12-801498-1.00002-4.

Hajimorad, M. R., S. A. Ghabrial, and M. J. Roossinck. "De Novo Emergence of a Novel Satellite RNA of Cucumber Mosaic Virus Following Serial Passages of the Virus Derived from RNA Transcripts." *Archives of Virology* 154, no. 1 (January 2009): 137–140. https://doi.org/10.1007/s00705-008-0280-x.

Hammond, Rosemarie W. "Economic Significance of Viroids in Vegetable and Field Crops." In *Viroids and Satellites* (Elsevier, 2017), 5–13. https://doi.org/10.1016/B978-0-12-801498-1.00001-2.

———. "Seed, Pollen, and Insect Transmission of Viroids." In *Viroids and Satellites* (Elsevier, 2017), 521–530. https://doi.org/10.1016/B978-0-12-801498-1.00048-6.

Hammond, Rosemarie W., and Natalia Kovalskaya. "Strategies to Introduce Resistance to Viroids." In *Viroids and Satellites* (Elsevier, 2017), 447–455. https://doi.org/10.1016/B978-0-12-801498-1.00042-5.

Han, Ziying, Carolina Alves, Severin Gudima, and John Taylor. "Intracellular Localization of Hepatitis Delta Virus Proteins in the Presence and Absence of Viral RNA Accumulation." *Journal of Virology* 83, no. 13 (July 2009): 6457–6463. https://doi.org/10.1128/JVI.00008-09.

Haruna, I., and S. Spiegelman. "Autocatalytic Synthesis of a Viral RNA in Vitro." *Science* 150, no. 3698 (November 1965): 884–886. https://doi.org/10.1126/science.150.3698.884.

Hataya, Tatsuji, Taro Tsushima, and Teruo Sano. "Hop Stunt Viroid." In *Viroids and Satellites* (Elsevier, 2017), 199–210. https://doi.org/10.1016/B978-0-12-801498-1.00019-X.

Hatcher, Eneida L., Sergey A. Zhdanov, Yiming Bao, Olga Blinkova, Eric P. Nawrocki, Yuri Ostapchuck, Alejandro A. Schäffer, and J. Rodney Brister. "Virus Variation Resource – Improved Response to Emergent Viral Outbreaks." *Nucleic Acids Research* 45, no. D1 (January 2017): D482–D490. https://doi.org/10.1093/nar/gkw1065.

Hedtke, Boris, Thomas Börner, and Andreas Weihe. "Mitochondrial and Chloroplast Phage-Type RNA Polymerases in Arabidopsis." *Science* 277, no. 5327 (August 1997): 809–811. https://doi.org/10.1126/science.277.5327.809.

Hegedus, Krisztina, Géza Dallmann, and Ervin Balázs. "The DNA Form of a Retroviroid-Like Element Is Involved in Recombination Events with Itself and with the Plant Genome." *Virology* 325, no. 2 (August 2004): 277–286. https://doi.org/10.1016/j.virol.2004.04.035.

Hepojoki, Jussi, Udo Hetzel, Sofia Paraskevopoulou, Christian Drosten, Balazs Harrach, F. Murilo Zerbini, Eugene V. Koonin, Mart Krupovic, Valerian V. Dolja, and Jens H. Kuhn. "Create One New Realm (*Ribozyviria*) Including One New Family (*Kolmioviridae*) Including Genus *Deltavirus* and Seven New Genera

for a Total of 15 Species." Proposal (International Committee on Taxonomy of Viruses, March 2020). http://doi.org/10.13140/RG.2.2.31235.43041.

Hetzel, Udo, Leonóra Szirovicza, Teemu Smura, Barbara Prähauser, Olli Vapalahti, Anja Kipar, and Jussi Hepojoki. "Identification of a Novel Deltavirus in Boa Constrictors." Edited by Anne Moscona. *mBio* 10, no. 2 (April 2019): e00014–19. https://doi.org/10.1128/mBio.00014-19.

Hodgson, R. A. J., G. C. Wall, and J. W. Randles. "Specific Identification of Coconut Tinangaja Viroid for Differential Field Diagnosis of Viroids in Coconut Palm." *Phytopathology®* 88, no. 8 (August 1998): 774–781. https://doi.org/10.1094/PHYTO.1998.88.8.774.

Hong, Shiao-Ya, and Pei-Jer Chen. "Phosphorylation of Serine 177 of the Small Hepatitis Delta Antigen Regulates Viral Antigenomic RNA Replication by Interacting with the Processive RNA Polymerase II." *Journal of Virology* 84, no. 3 (February 2010): 1430–1438. https://doi.org/10.1128/JVI.02083-09.

Horning, David P., and Gerald F. Joyce. "Amplification of RNA by an RNA Polymerase Ribozyme." *Proceedings of the National Academy of Sciences of the United States of America* 113, no. 35 (August 2016): 9786–9791. https://doi.org/10.1073/pnas.1610103113.

Hou, Wan-Ying, Shi-Fang Li, Zu-Jian Wu, Dong-Mei Jiang, and Teruo Sano. "Coleus Blumei Viroid 6: A New Tentative Member of the Genus Coleviroid Derived from Natural Genome Shuffling." *Archives of Virology* 154, no. 6 (June 2009): 993–997. https://doi.org/10.1007/s00705-009-0388-7.

Hsieh, S. Y., M. Chao, L. Coates, and J. Taylor. "Hepatitis Delta Virus Genome Replication: A Polyadenylated mRNA for Delta Antigen." *Journal of Virology* 64, no. 7 (July 1990): 3192. https://doi.org/10.1128/jvi.64.7.3192-3198.1990.

Hu, Chung-Chi, Yau-Heiu Hsu, and Na-Sheng Lin. "Satellite RNAs and Satellite Viruses of Plants." *Viruses* 1, no. 3 (December 2009): 1325–1350. https://doi.org/10.3390/v1031325.

Hu, Jianming. "Hepatitis B Virus Virology and Replication." In *Hepatitis B Virus in Human Diseases*, edited by Yun-Fan Liaw and Fabien Zoulim (Cham:

Springer International Publishing, 2016), 1–34. https://doi.org/10.1007/978-3-319-22330-8_1.

Hu, Y., P. A. Feldstein, P. J. Bottino, and R. A. Owens. "Role of the Variable Domain in Modulating Potato Spindle Tuber Viroid Replication." *Virology* 219, no. 1 (May 1996): 45–56. https://doi.org/10.1006/viro.1996.0221.

Huang, W. H., C. W. Chen, H. L. Wu, and P. J. Chen. "Post-Translational Modification of Delta Antigen of Hepatitis D Virus." *Current Topics in Microbiology and Immunology* 307 (2006): 91–112. https://doi.org/10.1007/3-540-29802-9_5.

Huang, Weichun, Leping Li, Jason R. Myers, and Gabor T. Marth. "ART: A Next-Generation Sequencing Read Simulator." *Bioinformatics* 28, no. 4 (February 2012): 593–594. https://doi.org/10.1093/bioinformatics/btr708.

Huang, Wen-Hung, Yen-Shun Chen, and Pei-Jer Chen. "Nucleolar Targeting of Hepatitis Delta Antigen Abolishes Its Ability to Initiate Viral Antigenomic RNA Replication." *Journal of Virology* 82, no. 2 (January 2008): 692–699. https://doi.org/10.1128/JVI.01155-07.

Huang, Ying-Wen, Chung-Chi Hu, Yau-Heiu Hsu, and Na-Sheng Lin. "Replication of Satellites." In *Viroids and Satellites* (Elsevier, 2017), 577–586. https://doi.org/10.1016/B978-0-12-801498-1.00053-X.

Jain, Nimit, Lucas R. Blauch, Michal R. Szymanski, Rhiju Das, Sindy K. Y. Tang, Y. Whitney Yin, and Andrew Z. Fire. "Transcription Polymerase–Catalyzed Emergence of Novel RNA Replicons." *Science* 368, no. 6487 (April 2020): eaay0688. https://doi.org/10.1126/science.aay0688.

Jalasvuori, Matti, and Eugene V. Koonin. "Classification of Prokaryotic Genetic Replicators: Between Selfishness and Altruism." *Annals of the New York Academy of Sciences* 1341 (April 2015): 96–105. https://doi.org/10.1111/nyas.12696.

Jandura, Allison, and Henry M. Krause. "The New RNA World: Growing Evidence for Long Noncoding RNA Functionality." *Trends in Genetics* 33, no. 10 (October 2017): 665–676. https://doi.org/10.1016/j.tig.2017.08.002.

Jayasinghe, Wikum H., Hangil Kim, Yusuke Nakada, and Chikara Masuta. "A Plant Virus Satellite RNA Directly Accelerates Wing Formation in Its Insect Vector for Spread." *Nature Communications* 12, no. 1 (December 2021): 7087. https://doi.org/10.1038/s41467-021-27330-4.

Jeancolas, Cyrille, Yoshiya J. Matsubara, Mykhailo Vybornyi, Camille N. Lambert, Alex Blokhuis, Thomas Alline, Andrew D. Griffiths, Sandeep Ameta, Sandeep Krishna, and Philippe Nghe. "RNA Diversification by a Self-Reproducing Ribozyme Revealed by Deep Sequencing and Kinetic Modelling." *Chemical Communications* 57, no. 61 (2021): 7517–7520. https://doi.org/10.1039/D1CC02290C.

Jimenez, Randi M., Eric Delwart, and Andrej Lupták. "Structure-Based Search Reveals Hammerhead Ribozymes in the Human Microbiome." *Journal of Biological Chemistry* 286, no. 10 (March 2011): 7737–7743. https://doi.org/10.1074/jbc.C110.209288.

Johnson, Marc T. J., Eric J. Carpenter, Zhijian Tian, Richard Bruskiewich, Jason N. Burris, Charlotte T. Carrigan, Mark W. Chase, et al. "Evaluating Methods for Isolating Total RNA and Predicting the Success of Sequencing Phylogenetically Diverse Plant Transcriptomes." Edited by Christopher Quince. *PLoS ONE* 7, no. 11 (November 2012): e50226. https://doi.org/10.1371/journal.pone.0050226.

Kalvari, Ioanna, Eric P Nawrocki, Nancy Ontiveros-Palacios, Joanna Argasinska, Kevin Lamkiewicz, Manja Marz, Sam Griffiths-Jones, et al. "Rfam 14: Expanded Coverage of Metagenomic, Viral and microRNA Families." *Nucleic Acids Research* 49, no. D1 (January 2021): D192–D200. https://doi.org/10.1093/nar/gkaa1047.

Kappagantu, Madhu, Mark E. Nelson, Jeff M. Bullock, Stephen T. Kenny, and Kenneth C. Eastwell. "Hop Stunt Viroid: Effects on Vegetative Growth and Yield of Hop Cultivars, and Its Distribution in Central Washington State." *Plant Disease* 101, no. 4 (April 2017): 607–612. https://doi.org/10.1094/PDIS-06-16-0884-RE.

Katoh, K., and D. M. Standley. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30, no. 4 (April 2013): 772–780. https://doi.org/10.1093/molbev/mst010.

Kazakov, S. A. "Ligation of the Hairpin Ribozyme in Cis Induced by Freezing and Dehydration." *RNA* 12, no. 3 (March 2006): 446–456. https://doi.org/10.1261/rna.2123506.

Keese, Paul, George Bruening, and Robert H. Symons. "Comparative Sequence and Structure of Circular RNAs from Two Isolates of Lucerne Transient Streak Virus." *FEBS Letters* 159, no. 1–2 (August 1983): 185–190. https://doi.org/10.1016/0014-5793(83)80443-8.

Keese, P, and R H Symons. "Domains in Viroids: Evidence of Intermolecular RNA Rearrangements and Their Contribution to Viroid Evolution." *Proceedings of the National Academy of Sciences of the United States of America* 82, no. 14 (July 1985): 4582–4586. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC390429/.

Kerpedjiev, Peter, Stefan Hammer, and Ivo L. Hofacker. "Forna (Force-Directed RNA): Simple and Effective Online RNA Secondary Structure Diagrams." *Bioinformatics* 31, no. 20 (October 2015): 3377–3379. https://doi.org/10.1093/bioinformatics/btv372.

Kim, Mikyeong, and Marilyn J. Roossinck. "Small Linear Satellite RNAs." In *Viroids and Satellites* (Elsevier, 2017), 649–658. https://doi.org/10.1016/B978-0-12-801498-1.00060-7.

Knight, Rob, Peter Maxwell, Amanda Birmingham, Jason Carnes, J Gregory Caporaso, Brett C Easton, Michael Eaton, et al. "PyCogent: A Toolkit for Making Sense from Sequence." *Genome Biology* 8, no. 8 (2007): R171. https://doi.org/10.1186/gb-2007-8-8-r171.

Koonin, Eugene V. *The Logic of Chance: The Nature and Origin of Biological Evolution* (Upper Saddle River, N.J: Pearson Education, 2012).

Koonin, Eugene V., and Valerian V. Dolja. "Virus World as an Evolutionary Network of Viruses and Capsidless Selfish Elements." *Microbiology and Molecular Biology Reviews* 78, no. 2 (June 2014): 278–303. https://doi.org/10.1128/MMBR.00049-13.

Koonin, Eugene V., Valerian V. Dolja, Mart Krupovic, Arvind Varsani, Yuri I. Wolf, Natalya Yutin, F. Murilo Zerbini, and Jens H. Kuhn. "Global Organization and Proposed Megataxonomy of the Virus World." *Microbiology and Molecular Biology Reviews* 84, no. 2 (May 2020): e00061–19. https://doi.org/10.1128/MMBR.00061-19.

Koonin, Eugene V., and Petro Starokadomskyy. "Are Viruses Alive? The Replicator Paradigm Sheds Decisive Light on an Old but Misguided Question." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 59 (October 2016): 125–134. https://doi.org/10.1016/j.shpsc.2016.02.016.

Koonin, Eugene V., and Yuri I. Wolf. "Genomics of Bacteria and Archaea: The Emerging Dynamic View of the Prokaryotic World." *Nucleic Acids Research* 36, no. 21 (December 2008): 6688–6719. https://doi.org/10.1093/nar/gkn668.

Koonin, Eugene V., Yuri I. Wolf, and Mikhail I. Katsnelson. "Inevitability of the Emergence and Persistence of Genetic Parasites Caused by Evolutionary Instability of Parasite-Free States." *Biology Direct* 12, no. 1 (December 2017): 31. https://doi.org/10.1186/s13062-017-0202-5.

Kos, A., R. Dijkema, A. C. Arnberg, P. H. van der Meide, and H. Schellekens. "The Hepatitis Delta (δ) Virus Possesses a Circular RNA." *Nature* 323, no. 6088 (October 1986): 558–560. https://doi.org/10.1038/323558a0.

Koster, J., and S. Rahmann. "Snakemake—a Scalable Bioinformatics Workflow Engine." *Bioinformatics* 28, no. 19 (October 2012): 2520–2522. https://doi.org/10.1093/bioinformatics/bts480.

Köster, Johannes. "Rust-Bio: A Fast and Safe Bioinformatics Library." *Bioinformatics* 32, no. 3 (February 2016): 444–446. https://doi.org/10.1093/bioinformatics/btv573.

Kruger, Kelly, Paula J. Grabowski, Arthur J. Zaug, Julie Sands, Daniel E. Gottschling, and Thomas R. Cech. "Self-Splicing RNA: Autoexcision and Autocyclization of the Ribosomal RNA Intervening Sequence of Tetrahymena." *Cell* 31, no. 1 (November 1982): 147–157. https://doi.org/10.1016/0092-8674(82)90414-7.

Krupovic, Mart, Valerian V. Dolja, and Eugene V. Koonin. "Origin of Viruses: Primordial Replicators Recruiting Capsids from Hosts." *Nature Reviews Microbiology* 17, no. 7 (July 2019): 449–458. https://doi.org/10.1038/s41579-019-0205-6.

———. "The LUCA and Its Complex Virome." *Nature Reviews. Microbiology* 18, no. 11 (November 2020): 661–670. https://doi.org/10.1038/s41579-020-0408-x.

Krupovic, Mart, Jens H. Kuhn, and Matthias G. Fischer. "A Classification System for Virophages and Satellite Viruses." *Archives of Virology* 161, no. 1 (January 2016): 233–247. https://doi.org/10.1007/s00705-015-2622-9.

Krupovic, Mart, Arvind Varsani, Darius Kazlauskas, Mya Breitbart, Eric Delwart, Karyna Rosario, Natalya Yutin, et al. "Cressdnaviricota: A Virus Phylum Unifying Seven Families of Rep-Encoding Viruses with Single-Stranded, Circular DNA Genomes." *Journal of Virology* 94, no. 12 (June 2020): e00582–20. https://doi.org/10.1128/JVI.00582-20.

Kuhn, David N., Andrew D. W. Geering, and Jonathan Dixon. "Avocado Sunblotch Viroid." In *Viroids and Satellites* (Elsevier, 2017), 297–305. https://doi.org/10.1016/B978-0-12-801498-1.00028-0.

Lafontaine, D. A., S. Mercure, V. Poisson, and J. P. Perreault. "The Viroid and Viroid-Like RNA Database." *Nucleic Acids Research* 26, no. 1 (January 1998): 190–191. https://doi.org/10.1093/nar/26.1.190.

Langmead, Ben, and Steven L Salzberg. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9, no. 4 (April 2012): 357–359. https://doi.org/10.1038/nmeth.1923.

Lee, Benjamin D., and Eugene V. Koonin. "Viroids and Viroid-like Circular RNAs: Do They Descend from Primordial Replicators?" *Life* 12, no. 1 (January 2022): 103. https://doi.org/10.3390/life12010103.

Lee, Benjamin D., Uri Neri, Caleb J. Oh, Peter Simmonds, and Eugene V. Koonin. "ViroidDB: A Database of Viroids and Viroid-Like Circular RNAs." *Nucleic Acids Research*, November 2021, gkab974. https://doi.org/10.1093/nar/gkab974.

Lee, Benjamin D., Uri Neri, Simon Roux, Yuri I. Wolf, Antonio Pedro Camargo, Mart Krupovic, Peter Simmonds, et al. "Mining Metatranscriptomes Reveals a Vast World of Viroid-Like Circular RNAs." *Cell* 186, no. 3 (February 2023): 646–661.e4. https://doi.org/10.1016/j.cell.2022.12.039.

Leichtfried, Thomas, Helga Reisenzein, Siegrid Steinkellner, and Richard A. Gottsberger. "Transmission Studies of the Newly Described Apple Chlorotic Fruit Spot Viroid Using a Combined RT-qPCR and Droplet Digital PCR Approach." *Archives of Virology* 165, no. 11 (2020): 2665–2671. https://doi.org/10.1007/s00705-020-04704-5.

Letunic, Ivica, and Peer Bork. "Interactive Tree Of Life (iTOL) V5: An Online Tool for Phylogenetic Tree Display and Annotation." *Nucleic Acids Research* 49, no. W1 (July 2021): W293–W296. https://doi.org/10.1093/nar/gkab301.

Li, Weizhong, and Adam Godzik. "Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences." *Bioinformatics* 22, no. 13 (July 2006): 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

Li, Wenjun, Kathleen R O'Neill, Daniel H Haft, Michael DiCuccio, Vyacheslav Chetvernin, Azat Badretdin, George Coulouris, et al. "RefSeq: Expanding the Prokaryotic Genome Annotation Pipeline Reach with Protein Family Model Curation." *Nucleic Acids Research* 49, no. D1 (January 2021): D1020–D1028. https://doi.org/10.1093/nar/gkaa1105.

Li, Yi-Jia, Thomas Macnaughton, Lu Gao, and Michael M. C. Lai. "RNA-Templated Replication of Hepatitis Delta Virus: Genomic and Antigenomic

RNAs Associate with Different Nuclear Bodies." *Journal of Virology* 80, no. 13 (July 2006): 6478–6486. https://doi.org/10.1128/JVI.02650-05.

Lima, M. I., M. E. N. Fonseca, R. Flores, and E. W. Kitajima. "Detection of Avocado Sunblotch Viroid in Chloroplasts of Avocado Leaves by in Situ Hybridization." *Archives of Virology* 138, no. 3–4 (September 1994): 385–390. https://doi.org/10.1007/BF01379142.

Ling, Kai-Shu. "Decontamination Measures to Prevent Mechanical Transmission of Viroids." In *Viroids and Satellites* (Elsevier, 2017), 437–445. https://doi.org/10.1016/B978-0-12-801498-1.00041-3.

Ling, K.-S., R. Li, D. R. Panthee, and R. G. Gardner. "First Report of Potato Spindle Tuber Viroid Naturally Infecting Greenhouse Tomatoes in North Carolina." *Plant Disease* 97, no. 1 (January 2013): 148. https://doi.org/10.1094/PDIS-07-12-0679-PDN.

Ling, K.-S., and D. Sfetcu. "First Report of Natural Infection of Greenhouse Tomatoes by Potato Spindle Tuber Viroid in the United States." *Plant Disease* 94, no. 11 (November 2010): 1376. https://doi.org/10.1094/PDIS-07-10-0516.

Linnakoski, Riikka, Suvi Sutela, Martin P. A. Coetzee, Tuan A. Duong, Igor N. Pavlov, Yulia A. Litovka, Jarkko Hantula, Brenda D. Wingfield, and Eeva J. Vainio. "Armillaria Root Rot Fungi Host Single-Stranded RNA Viruses." *Scientific Reports* 11, no. 1 (December 2021): 7336. https://doi.org/10.1038/s41598-021-86343-7.

Liu, Yijin, Timothy J. Wilson, and David M. J. Lilley. "The Structure of a Nucleolytic Ribozyme That Employs a Catalytic Metal Ion." *Nature Chemical Biology* 13, no. 5 (May 2017): 508–513. https://doi.org/10.1038/nchembio.2333.

López-Carrasco, Amparo, Cristina Ballesteros, Vicente Sentandreu, Sonia Delgado, Selma Gago-Zachert, Ricardo Flores, and Rafael Sanjuán. "Different Rates of Spontaneous Mutation of Chloroplastic and Nuclear Viroids as Determined by High-Fidelity Ultra-Deep Sequencing." Edited by Edward C. Holmes. *PLOS Pathogens* 13, no. 9 (September 2017): e1006547. https://doi.org/10.1371/journal.ppat.1006547.

Lorenz, Ronny, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. "ViennaRNA Package 2.0." *Algorithms for Molecular Biology* 6, no. 1 (December 2011): 26. https://doi.org/10.1186/1748-7188-6-26.

Macke, T. J. "RNAMotif, an RNA Secondary Structure Definition and Search Algorithm." *Nucleic Acids Research* 29, no. 22 (November 2001): 4724–4735. https://doi.org/10.1093/nar/29.22.4724.

Mackie, A. E., B. A. Coutts, M. J. Barbetti, B. C. Rodoni, S. J. McKirdy, and R. A. C. Jones. "Potato Spindle Tuber Viroid: Stability on Common Surfaces and Inactivation With Disinfectants." *Plant Disease* 99, no. 6 (June 2015): 770–775. https://doi.org/10.1094/PDIS-09-14-0929-RE.

MacNaughton, T. B., E. J. Gowans, S. P. McNamara, and C. J. Burrell. "Hepatitis Delta Antigen Is Necessary for Access of Hepatitis Delta Virus RNA to the Cell Transcriptional Machinery but Is Not Part of the Transcriptional Complex." *Virology* 184, no. 1 (September 1991): 387–390. https://doi.org/10.1016/0042-6822(91)90855-6.

Macnaughton, T. B., and M. M. C. Lai. "HDV RNA Replication: Ancient Relic or Primer?" *Current Topics in Microbiology and Immunology* 307 (2006): 25–45. https://doi.org/10.1007/3-540-29802-9_2.

Macnaughton, Thomas B., Stephanie T. Shi, Lucy E. Modahl, and Michael M. C. Lai. "Rolling Circle Replication of Hepatitis Delta Virus RNA Is Carried Out by Two Different Cellular RNA Polymerases." *Journal of Virology* 76, no. 8 (April 2002): 3920–3927. https://doi.org/10.1128/jvi.76.8.3920-3927.2002.

Madigan, Michael T., Deborah O. Jung, Elizabeth A. Karr, W. Matthew Sattley, Laurie A. Achenbach, and Marcel T. J. van der Meer. "Diversity of Anoxygenic Phototrophs in Contrasting Extreme Environments." In *Geothermal Biology and Geochemistry in Yellowstone National Park*, edited by William Parks Inskeep and Timothy R. McDermott (Bozeman, MT: Thermal Biology Institute, 2005), 203–219.

Marais, Armelle, Aurélia Nivault, Chantal Faure, Sébastien Theil, Gwenaëlle Comont, Thierry Candresse, and Marie-France Corio-Costet. "Determination of the Complete Genomic Sequence of Neofusicoccum Luteum Mitovirus 1 (NLMV1), a Novel Mitovirus Associated with a Phytopathogenic Botryosphaeriaceae." *Archives of Virology* 162, no. 8 (August 2017): 2477–2480. https://doi.org/10.1007/s00705-017-3338-9.

Marçais, Guillaume, and Carl Kingsford. "A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of k-Mers." *Bioinformatics* 27, no. 6 (March 2011): 764–770. https://doi.org/10.1093/bioinformatics/btr011.

Mascia, Tiziana, and Donato Gallitelli. "Economic Significance of Satellites." In *Viroids and Satellites* (Elsevier, 2017), 555–563. https://doi.org/10.1016/B978-0-12-801498-1.00051-6.

Masuta, Chikara, and Hanako Shimura. "Satellite RNAs: Their Involvement in Pathogenesis and RNA Silencing." In *Viroids and Satellites* (Elsevier, 2017), 587–596. https://doi.org/10.1016/B978-0-12-801498-1.00054-1.

Mathews, D E, and R D Durbin. "Tagetitoxin Inhibits RNA Synthesis Directed by RNA Polymerases from Chloroplasts and Escherichia Coli." *Journal of Biological Chemistry* 265, no. 1 (January 1990): 493–498. https://doi.org/10.1016/S0021-9258(19)40258-5.

McCutcheon, John P., and Nancy A. Moran. "Extreme Genome Reduction in Symbiotic Bacteria." *Nature Reviews. Microbiology* 10, no. 1 (November 2011): 13–26. https://doi.org/10.1038/nrmicro2670.

McKinney, Wes. "Data Structures for Statistical Computing in Python" (Austin, Texas, 2010), 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.

Meer, Marcel T. J. van der, Christian G. Klatt, Jason Wood, Donald A. Bryant, Mary M. Bateson, Laurens Lammerts, Stefan Schouten, Jaap S. Sinninghe Damsté, Michael T. Madigan, and David M. Ward. "Cultivation and Genomic, Nutritional, and Lipid Biomarker Characterization of Roseiflexus Strains Closely Related to Predominant In Situ Populations Inhabiting Yellowstone

Hot Spring Microbial Mats." *Journal of Bacteriology* 192, no. 12 (June 2010): 3033–3042. https://doi.org/10.1128/JB.01610-09.

Melsted, Páll, and Jonathan K Pritchard. "Efficient Counting of k-Mers in DNA Sequences Using a Bloom Filter." *BMC Bioinformatics* 12, no. 1 (December 2011): 333. https://doi.org/10.1186/1471-2105-12-333.

Mendes, Marta, Daniel Pérez-Hernandez, Jesús Vázquez, Ana V. Coelho, and Celso Cunha. "Proteomic Changes in HEK-293 Cells Induced by Hepatitis Delta Virus Replication." *Journal of Proteomics* 89 (August 2013): 24–38. https://doi.org/10.1016/j.jprot.2013.06.002.

Mills, D. R., F. R. Kramer, and S. Spiegelman. "Complete Nucleotide Sequence of a Replicating RNA Molecule: The Sequence Suggests How Nucleic Acids Exhibit Phenotypes for Selection and Can Evolve to Greater Complexity." *Science* 180, no. 4089 (June 1973): 916–927. https://doi.org/10.1126/science.180.4089.916.

Modahl, Lucy E., and Michael M. C. Lai. "The Large Delta Antigen of Hepatitis Delta Virus Potently Inhibits Genomic but Not Antigenomic RNA Synthesis: A Mechanism Enabling Initiation of Viral Replication." *Journal of Virology* 74, no. 16 (August 2000): 7375–7380. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC112257/.

Modahl, Lucy E., Thomas B. Macnaughton, Nongliao Zhu, Deborah L. Johnson, and Michael M. C. Lai. "RNA-Dependent Replication and Transcription of Hepatitis Delta Virus RNA Involve Distinct Cellular RNA Polymerases." *Molecular and Cellular Biology* 20, no. 16 (August 2000): 6030–6039. https://doi.org/10.1128/MCB.20.16.6030-6039.2000.

Moelling, Karin, and Felix Broecker. "Viroids and the Origin of Life." *International Journal of Molecular Sciences* 22, no. 7 (March 2021): 3476. https://doi.org/10.3390/ijms22073476.

Mölder, Felix, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, et al. "Sustainable Data Analysis with Snakemake." *F1000Research* 10 (January 2021): 33. https://doi.org/10.12688/f1000research.29032.1.

Molina-Serrano, Diego, Loreto Suay, María L. Salvador, Ricardo Flores, and José-Antonio Daròs. "Processing of RNAs of the Family Avsunviroidae in Chlamydomonas Reinhardtii Chloroplasts." *Journal of Virology* 81, no. 8 (April 2007): 4363–4366. https://doi.org/10.1128/JVI.02556-06.

Moraleda, Gloria, and John Taylor. "Host RNA Polymerase Requirements for Transcription of the Human Hepatitis Delta Virus Genome." *Journal of Virology* 75, no. 21 (November 2001): 10161–10169. https://doi.org/10.1128/JVI.75.21.10161-10169.2001.

Moran, Nancy A. "Microbial Minimalism." *Cell* 108, no. 5 (March 2002): 583–586. https://doi.org/10.1016/S0092-8674(02)00665-7.

Morgulis, Aleksandr, E. Michael Gertz, Alejandro A. Schäffer, and Richa Agarwala. "A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences." *Journal of Computational Biology* 13, no. 5 (June 2006): 1028–1040. https://doi.org/10.1089/cmb.2006.13.1028.

Mosig, Axel, Ivo Hofacker, and Peter F Stadler. "Comparative Analysis of Cyclic Sequences: Viroids and Other Small Circular RNAs." In *German Conference on Bioinformatics*, 2006, 83:93–102.

Muhire, Brejnev Muhizi, Arvind Varsani, and Darren Patrick Martin. "SDT: A Virus Classification Tool Based on Pairwise Sequence Alignment and Identity Calculation." Edited by Jens H. Kuhn. *PLoS ONE* 9, no. 9 (September 2014): e108277. https://doi.org/10.1371/journal.pone.0108277.

Mühlbach, Hans-P., and Heinz L. Sänger. "Viroid Replication Is Inhibited by α-Amanitin." *Nature* 278, no. 5700 (March 1979): 185–188. https://doi.org/10.1038/278185a0.

Munson-McGee, Jacob H., Shengyun Peng, Samantha Dewerff, Ramunas Stepanauskas, Rachel J. Whitaker, Joshua S. Weitz, and Mark J. Young. "A Virus or More in (Nearly) Every Cell: Ubiquitous Networks of Virus–Host Interactions in Extreme Environments." *The ISME Journal* 12, no. 7 (July 2018): 1706–1714. https://doi.org/10.1038/s41396-018-0071-7.

Nahas, Michelle K, Timothy J Wilson, Sungchul Hohng, Kaera Jarvie, David M J Lilley, and Taekjip Ha. "Observation of Internal Cleavage and Ligation Reactions of a Ribozyme." *Nature Structural & Molecular Biology* 11, no. 11 (November 2004): 1107–1113. https://doi.org/10.1038/nsmb842.

Naoi, Takashi, and Tatsuji Hataya. "Tolerance Even to Lethal Strain of Potato Spindle Tuber Viroid Found in Wild Tomato Species Can Be Introduced by Crossing." *Plants* 10, no. 3 (March 2021): 575. https://doi.org/10.3390/plants10030575.

Navarro, Beatriz, Ricardo Flores, and Francesco Di Serio. "Advances in Viroid-Host Interactions." *Annual Review of Virology* 8, no. 1 (September 2021): 305–325. https://doi.org/10.1146/annurev-virology-091919-092331.

Navarro, Beatriz, Andreas Gisel, Maria Elena Rodio, Sonia Delgado, Ricardo Flores, and Francesco Di Serio. "Small RNAs Containing the Pathogenic Determinant of a Chloroplast-Replicating Viroid Guide the Degradation of a Host mRNA as Predicted by RNA Silencing." *The Plant Journal* 70, no. 6 (June 2012): 991–1003. https://doi.org/10.1111/j.1365-313X.2012.04940.x.

Navarro, Beatriz, Luisa Rubino, and Francesco Di Serio. "Small Circular Satellite RNAs." In *Viroids and Satellites* (Elsevier, 2017), 659–669. https://doi.org/10.1016/B978-0-12-801498-1.00061-9.

Navarro, J.-A. "Characterization of the Initiation Sites of Both Polarity Strands of a Viroid RNA Reveals a Motif Conserved in Sequence and Structure." *The EMBO Journal* 19, no. 11 (June 2000): 2662–2670. https://doi.org/10.1093/emboj/19.11.2662.

Navarro, José-Antonio, Antonio Vera, and Ricardo Flores. "A Chloroplastic RNA Polymerase Resistant to Tagetitoxin Is Involved in Replication of Avocado Sunblotch Viroid." *Virology* 268, no. 1 (March 2000): 218–225. https://doi.org/10.1006/viro.1999.0161.

Nawrocki, E. P., and S. R. Eddy. "Infernal 1.1: 100-Fold Faster RNA Homology Searches." *Bioinformatics* 29, no. 22 (November 2013): 2933–2935. https://doi.org/10.1093/bioinformatics/btt509.

Neri, Uri, Yuri I. Wolf, Simon Roux, Antonio Pedro Camargo, Benjamin Lee, Darius Kazlauskas, I. Min Chen, et al. "Expansion of the Global RNA Virome Reveals Diverse Clades of Bacteriophages." *Cell* 185, no. 21 (October 2022): 4023–4037.e18. https://doi.org/10.1016/j.cell.2022.08.023.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies." *Molecular Biology and Evolution* 32, no. 1 (January 2015): 268–274. https://doi.org/10.1093/molbev/msu300.

Nie, Xianzhou, and Rudra P. Singh. "Coleus Blumei Viroids." In *Viroids and Satellites* (Elsevier, 2017), 289–295. https://doi.org/10.1016/B978-0-12-801498-1.00027-9.

Niro, Grazia Anna, and Antonina Smedile. "Current Concept in the Pathophysiology of Hepatitis Delta Infection." *Current Infectious Disease Reports* 14, no. 1 (February 2012): 9–14. https://doi.org/10.1007/s11908-011-0233-5.

Nohales, María-Ángeles, Ricardo Flores, and José-Antonio Daròs. "Viroid RNA Redirects Host DNA Ligase 1 to Act as an RNA Ligase." *Proceedings of the National Academy of Sciences* 109, no. 34 (August 2012): 13805–13810. https://doi.org/10.1073/pnas.1206187109.

Nohales, María-Ángeles, Diego Molina-Serrano, Ricardo Flores, and José-Antonio Daròs. "Involvement of the Chloroplastic Isoform of tRNA Ligase in the Replication of Viroids Belonging to the Family Avsunviroidae." *Journal of Virology* 86, no. 15 (August 2012): 8269–8276. https://doi.org/10.1128/JVI.00629-12.

One Thousand Plant Transcriptomes Initiative. "One Thousand Plant Transcriptomes and the Phylogenomics of Green Plants." *Nature* 574, no. 7780 (October 2019): 679–685. https://doi.org/10.1038/s41586-019-1693-2.

Paez-Espino, David, Emiley A. Eloe-Fadrosh, Georgios A. Pavlopoulos, Alex D. Thomas, Marcel Huntemann, Natalia Mikhailova, Edward Rubin, Natalia N. Ivanova, and Nikos C. Kyrpides. "Uncovering Earth's Virome." *Nature* 536, no. 7617 (August 2016): 425–430. https://doi.org/10.1038/nature19094.

Palukaitis, Peter. “Satellite Taxonomy.” In *Viroids and Satellites* (Elsevier, 2017), 615–622. https://doi.org/10.1016/B978-0-12-801498-1.00057-7.

Pan, Calvin, Robert Gish, Ira M. Jacobson, Ke-Qin Hu, Heiner Wedemeyer, and Paul Martin. “Diagnosis and Management of Hepatitis Delta Virus Infection.” *Digestive Diseases and Sciences* 68, no. 8 (August 2023): 3237–3248. https://doi.org/10.1007/s10620-023-07960-y.

Parakh, Dattaraj B., Shuifang Zhu, and Teruo Sano. “Geographical Distribution of Viroids in South, Southeast, and East Asia.” In *Viroids and Satellites* (Elsevier, 2017), 507–518. https://doi.org/10.1016/B978-0-12-801498-1.00047-4.

Paraskevopoulou, Sofia, Fabian Pirzer, Nora Goldmann, Julian Schmid, Victor Max Corman, Lina Theresa Gottula, Simon Schroeder, et al. “Mammalian Deltavirus Without Hepadnavirus Coinfection in the Neotropical Rodent Proechimys Semispinosus.” *Proceedings of the National Academy of Sciences* 117, no. 30 (July 2020): 17977–17983. https://doi.org/10.1073/pnas.2006750117.

Pei, Jimin, and Nick V. Grishin. “PROMALS3D: Multiple Protein Sequence Alignment Enhanced with Evolutionary and Three-Dimensional Structural Information.” In *Multiple Sequence Alignment Methods*, edited by David J Russell (Totowa, NJ: Humana Press, 2014), 1079:263–271. https://doi.org/10.1007/978-1-62703-646-7_17.

Pena, M. de la, and I. Garcia-Robles. “Ubiquitous Presence of the Hammerhead Ribozyme Motif Along the Tree of Life.” *RNA* 16, no. 10 (October 2010): 1943–1950. https://doi.org/v.

Peña, Marcos de la. “Circular RNAs Biogenesis in Eukaryotes Through Self-Cleaving Hammerhead Ribozymes.” *Advances in Experimental Medicine and Biology* 1087 (2018): 53–63. https://doi.org/10.1007/978-981-13-1426-1_5.

Peña, Marcos de la, Raquel Ceprián, John L. Casey, and Amelia Cervera. “Hepatitis Delta Virus-Like Circular RNAs from Diverse Metazoans Encode Conserved Hammerhead Ribozymes.” *Virus Evolution* 7, no. 1 (January 2021): veab016. https://doi.org/10.1093/ve/veab016.

Peña, Marcos de la, Raquel Ceprián, and Amelia Cervera. "A Singular and Widespread Group of Mobile Genetic Elements: RNA Circles with Autocatalytic Ribozymes." *Cells* 9, no. 12 (November 2020): E2555. https://doi.org/10.3390/cells9122555.

Peña, Marcos de la, and Amelia Cervera. "Circular RNAs with Hammerhead Ribozymes Encoded in Eukaryotic Genomes: The Enemy at Home." *RNA Biology* 14, no. 8 (August 2017): 985–991. https://doi.org/10.1080/15476286.2017.1321730.

Perez-Vargas, Jimena, Fouzia Amirache, Bertrand Boson, Chloé Mialon, Natalia Freitas, Camille Sureau, Floriane Fusil, and François-Loïc Cosset. "Enveloped Viruses Distinct from HBV Induce Dissemination of Hepatitis D Virus in Vivo." *Nature Communications* 10 (May 2019): 2098. https://doi.org/10.1038/s41467-019-10117-z.

Perreault, Jonathan, Zasha Weinberg, Adam Roth, Olivia Popescu, Pascal Chartrand, Gerardo Ferbeyre, and Ronald R. Breaker. "Identification of Hammerhead Ribozymes in All Domains of Life Reveals Novel Structural Variations." Edited by Wyeth W. Wasserman. *PLoS Computational Biology* 7, no. 5 (May 2011): e1002031. https://doi.org/10.1371/journal.pcbi.1002031.

Piazzolla, P., C. Vovlas, and L. Rubino. "Symptom Regulation Induced by Chicory Yellow Mottle Virus Satellite-Like RNA." *Journal of Phytopathology* 115, no. 2 (February 1986): 124–129. https://doi.org/10.1111/j.1439-0434.1986.tb00868.x.

Podleckis, Edward V. "Geographical Distribution of Viroids in the Americas." In *Viroids and Satellites* (Elsevier, 2017), 459–472. https://doi.org/10.1016/B978-0-12-801498-1.00043-7.

Polson, Andrew G., I. I. I. Herbert L. Ley, Brenda L. Bass, and John L. Casey. "Hepatitis Delta Virus RNA Editing Is Highly Specific for the Amber/W Site and Is Suppressed by Hepatitis Delta Antigen." *Molecular and Cellular Biology* 18, no. 4 (April 1998): 1919. https://doi.org/10.1128/mcb.18.4.1919.

Qi, Yijun, and Biao Ding. "Differential Subnuclear Localization of RNA Strands of Opposite Polarity Derived from an Autonomously Replicating Viroid." *The Plant Cell* 15, no. 11 (November 2003): 2566–2577. https://doi.org/10.1105/tpc.016576.

Qin, Yangmei, Tingting Xu, Wenbo Lin, Qingjie Jia, Qiushun He, Ke Liu, Juan Du, et al. "Reference-Free and de Novo Identification of Circular RNAs." Preprint, April 2020. https://doi.org/10.1101/2020.04.21.050617.

Quake, Stephen R. "The Cell as a Bag of RNA." *Trends in Genetics* 37, no. 12 (December 2021): 1064–1068. https://doi.org/10.1016/j.tig.2021.08.003.

Rao, A. L. N., and Kriton Kalantidis. "Virus-Associated Small Satellite RNAs and Viroids Display Similarities in Their Replication Strategies." *Virology* 479–480 (May 2015): 627–636. https://doi.org/10.1016/j.virol.2015.02.018.

Raymer, W. B., and T. O. Diener. "Potato Spindle Tuber Virus: A Plant Virus with Properties of a Free Nucleic Acid." *Virology* 37, no. 3 (March 1969): 343–350. https://doi.org/10.1016/0042-6822(69)90218-9.

Rice, P., I. Longden, and A. Bleasby. "EMBOSS: The European Molecular Biology Open Software Suite." *Trends in Genetics* 16, no. 6 (June 2000): 276–277. https://doi.org/10.1016/s0168-9525(00)02024-2.

Rizzetto, M., M. G. Canese, S. Aricò, O. Crivelli, C. Trepo, F. Bonino, and G. Verme. "Immunofluorescence Detection of New Antigen-Antibody System (Delta/Anti-Delta) Associated to Hepatitis B Virus in Liver and in Serum of HBsAg Carriers." *Gut* 18, no. 12 (December 1977): 997–1003. https://doi.org/10.1136/gut.18.12.997.

Rocheleau, Lynda, and Martin Pelchat. "The Subviral RNA Database: A Toolbox for Viroids, the Hepatitis Delta Virus and Satellite RNAs Research." *BMC Microbiology* 6 (March 2006): 24. https://doi.org/10.1186/1471-2180-6-24.

Rodio, Maria-Elena, Sonia Delgado, Angelo De Stradis, María-Dolores Gómez, Ricardo Flores, and Francesco Di Serio. "A Viroid RNA with a Specific Structural Motif Inhibits Chloroplast Development." *The Plant Cell* 19, no. 11 (December 2007): 3610–3626. https://doi.org/10.1105/tpc.106.049775.

Rodriguez, M. Judith B., Ganesan Vadamalai, and John W. Randles. "Economic Significance of Palm Tree Viroids." In *Viroids and Satellites* (Elsevier, 2017), 39–49. https://doi.org/10.1016/B978-0-12-801498-1.00004-8.

Romeo, Raffaella, Ersilio Del Ninno, Mariagrazia Rumi, Antonio Russo, Angelo Sangiovanni, Roberto de Franchis, Guido Ronchi, and Massimo Colombo. "A 28-Year Study of the Course of Hepatitis Delta Infection: A Risk Factor for Cirrhosis and Hepatocellular Carcinoma." *Gastroenterology* 136, no. 5 (May 2009): 1629–1638. https://doi.org/10.1053/j.gastro.2009.01.052.

Roth, Adam, Zasha Weinberg, Andy G. Y. Chen, Peter B. Kim, Tyler D. Ames, and Ronald R. Breaker. "A Widespread Self-Cleaving Ribozyme Class Is Revealed by Bioinformatics." *Nature Chemical Biology* 10, no. 1 (January 2014): 56–60. https://doi.org/10.1038/nchembio.1386.

Rumpf, Andreas. *Mastering Nim: A Complete Guide to the Programming Language*, 2022.

Ryu, W S, H J Netter, M Bayer, and J Taylor. "Ribonucleoprotein Complexes of Hepatitis Delta Virus." *Journal of Virology* 67, no. 6 (June 1993): 3281–3287. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC237669/.

Salehi-Ashtiani, Kourosh, Andrej Lupták, Alexander Litovchick, and Jack W. Szostak. "A Genomewide Search for Ribozymes Reveals an HDV-Like Sequence in the Human CPEB3 Gene." *Science* 313, no. 5794 (September 2006): 1788–1792. https://doi.org/10.1126/science.1129308.

Salehi-Ashtiani, Kourosh, and Jack W. Szostak. "In Vitro Evolution Suggests Multiple Origins for the Hammerhead Ribozyme." *Nature* 414, no. 6859 (November 2001): 82–84. https://doi.org/10.1038/35102081.

Sanger, H. L., G. Klotz, D. Riesner, H. J. Gross, and A. K. Kleinschmidt. "Viroids Are Single-Stranded Covalently Closed Circular RNA Molecules Existing as Highly Base-Paired Rod-Like Structures." *Proceedings of the National Academy of Sciences* 73, no. 11 (November 1976): 3852–3856. https://doi.org/10.1073/pnas.73.11.3852.

Sanjuán, Rafael. "The Social Life of Viruses." *Annual Review of Virology* 8, no. 1 (September 2021): 183–199. https://doi.org/10.1146/annurev-virology-091919-071712.

Sato, Shuji, Cromwell Cornillez-Ty, and David W. Lazinski. "By Inhibiting Replication, the Large Hepatitis Delta Antigen Can Indirectly Regulate Amber/W Editing and Its Own Expression." *Journal of Virology* 78, no. 15 (August 2004): 8120–8134. https://doi.org/10.1128/JVI.78.15.8120-8134.2004.

Sayers, Eric W, Mark Cavanaugh, Karen Clark, Kim D Pruitt, Conrad L Schoch, Stephen T Sherry, and Ilene Karsch-Mizrachi. "GenBank." *Nucleic Acids Research* 49, no. D1 (January 2021): D92–D96. https://doi.org/10.1093/nar/gkaa1023.

Schindler, I.-M., and H.-P. Mühlbach. "Involvement of Nuclear DNA-dependent RNA Polymerases in Potato Spindle Tuber Viroid Replication: A Reevaluation." *Plant Science* 84, no. 2 (January 1992): 221–229. https://doi.org/10.1016/0168-9452(92)90138-C.

Seith, Daniel D., Jamie L. Bingaman, Andrew J. Veenis, Aileen C. Button, and Philip C. Bevilacqua. "Elucidation of Catalytic Strategies of Small Nucleolytic Ribozymes From Comparative Analysis of Active Sites." *ACS Catalysis* 8, no. 1 (January 2018): 314–327. https://doi.org/10.1021/acscatal.7b02976.

Senkevich, Tatiana G., Natalya Yutin, Yuri I. Wolf, Eugene V. Koonin, and Bernard Moss. "Ancient Gene Capture and Recent Gene Loss Shape the Evolution of Orthopoxvirus-Host Interaction Genes." *mBio* 12, no. 4 (July 2021): e01495–21. https://doi.org/10.1128/mBio.01495-21.

Serra, Pedro, Alberto Carbonell, Beatriz Navarro, Selma Gago-Zachert, Shifang Li, Francesco Di Serio, and Ricardo Flores. "Symptomatic Plant Viroid Infections in Phytopathogenic Fungi: A Request for a Critical Reassessment." *Proceedings of the National Academy of Sciences* 117, no. 19 (May 2020): 10126–10128. https://doi.org/10.1073/pnas.1922249117.

Serra, Pedro, Amber Messmer, Daniel Sanderson, Delano James, and Ricardo Flores. "Apple Hammerhead Viroid-Like RNA Is a Bona Fide Viroid: Autonomous Replication and Structural Features Support Its Inclusion as a New Member in the Genus Pelamoviroid." *Virus Research* 249 (April 2018): 8–15. https://doi.org/10.1016/j.virusres.2018.03.001.

Seward, Julian. "Bzip2 and Libbzip2," December 2001. https://perso.crans.org/pklein/Info/emacs/manual.pdf.

Shen, Wei, Shuai Le, Yan Li, and Fuquan Hu. "SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation." Edited by Quan Zou. *PLOS ONE* 11, no. 10 (October 2016): e0163962. https://doi.org/10.1371/journal.pone.0163962.

Simmonds, Peter, Mike J. Adams, Mária Benkő, Mya Breitbart, J. Rodney Brister, Eric B. Carstens, Andrew J. Davison, et al. "Virus Taxonomy in the Age of Metagenomics." *Nature Reviews Microbiology* 15, no. 3 (March 2017): 161–168. https://doi.org/10.1038/nrmicro.2016.177.

Singh, Urminder, and Eve Syrkin Wurtele. "Orfipy: A Fast and Flexible Tool for Extracting ORFs." Edited by Inanc Birol. *Bioinformatics* 37, no. 18 (September 2021): 3019–3020. https://doi.org/10.1093/bioinformatics/btab090.

Soliman, T., M. C. M. Mourits, A. G. J. M. Oude Lansink, and W. van der Werf. "Quantitative Economic Impact Assessment of an Invasive Plant Disease Under Uncertainty – A Case Study for Potato Spindle Tuber Viroid (PSTVd) Invasion into the European Union." *Crop Protection* 40 (October 2012): 28–35. https://doi.org/10.1016/j.cropro.2012.04.019.

Spiegelman, S. "An Approach to the Experimental Analysis of Precellular Evolution." *Quarterly Reviews of Biophysics* 4, no. 2–3 (August 1971): 213–253. https://doi.org/10.1017/S0033583500000639.

Spiegelman, S., I. Haruna, N. R. Pace, D. R. Mills, D. H. L. Bishop, J. R. Claybrook, and R. Peterson. "Studies in the Replication of Viral RNA." *Journal of Cellular Physiology* 70, no. S1 (1967): 35–64. https://doi.org/10.1002/jcp.1040700405.

Steinegger, Martin, and Johannes Söding. "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets." *Nature Biotechnology* 35, no. 11 (November 2017): 1026–1028. https://doi.org/10.1038/nbt.3988.

Stollar, B. David, and T. O. Diener. "Potato Spindle Tuber Viroid: V. Failure of Immunological Tests to Disclose Double-Stranded RNA or RNA-DNA Hybrids." *Virology* 46, no. 1 (October 1971): 168–170. https://doi.org/10.1016/0042-6822(71)90017-1.

Sun, Liying, and Ahmed Hadidi. "Mycoviroids: Fungi as Hosts and Vectors of Viroids." *Cells* 11, no. 8 (April 2022): 1335. https://doi.org/10.3390/cells11081335.

Sun, Mike, Susan Siemsen, William Campbell, Pablo Guzman, Robert Davidson, Jonathan L. Whitworth, Terry Bourgoin, et al. "Survey of Potato Spindle Tuber Viroid in Seed Potato Growing Areas of the United States." *American Journal of Potato Research* 81, no. 3 (May 2004): 227–231. https://doi.org/10.1007/BF02871753.

Sureau, Camille, and Francesco Negro. "The Hepatitis Delta Virus: Replication and Pathogenesis." *Journal of Hepatology* 64, no. 1 (April 2016): S102–S116. https://doi.org/10.1016/j.jhep.2016.02.013.

Sutela, Suvi, Marco Forgia, Eeva J Vainio, Marco Chiapello, Stefania Daghino, Marta Vallino, Elena Martino, Mariangela Girlanda, Silvia Perotto, and Massimo Turina. "The Virome from a Collection of Endomycorrhizal Fungi Reveals New Viral Taxa with Unprecedented Genome Organization." *Virus Evolution* 6, no. 2 (July 2020): veaa076. https://doi.org/10.1093/ve/veaa076.

Suzek, B. E., Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, and the UniProt Consortium. "UniRef Clusters: A Comprehensive and Scalable Alternative for Improving Sequence Similarity Searches." *Bioinformatics* 31, no. 6 (March 2015): 926–932. https://doi.org/10.1093/bioinformatics/btu739.

Syller, Jerzy, Waldemar Marczewski, and Jerzy Pawłowicz. "Transmission by Aphids of Potato Spindle Tuber Viroid Encapsidated by Potato Leafroll Lu-

teovirus Particles." *European Journal of Plant Pathology* 103, no. 3 (1997): 285–289. https://doi.org/10.1023/A:1008648822190.

Szathmáry, Eörs. "The Evolution of Replicators." Edited by B. Charlesworth and P. H. Harvey. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 355, no. 1403 (November 2000): 1669–1676. https://doi.org/10.1098/rstb.2000.0730.

Szathmáry, Eörs, and John Maynard Smith. "From Replicators to Reproducers: The First Major Transitions Leading to Life." *Journal of Theoretical Biology* 187, no. 4 (August 1997): 555–571. https://doi.org/10.1006/jtbi.1996.0389.

Szirovicza, Leonora, Udo Hetzel, Anja Kipar, Luis Martinez-Sobrido, Olli Vapalahti, and Jussi Hepojoki. "Snake Deltavirus Utilizes Envelope Proteins of Different Viruses To Generate Infectious Particles." *mBio* 11, no. 2 (March 2020): e03250–19. https://doi.org/10.1128/mBio.03250-19.

Tadini, Luca, Nicolaj Jeran, Carlotta Peracchio, Simona Masiero, Monica Colombo, and Paolo Pesaresi. "The Plastid Transcription Machinery and Its Coordination with the Expression of Nuclear Genome: Plastid-Encoded Polymerase, Nuclear-Encoded Polymerase and the Genomes Uncoupled 1-Mediated Retrograde Communication." *Philosophical Transactions of the Royal Society B: Biological Sciences* 375, no. 1801 (May 2020): 20190399. https://doi.org/10.1098/rstb.2019.0399.

Taylor, J. M. "Structure and Replication of Hepatitis Delta Virus RNA." *Current Topics in Microbiology and Immunology* 307 (2006): 1–23. https://doi.org/10.1007/3-540-29802-9_1.

Thomas, Scott C., Kevin O. Tamadonfar, Cale O. Seymour, Dengxun Lai, Jeremy A. Dodsworth, Senthil K. Murugapiran, Emiley A. Eloe-Fadrosh, Paul Dijkstra, and Brian P. Hedlund. "Position-Specific Metabolic Probing and Metagenomics of Microbial Communities Reveal Conserved Central Carbon Metabolic Network Activities at High Temperatures." *Frontiers in Microbiology* 10 (July 2019): 1427. https://doi.org/10.3389/fmicb.2019.01427.

Tjhung, Katrina F., Maxim N. Shokhirev, David P. Horning, and Gerald F. Joyce. "An RNA Polymerase Ribozyme That Synthesizes Its Own Ancestor." *Proceedings of the National Academy of Sciences* 117, no. 6 (February 2020): 2906–2913. https://doi.org/10.1073/pnas.1914282117.

Traag, V. A., L. Waltman, and N. J. van Eck. "From Louvain to Leiden: Guaranteeing Well-Connected Communities." *Scientific Reports* 9, no. 1 (December 2019): 5233. https://doi.org/10.1038/s41598-019-41695-z.

Tseligka, Eirini D., Sophie Clément, and Francesco Negro. "HDV Pathogenesis: Unravelling Ariadne's Thread." *Viruses* 13, no. 5 (April 2021): 778. https://doi.org/10.3390/v13050778.

Tseng, Chung-Hsin, and Michael M. Lai. "Hepatitis Delta Virus RNA Replication." *Viruses* 1, no. 3 (November 2009): 818–831. https://doi.org/10.3390/v1030818.

Turina, Massimo, Benjamin D. Lee, Sead Sabanadzovic, Eeva Vainio, Beatriz Navarro, Peter Simmonds, Jens H. Kuhn, Marcos de la Peña, Leticia Botella, and Mart Krupovic. "Create One New Phylum, *Ambiviricota*, Including One New Class, One New Order, Four New Families, Four New Genera, and 20 New Species, in Kingdom *Orthornavirae* (Realm *Riboviria*)." Proposal (International Committee on Taxonomy of Viruses, April 2023). https://ictv.global/system/files/proposals/approved/Fungal_and_Protist_viruses/2023.007F.Ambiviricota_nphy.docx.

Vadamalai, Ganesan, Sathis S. Thanarajoo, Hendry Joseph, Lih L. Kong, and John W. Randles. "Coconut Cadang-Cadang Viroid and Coconut Tinangaja Viroid." In *Viroids and Satellites* (Elsevier, 2017), 263–273. https://doi.org/10.1016/B978-0-12-801498-1.00025-5.

Vera, A., J. A. Daròs, R. Flores, and C. Hernández. "The DNA of a Plant Retroviroid-Like Element Is Fused to Different Sites in the Genome of a Plant Pararetrovirus and Shows Multiple Forms with Sequence Deletions." *Journal of Virology* 74, no. 22 (November 2000): 10390–10400. https://doi.org/10.1128/jvi.74.22.10390-10400.2000.

Verhoeven, Jacobus Th.J., Rosemarie W. Hammond, and Giuseppe Stancanelli. "Economic Significance of Viroids in Ornamental Crops." In *Viroids and Satellites* (Elsevier, 2017), 27–38. https://doi.org/10.1016/B978-0-12-801498-1.00003-6.

Verhoeven, Jacobus Th.J., Ellis T. M. Meekes, Johanna W. Roenhorst, Ricardo Flores, and Pedro Serra. "Dahlia Latent Viroid." In *Viroids and Satellites* (Elsevier, 2017), 211–216. https://doi.org/10.1016/B978-0-12-801498-1.00020-6.

Walker, D. Roland, and Eugene V. Koonin. "SEALS: A System for Easy Analysis of Lots of Sequences." In *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology* (AAAI Press, 1997), 333–339.

Walker, Peter J., Stuart G. Siddell, Elliot J. Lefkowitz, Arcady R. Mushegian, Evelien M. Adriaenssens, Poliane Alfenas-Zerbini, Andrew J. Davison, et al. "Changes to Virus Taxonomy and to the International Code of Virus Classification and Nomenclature Ratified by the International Committee on Taxonomy of Viruses (2021)." *Archives of Virology* 166, no. 9 (September 2021): 2633–2648. https://doi.org/10.1007/s00705-021-05156-1.

Wan, Yu. "readSimulator," December 2019. https://github.com/wanyuac/readSimulator.

Wang, Kang-Sheng, Qui-Lim Choo, Amy J. Weiner, Jing-Hsiung Ou, Richard C. Najarian, Richard M. Thayer, Guy T. Mullenbach, Katherine J. Denniston, John L. Gerin, and Michael Houghton. "Structure, Sequence and Expression of the Hepatitis Delta (δ) Viral Genome." *Nature* 323, no. 6088 (October 1986): 508–514. https://doi.org/10.1038/323508a0.

Wang, Ying. "Current View and Perspectives in Viroid Replication." *Current Opinion in Virology* 47 (April 2021): 32–37. https://doi.org/10.1016/j.coviro.2020.12.004.

Ward, L. I., J. Tang, S. Veerakone, B. D. Quinn, S. J. Harper, C. Delmiglio, and G. R. G. Clover. "First Report of Potato Spindle Tuber Viroid in Cape Gooseberry

(Physalis Peruviana) in New Zealand." *Plant Disease* 94, no. 4 (April 2010): 479. https://doi.org/10.1094/PDIS-94-4-0479A.

Warren, J. G., J. Mercado, and D. Grace. "Occurrence of Hop Latent Viroid Causing Disease in Cannabis Sativa in California." *Plant Disease* 103, no. 10 (October 2019): 2699–2699. https://doi.org/10.1094/PDIS-03-19-0530-PDN.

Webb, Chiu-Ho T., Nathan J. Riccitelli, Dana J. Ruminski, and Andrej Lupták. "Widespread Occurrence of Self-Cleaving Ribozymes." *Science* 326, no. 5955 (November 2009): 953. https://doi.org/10.1126/science.1178084.

Wedemeyer, Heiner, and Michael P. Manns. "Epidemiology, Pathogenesis and Management of Hepatitis D: Update and Challenges Ahead." *Nature Reviews Gastroenterology & Hepatology* 7, no. 1 (January 2010): 31–40. https://doi.org/10.1038/nrgastro.2009.205.

Wei, Shuang, Ruiling Bian, Ida Bagus Andika, Erbo Niu, Qian Liu, Hideki Kondo, Liu Yang, et al. "Reply to Serra Et Al.: Nucleotide Substitutions in Plant Viroid Genomes That Multiply in Phytopathogenic Fungi." *Proceedings of the National Academy of Sciences* 117, no. 19 (May 2020): 10129–10130. https://doi.org/10.1073/pnas.2001670117.

———, et al. "Symptomatic Plant Viroid Infections in Phytopathogenic Fungi." *Proceedings of the National Academy of Sciences* 116, no. 26 (June 2019): 13042–13050. https://doi.org/10.1073/pnas.1900762116.

Weinberg, Christina E, V Janett Olzog, Iris Eckert, and Zasha Weinberg. "Identification of over 200-Fold More Hairpin Ribozymes Than Previously Known in Diverse Circular RNAs." *Nucleic Acids Research* 49, no. 11 (June 2021): 6375–6388. https://doi.org/10.1093/nar/gkab454.

Weinberg, Christina E, Zasha Weinberg, and Christian Hammann. "Novel Ribozymes: Discovery, Catalytic Mechanisms, and the Quest to Understand Biological Function." *Nucleic Acids Research* 47, no. 18 (October 2019): 9480–9494. https://doi.org/10.1093/nar/gkz737.

Weinberg, Zasha, Peter B. Kim, Tony H. Chen, Sanshu Li, Kimberly A. Harris, Christina E. Lünse, and Ronald R. Breaker. "New Classes of Self-Cleaving

Ribozymes Revealed by Comparative Genomics Analysis." *Nature Chemical Biology* 11, no. 8 (August 2015): 606–610. https://doi.org/10.1038/nchembio.1846.

Wernegreen, Jennifer J. "For Better or Worse: Genomic Consequences of Intracellular Mutualism and Parasitism." *Current Opinion in Genetics & Development* 15, no. 6 (December 2005): 572–583. https://doi.org/10.1016/j.gde.2005.09.013.

Wheeler, Travis J., and Sean R. Eddy. "Nhmmer: DNA Homology Search with Profile HMMs." *Bioinformatics (Oxford, England)* 29, no. 19 (October 2013): 2487–2489. https://doi.org/10.1093/bioinformatics/btt403.

Wickham, Hadley. *Ggplot2* Use R! (Cham: Springer International Publishing, 2016). https://doi.org/10.1007/978-3-319-24277-4.

Wille, Michelle, Hans Netter, Margaret Littlejohn, Lilly Yuen, Mang Shi, John-Sebastian Eden, Marcel Klaassen, Edward Holmes, and Aeron Hurt. "A Divergent Hepatitis D-Like Agent in Birds." *Viruses* 10, no. 12 (December 2018): 720. https://doi.org/10.3390/v10120720.

Witte, C.-P., Q. H. Le, T. Bureau, and A. Kumar. "Terminal-Repeat Retrotransposons in Miniature (TRIM) Are Involved in Restructuring Plant Genomes." *Proceedings of the National Academy of Sciences* 98, no. 24 (November 2001): 13778–13783. https://doi.org/10.1073/pnas.241341898.

Wolf, Yuri I., and Eugene V. Koonin. "Genome Reduction as the Dominant Mode of Evolution." *BioEssays* 35, no. 9 (September 2013): 829–837. https://doi.org/10.1002/bies.201300037.

Wolf, Yuri I., Sukrit Silas, Yongjie Wang, Shuang Wu, Michael Bocek, Darius Kazlauskas, Mart Krupovic, Andrew Fire, Valerian V. Dolja, and Eugene V. Koonin. "Doubling of the Known Set of RNA Viruses by Metagenomic Analysis of an Aquatic Virome." *Nature Microbiology*, July 2020. https://doi.org/10.1038/s41564-020-0755-4.

Wu, Jian, and David M. Bisaro. "Biased Pol II Fidelity Contributes to Conservation of Functional Domains in the Potato Spindle Tuber Viroid Genome." *PLoS*

Pathogens 16, no. 12 (December 2020): e1009144. https://doi.org/10.1371/journal.ppat.1009144.

Wu, Jian-Guo, Wen-Jun Lu, and Po Tien. "Multiplication of Velvet Tobacco Mottle Virus in Nicotiana Clevelandii Protoplasts Is Resistant to α-Amanitin." *Journal of General Virology* 67, no. 12 (December 1986): 2757–2762. https://doi.org/10.1099/0022-1317-67-12-2757.

Wu, Jian, Cuiji Zhou, James Li, Chun Li, Xiaorong Tao, Neocles B Leontis, Craig L Zirbel, David M Bisaro, and Biao Ding. "Functional Analysis Reveals G/U Pairs Critical for Replication and Trafficking of an Infectious Non-Coding Viroid RNA." *Nucleic Acids Research* 48, no. 6 (February 2020): 3134–3155. https://doi.org/10.1093/nar/gkaa100.

Wu, Qingfa, Ying Wang, Mengji Cao, Vitantonio Pantaleo, Joszef Burgyan, Wan-Xiang Li, and Shou-Wei Ding. "Homology-Independent Discovery of Replicating Pathogenic Circular RNAs by Deep Sequencing and a New Computational Algorithm." *Proceedings of the National Academy of Sciences* 109, no. 10 (March 2012): 3938–3943. https://doi.org/10.1073/pnas.1117815109.

Wüsthoff, Kevin-Phil, and Gerhard Steger. "Conserved Motifs and Domains in Members of Pospiviroidae." *Cells* 11, no. 2 (January 2022): 230. https://doi.org/10.3390/cells11020230.

Yamaguchi, Yuki, Julija Filipovska, Keiichi Yano, Akiko Furuya, Naoto Inukai, Takashi Narita, Tadashi Wada, Seiji Sugimoto, Maria M. Konarska, and Hiroshi Handa. "Stimulation of RNA Polymerase II Elongation by Hepatitis Delta Antigen." *Science* 293, no. 5527 (July 2001): 124–127. https://doi.org/10.1126/science.1057925.

Yamaguchi, Yuki, and Hiroshi Handa. "Hepatitis Delta Antigen and RNA Polymerase II." In *Madame Curie Bioscience Database* (Landes Bioscience, 2013). https://www.ncbi.nlm.nih.gov/books/NBK6547/.

Yamaguchi, Yuki, Takashi Mura, Sittinan Chanarat, Sachiko Okamoto, and Hiroshi Handa. "Hepatitis Delta Antigen Binds to the Clamp of RNA Polymerase II

and Affects Transcriptional Fidelity." *Genes to Cells* 12, no. 7 (2007): 863–875. https://doi.org/10.1111/j.1365-2443.2007.01094.x.

Zahid, Kiran, Jian-Hua Zhao, Neil A. Smith, Ulrike Schumann, Yuan-Yuan Fang, Elizabeth S. Dennis, Ren Zhang, Hui-Shan Guo, and Ming-Bo Wang. "Nicotiana Small RNA Sequences Support a Host Genome Origin of Cucumber Mosaic Virus Satellite RNA." *PLoS Genetics* 11, no. 1 (January 2015): e1004906. https://doi.org/10.1371/journal.pgen.1004906.

Zayed, Ahmed A., James M. Wainaina, Guillermo Dominguez-Huerta, Eric Pelletier, Jiarong Guo, Mohamed Mohssen, Funing Tian, et al. "Cryptic and Abundant Marine Viruses at the Evolutionary Origins of Earth's RNA Virome." *Science* 376, no. 6589 (April 2022): 156–162. https://doi.org/10.1126/science.abm5847.

Zhang, Chunfen, Honggao Yan, and Zachary F. Burton. "Combinatorial Control of Human RNA Polymerase II (RNAP II) Pausing and Transcript Cleavage by Transcription Factor IIF, Hepatitis Delta Antigen, and Stimulatory Factor II." *The Journal of Biological Chemistry* 278, no. 50 (December 2003): 50101–50111. https://doi.org/10.1074/jbc.M307590200.

Zhang, Zhixiang, Shuishui Qi, Nan Tang, Xinxin Zhang, Shanshan Chen, Pengfei Zhu, Lin Ma, et al. "Discovery of Replicating Circular RNAs by RNA-Seq and Computational Algorithms." Edited by Biao Ding. *PLOS Pathogens* 10, no. 12 (December 2014): e1004553. https://doi.org/10.1371/journal.ppat.1004553.

Zheludev, Ivan N., Robert C. Edgar, Maria Jose Lopez-Galiano, Marcos De La Peña, Artem Babaian, Ami S. Bhatt, and Andrew Z. Fire. "Viroid-Like Colonists of Human Microbiomes." *Cell* 187, no. 23 (November 2024): 6521–6536.e18. https://doi.org/10.1016/j.cell.2024.09.033.

Zhong, Xuehua, Anthony J. Archual, Amy A. Amin, and Biao Ding. "A Genomic Map of Viroid RNA Motifs Critical for Replication and Systemic Trafficking." *The Plant Cell* 20, no. 1 (2008): 35–47. https://www.jstor.org/stable/25224211.

Zuccola, Harmon J, James E Rozzelle, Stanley M Lemon, Bruce W Erickson, and James M Hogle. "Structural Basis of the Oligomerization of Hepatitis Delta Antigen." *Structure* 6, no. 7 (July 1998): 821–830. https://doi.org/10.1016/S0969-2126(98)00084-7.

Statement on re-use of published materials

The content of this thesis is original except as noted below.

- Chapter 1 re-uses text and figures with modifications from Lee et al.[1] and Lee and Koonin[2].
- Chapter 2 reuses text and figures with modifications from Lee et al.[3]
- Chapter 3 reuses text with significant expansion from Benjamin D. Lee et al.[4]
- Chapter 4 reuses text and figures with modifications from Lee et al.[5]
- Section 6.7 and Section 6.8 re-uses text from Lee and Koonin[6] with extensive modifications.

All coauthors of the papers included in this thesis have given permission for their work to be included.

[1] "ViroidDB."
[2] "Viroids and Viroid-like Circular RNAs."
[3] "ViroidDB."
[4] "Mining Metatranscriptomes Reveals a Vast World of Viroid-Like Circular RNAs," *Cell* 186, no. 3 (February 2023): 646–661.e4, doi:10.1016/j.cell.2022.12.039.
[5] "Mining Metatranscriptomes Reveals a Vast World of Viroid-Like Circular RNAs."
[6] "Viroids and Viroid-like Circular RNAs."