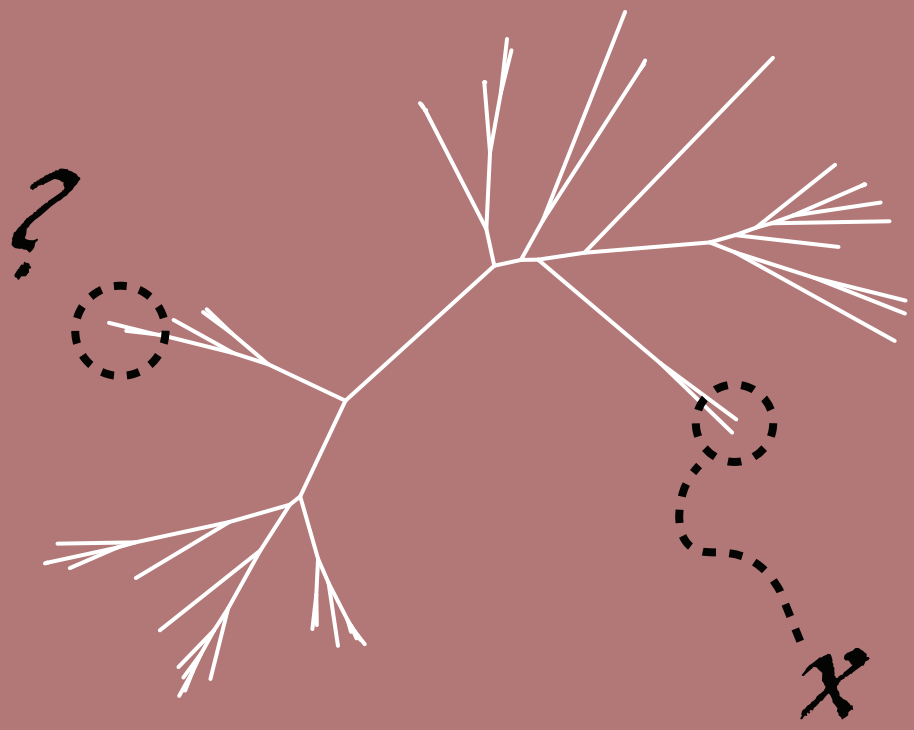
Computational discovery of viruses and their hosts

Kinsella, C.M.

Publication date
2023
Document Version
Final published version

Computational discovery of viruses and their hosts



Cormac M. Kinsella

Computational discovery of viruses and their hosts

Cormac M. Kinsella

ISBN: 978-94-6483-273-0

Layout and cover design: Cormac M. Kinsella

Chapter facing art: Kristel Parv Kinsella, inspired by the works of J. R. R. Tolkien

Printing: Ridderprint, the Netherlands

The research reported in this doctoral thesis received financial assistance from the European Union's Horizon 2020 research and innovation programme, under the Marie Skłodowska-Curie Actions grant agreement no. 721367 (HONOURs). Financial support for the printing of this thesis was kindly provided by the Amsterdam UMC.

Computational discovery of viruses and their hosts

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op maandag 11 september 2023, te 14.00 uur

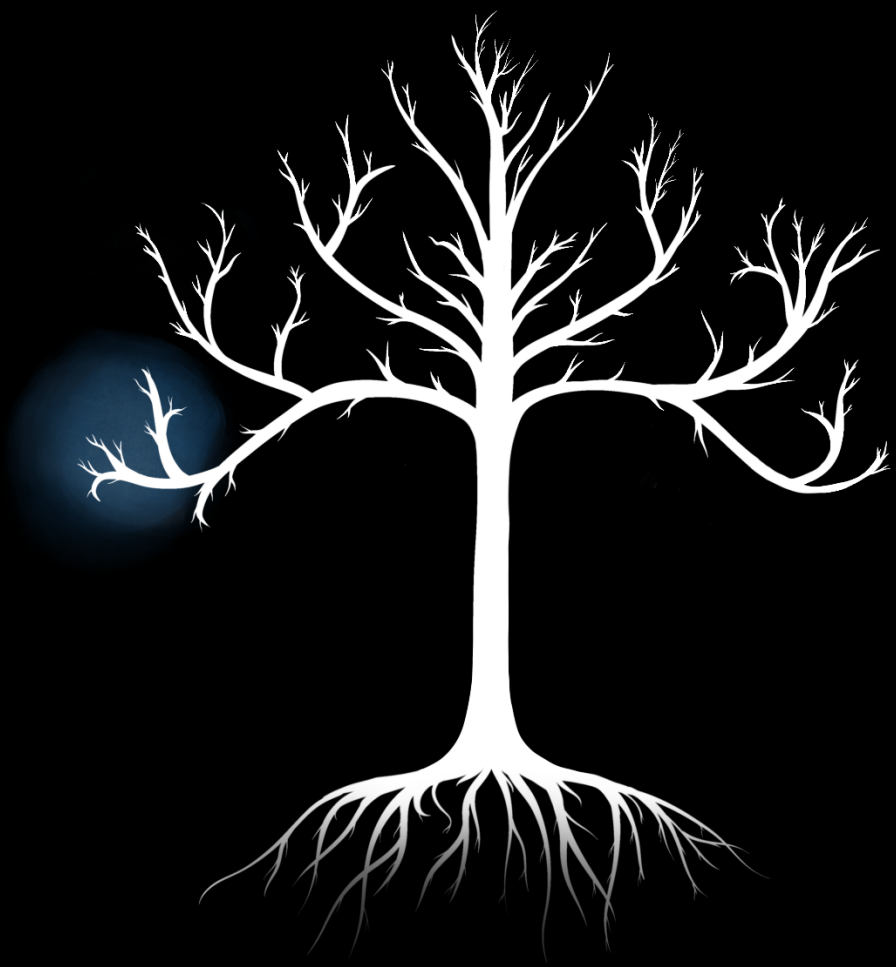
door Cormac Michael Kinsella
geboren te Harrow

Promotiecommissie

Promotor:	dr. C.M. van der Hoek	AMC-UvA
Copromotores:	prof. dr. B. Berkhout	AMC-UvA
	dr. A. Bart	Tergooi Ziekenhuis
Overige leden:	prof. dr. M.D. de Jong	AMC-UvA
	prof. dr. C.A. Russell	AMC-UvA
	prof. dr. M.P.G. Koopmans	Erasmus Universiteit Rotterdam
	dr. M. Krupovic	Institut Pasteur
	dr. J. Matthijnssens	KU Leuven

Faculteit der Geneeskunde

Table of contents



Chapter 1

General introduction and scope of this thesis

The discovery of viruses, a distinct class of disease agents

'Virus', derived from a Latin word meaning poison, has been used to non-specifically describe infectious disease agents for centuries[1]. When scientists in the 1800s came to understand that some microbes could cause disease, a flurry of cellular pathogens were isolated in pure culture by growing them on nutrient-rich matrices, allowing their associations to disease to be directly tested under experimental conditions[2]. An assumption that culturable bacteria, fungi, and protists caused all infectious diseases took root. Usage of the term 'virus' remained non-specific into the early 1900s, with apparent oxymorons such as 'bacterial viruses' appearing[3] – meaning 'bacterial agents of disease' – not 'viruses infecting bacteria' as we might now understand it. However, in 1898 a key conceptual leap was made that would shape the modern conception of viruses, namely that a category of disease agents distinct from bacteria existed. First, work by Friedrich Loeffler and Paul Frosch showed that the causative agent of foot and mouth disease could pass through filters capable of holding back all known bacterial cells[4]. They postulated a very small, particulate agent of disease that was capable of replication (i.e., not a toxin). Secondly, Dutch microbiologist Martinus Beijerinck showed that the agent causing tobacco mosaic disease could also pass filters[5]. Beijerinck proposed a non-bacterial identity for the agent, though he considered it to be liquid-like, or as he called it: "contagious living fluid". A new class of agents known as 'filterable viruses' were thus recognised, and over the following decades non-specific usage of the terminology faded, until 'filterable' was also eventually dropped.

What defines a virus?

We now understand that viruses are not liquid-like, instead they are made up of infectious particles called **virions**. The small size of most virions explains why they can pass fine filters, though size does not define them. In fact, so-called 'giant viruses' have been found that are larger than the smallest bacteria[6,7]. More fundamentally, viruses are **acellular** but require cells to replicate, as they lack some of the necessary machinery for producing further generations. They are thus **obligate intracellular parasites** of host replication machinery, and **must transmit between host cells** to gain access to this. Virions represent individual virus units, such that in some cases a single virion can produce a new infection. At the least, virions possess a **genome** or genome segment of RNA or DNA, and some **proteins** encoded by that genome. While these features define most known viruses, biological discoveries regularly complicate attempts at an all-encompassing yet restrictive definition. For example, one definition[8] splits biological entities into either ribosome-encoding or capsid-encoding forms, i.e., cellular life and viruses respectively. However, viruses that lack capsids and encode other proteins are now known[9], excluding them from this definition, and also from the viroids (virus-like elements that do not encode protein). Dropping the capsid requirement of the definition opens the door to other selfish genetic elements usually considered distinct from viruses, such as some transposons or plasmids. A clean definition is likely elusive, and given that viruses are a polyphyletic group (i.e., they did not all evolve from a single common ancestor) this should be expected. Individual

discoveries should therefore be evaluated in terms of how much their genetic relationships and biological behaviours overlap with those considered typically viral.

The development of virus discovery techniques

The visible effects of viruses have long been readily apparent to humans[10,11], likely since our origin[12]. Experimentation with viruses also began before their nature was understood, for example Edward Jenner's work on smallpox vaccination in the 1700s[13]. Virus discovery as a field arguably began with Loeffler, Frosch, and Beijerinck's conclusions regarding filterable viruses[4,5]. By 1912, application of filtration techniques resulted in the discovery of at least 17 distinct viruses[14,15], though detection and study was only possible via the diseases they induced. The subsequent development of virus discovery was tied to technological innovations enabling deeper characterisation and thus categorisation of filterable agents. Key early advances were the 1935 crystallisation of tobacco mosaic virus (TMV)[16], the 1937 discovery of viral nucleic acids[17], the 1939 electron microscope analysis of TMV[18], and the 1941 application of X-ray crystallography techniques[19]. These enabled analysis of virus biochemistry and morphology.

Viruses only replicate in host cells, so early attempts to produce pure virus cultures in nutrient media were unsuccessful. Early propagation was done in whole organisms or eggs, and this had multiple drawbacks including bacterial contamination of stocks[20]. It was during a negative experiment aiming to grow pure vaccinia virus that Frederick Twort inadvertently established the first virus culture, though it was not vaccinia. Reporting in 1915[21], Twort noticed that colonies of growing bacterial contaminants were killed off by a filterable, dilutable, infectious agent that could be propagated between colonies. Subsequent work from 1917 by Félix d'Hérelle named the 'bacteriophages' and properly established virus culture in bacterial cells, and specifically the plaque assay, as vital tools in virus research and discovery[22]. As eukaryotic tissue and cell culture techniques developed later in the 1900s, many viruses were discovered by inoculating cultures with infectious material and isolating agents[23–25]. Cell, tissue, or host tropism could also be tested using panels of different cell cultures[25], something that Twort already comprehended in 1915 when testing bacteriophage host tropism[21]. With advances in immunology, the possibility to characterise isolated viruses by their antigenic or serological properties also developed[26], and with this came the ability to test for viruses using immunoassays[25,27]. While two agents may share similar morphology and cytopathic effects, different responses to antibodies could distinguish 'serotypes'.

By the 1970s scientists already had powerful tools to find and characterise new pathogenic viruses, but a revolution in molecular biology was underway. Restriction enzymes that cut DNA in specific locations had been isolated[28], vital components of molecular cloning techniques that enabled amplification of specific nucleic acids[29]. In 1977 Frederick Sanger refined a technique for DNA sequencing and the first ever virus genome sequence was published, φX174[30,31]. This would eventually allow determination of comparative virus

relationships, but did not immediately overhaul virus discovery methods, as it required pure input DNA at high copy number, and was therefore limited to viruses established in culture or cloned fragments. In the 1980s the polymerase chain reaction (PCR) method was developed[32,33], which enabled amplification of specific DNA sequences via multiple cycles of *in vitro* reactions. Because PCR utilises 'primer' sequences that match sections of a target, it could also be used to detect closely related targets[34]. Primers designed to target sequences highly conserved across an entire viral lineage have often been used to detect unknown members of the group[35]. However, detection range is limited by design, and more divergent viruses will not be found.

To solve this, advanced molecular biology techniques agnostic to virus sequence were applied. These included shotgun cloning, wherein total DNA from a sample was randomly sheared, and fragments were then cloned and Sanger sequenced[36,37]. As this could be applied to mixed samples containing nucleic acids from multiple organisms, it became known as 'metagenomics'[37]. Representational difference analysis was another approach[38], which disproportionately amplified nucleic acids found in one sample but not another (i.e., a virus found in a test sample, but not in a control sample). Similarly, techniques such as sequence-independent single primer amplification (SISPA) and virus discovery based on cDNA-amplified fragment length polymorphism (VIDISCA) used restriction enzymes to digest nucleic acids in control and test samples before amplification, with different nucleic acid fragments then visualised by gel electrophoresis[39,40]. Samples containing a new virus displayed unique nucleic acid fragments, which were then excised from the gel, cloned, and sequenced. Inclusion of a reverse transcription step converting RNA virus genomes to DNA enabled detection of either genome type, and further laboratory techniques could non-specifically enrich virus nucleic acids relative to background. These included centrifugation of samples to remove heavier cell debris, filtration of supernatants to remove other large particles, treatment with nucleases such as DNase to digest naked host chromosomal DNA, and use of selective primers during reverse transcription to reduce host ribosomal RNA levels[39–42].

Virus discovery with high-throughput sequencing

Despite the maturation of virology during the 1900s, key issues remained at the turn of the millennium. One of these, discussed by Twort even in 1915[21], was efficient identification of viruses that do not cause visible disease or cytopathic effect, and relatedly, how to find viruses infecting host species difficult to isolate in cell culture. While molecular techniques offered promising solutions, they remained low-throughput and logistically complex[36,38–40]. It would be the development of high-throughput sequencing (HTS) platforms in the 2000s[43] that precipitated a major leap forward for virus discovery. Also known as massively parallel sequencing or next-generation sequencing, HTS techniques allow simultaneous sequencing of millions of DNA fragments in a processed sample known as a 'library'. As the fragments overlap in their sequence content, they can be computationally 'assembled' together into longer sequences[44], including whole virus genomes. Using sequence similarity detection

algorithms such as the basic local alignment search tool (BLAST)[45], novel virus genomes can be identified. Because HTS requires no prior knowledge of target sequences and no cloning, it was readily integrated with metagenomic approaches[46] (i.e., metagenomic HTS), enabling discovery of apathogenic or unculturable viruses from any environment[47]. Complicating this, sequenced genomes can remain undetected if they are highly divergent from known viruses. While fast and sensitive protein similarity detection algorithms[48–50] and even protein structure-based comparison tools[51] have pushed the limits of remote homology detection, scientists have not yet charted all virus sequence 'dark matter'.

Today, virus discovery techniques such as VIDISCA have been updated to take advantage of HTS technology (i.e., VIDISCA-NGS[42]), while further techniques have been developed[52–54]. Overall, the importance of metagenomic HTS is such that it spawned the age of 'viromic' studies, aiming to sequence all viral genomes in a particular individual, community, or environment. The vast increase in data processing requirements drove advances in computational algorithms used in sequence analysis, and together these technologies have enabled discovery of hundreds to hundreds of thousands of virus genomes even within single reports[55–57]. With virus genome discovery now far outpacing the ability to characterise individual viruses in the laboratory, the International Committee on Taxonomy of Viruses (ICTV) recently took the step of allowing assignment of virus taxonomy to sequences acquired using metagenomic HTS alone[58]. Further, moving away from traditional characterisation metrics such as phenotype, taxonomy is now recommended to centre around monophyletic evolutionary relationships, in effect prioritising genomic sequence information[59].

The host identity problem

Over most of the history of virology, the identity of host species has been self-evident, because virus discovery efforts began with a host disease. With the metagenomic HTS revolution, this 'host first' identification order is reversed for most new viruses[58,60]. Many viruses today have a known genome sequence but an unknown host, referred to in parts of this thesis as 'stray viruses'. At first glance this problem might appear simple; for example, we may conclude a novel virus discovered in the intestines of a person is a human-infecting virus. However, this is not always true. Microbe cells outnumber mammal cells in humans[61], and all of these can suffer virus infections. Many eukaryotic parasites live in mammalian guts[62], and food contains numerous viruses capable of transiting the digestive system[63]. Most environments are analogous, in that the potential host diversity is high, and links between individual viruses and their specific hosts are obscured. This is an important challenge to solve, as without host information we cannot clearly conclude the medical or veterinary importance of stray viruses, and cannot contextualise their evolution.

Laboratory approaches to solve host identities vary in their utility. Attempting to isolate a stray virus in cell culture may be suitable when a specific host is suspected[64], but is otherwise low-throughput and unlikely to succeed. Many potential host taxa have never

been isolated in culture, and no single laboratory maintains all established culture systems. More promisingly, library preparation techniques that compartmentalise samples at the level of single cells before sequencing allow capture of viruses inside specific identifiable organisms[65]. Other approaches such as proximity ligation link physically close nucleic acids[66] and can thus show which organism a virus is in. Methodologies include hybridisation of viral mRNA to host rRNA before sequencing[67], and Hi-C[64]. As these techniques are done upstream of sequencing, they do not offer a solution for stray viruses identified using conventional HTS, i.e., the majority.

For stray viruses, computational methods of host identification are currently the most appropriate. Phylogenetic analysis is often used to find the most closely related virus with a known host, as host tropism is generally a conserved feature of viruses, allowing educated predictions[60]. Viruses often coevolve with their hosts, resulting in similar evolutionary branching patterns that may hold for millions of years[68]. However, accuracy of inferences depends on the degree of host switching in the lineage, the viral host range, and the degree of relatedness to viruses with determined hosts. Furthermore, it requires prior knowledge of some host identities across the viral lineage, information which is often absent. Many other approaches utilise similar prior knowledge[69,70]. For example, machine learning approaches train algorithms by analysing many genome sequences of viruses with known hosts, and then apply this to predict hosts in unknown cases[71]. This can be effective for lineages in which many host relationships are already known[72], but it will never predict a host that does not occur in the training data. If available, host genome assemblies can partly solve these issues. Viruses occasionally leave genomic traces in host genomes, and detecting these can directly link virus lineages to hosts. In prokaryotic hosts, bacteriophage sequences are sometimes incorporated into clustered regularly interspaced short palindromic repeats (CRISPRs) for use in antiviral defence. Detecting CRISPR similarity to exogenous bacteriophages allows host inference[73]. In eukaryotic hosts that lack CRISPR, endogenous viral elements (EVEs) may offer an equivalent line of evidence. EVEs are occasionally generated upon infection of host germline cells, and can be vertically inherited as part of the genome for millions of years, allowing investigation of virus host ranges[74].

A host inference study system: the *Cressdnaviricota*

As mentioned above, the first virus sequenced was φX174, which has a circular genome of single-stranded (ss)DNA and infects a prokaryote. This genomic arrangement was previously thought extremely rare for viruses infecting eukaryotes. During the 1970s and 1980s two plant-pathogenic lineages were identified, the geminiviruses and nanoviruses[75,76]. Both were notable for their small virion sizes, between 15 and 20 nanometers in diameter. Upon genome sequencing the two lineages were found to share a homologous *Rep* gene, indicating common ancestry between them[77]. In 1974 the only lineage known to infect vertebrates was found, the circoviruses[78,79]. Considerable interest in the group was raised when a globally important disease of pigs (postweaning multisystemic wasting syndrome) was found to be circovirus-induced[80]. In 2005 and 2010 additional

lineages causing cell lysis of diatoms and debilitation of a fungus were found, the bacilladnaviruses and genomoviruses respectively[81,82]. United by a similar genome organisation and a homologous *Rep* gene encoding a protein with both an endonuclease and a helicase domain, the acronym CRESS DNA (circular Rep-encoding single-stranded DNA) virus was coined to refer to them collectively[83]. Application of rolling circle amplification to enrich circular DNAs and metagenomic analysis gradually revealed CRESS viruses were widespread and diverse[54,83–88], and numerous stray CRESS viruses have been found, including in association to disease[89–92]. At the outset of this thesis in November 2017, the five lineages mentioned above were all officially accepted families (named *Geminiviridae*, *Nanoviridae*, *Circoviridae*, *Bacilladnaviridae*, and *Genomoviridae*), and the unofficial family *Kirkoviridae* was proposed in the literature[89]. During work on this thesis, the *Smacoviridae*[93,94], *Redondoviridae*[90], and *Metaxyviridae*[95] were described by other authors and accepted as official families, while the unofficial lineages CRESSV1 to CRESSV6 were reported[96], and likely represent further family-level clusters. In recognition of this rapidly expanding diversity, the virus phylum *Cressdnaviricota* was recently established[97]. Housing many stray virus lineages – including some associated to disease – the phylum represents an appropriate study system to develop host inference techniques.

Scope of this thesis

The aims of this thesis were to develop and apply computational approaches to both the discovery of viruses and the identification of their hosts. While the *Cressdnaviricota* were a major focus of this work, the overarching goal was to address challenges common across the virus discovery field. The intention is that this thesis will contribute to understanding the evolutionary history and biology of additional virus groups, and their current roles in disease.

Previous work in our laboratory established the library-preparation method VIDISCA-NGS as a powerful tool for enrichment and discovery of viruses. We developed a novel computational workflow for analysis of VIDISCA-NGS data, reported in **chapter 2**. In addition to field-standard sequence-similarity based approaches, the workflow was designed to leverage the reproducible production of specific restriction fragments from a given DNA template. The resulting 'cluster-profiling analysis' enabled identification of virus-like sequences even in the absence of detectable sequence similarity.

Application of the resulting computational workflow led to the discovery of previously unknown cressdnaviruses in human stool, reported in **chapter 3**. Determination of their genetic relationships revealed three families, which we named *Naryaviridae*, *Nenyaviridae*, and *Vilyaviridae*, now officially recognised by the ICTV[98]. To identify their hosts, we applied case-control analyses of human stool samples, alongside analyses of host EVEs and small RNAs, and virus recombination. Hosts were identified as members of the important human parasite genera *Entamoeba* and *Giardia*.

Building upon this work, we aimed to develop a computational workflow that required no training data and was capable of virus host prediction in the absence of host genome assemblies, reported in **chapter 4**. Focusing on cressdnaviruses, we first phylogenetically characterised additional unclassified lineages, resolving lineages CRESSV7 to CRESSV39. Examining disease-associated lineages found in the gastrointestinal tracts of humans and pigs, we predicted hosts of four, namely the *Redondoviridae* with *Entamoeba gingivalis*, *Kirkoviridae* with parabasalids including *Dientamoeba*, CRESSV1 with *Blastocystis*, and CRESSV19 with *Endolimax*.

Horizontal gene transfer from viruses to hosts occasionally generates EVEs, which are useful for determination of virus host relationships. In **chapter 5**, we extended this concept to horizontal gene transfer between viruses, in a case where the host of one virus lineage was already known. We showed the cressdnavirus lineage CRESSV3 donated *Rep* genes to avipoxviruses, large dsDNA pathogens of birds and other saurians. This implied saurian hosts for CRESSV3, only the second cressdnavirus lineage after the *Circoviridae* recognised to infect vertebrates. We renamed this unofficial lineage as the family *Draupnirviridae*, and provided evidence that they first infected saurian hosts over 100 million years ago.

Some cressdnaviruses infecting fungi can induce debilitation and hypovirulence effects. In **chapter 6**, we carried out a virus discovery project on isolates of human-pathogenic fungi looking for further new species. While we did not identify cressdnaviruses infecting fungi, we did find a wide diversity of new RNA viruses in the cultures, including one from a lineage never previously confirmed as fungus-infecting.

In **chapter 7**, the results are evaluated and possibilities for future work are discussed.

References

1. Horzinek, M. C. The birth of virology. Antonie Van Leeuwenhoek 71, 15–20 (1997).

2. Blevins, S. M. & Bronze, M. S. Robert Koch and the 'golden age' of bacteriology. Int. J. Infect. Dis. 14, e744–e751 (2010).

3. Rosenau, M. J. The inefficiency of bacterial viruses in the extermination of rats. in The rat and its relation to the public health (Public Health and Marine-Hospital Service of the United States, 1910).

4. Witz, J. A reappraisal of the contribution of Friedrich Loeffler to the development of the modern concept of virus. Arch. Virol. 143, 2261–2263 (1998).

5. Beijerinck, M. W. Über ein contagium vivum fluidum als Ursache der Fleckenkrankheit der Tabaksblatter. Verh. der Koninklyke Akad. van Wettenschappen te Amsterdam 65, 3–21 (1898).

6. Legendre, M. et al. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. Proc. Natl. Acad. Sci. 111, 4274–4279 (2014).

7. La Scola, B. et al. A giant virus in amoebae. Science 299, 2033 (2003).

8. Raoult, D. & Forterre, P. Redefining viruses: Lessons from Mimivirus. Nat. Rev. Microbiol. 6, 315–319 (2008).

9. Ayllón, M. A. et al. ICTV virus taxonomy profile: Botourmiaviridae. J. Gen. Virol. 101, 454 (2020).

10. Saunders, K., Bedford, I. D., Yahara, T. & Stanley, J. The earliest recorded plant virus disease. Nature 422, 831–831 (2003).

11. Strouhal, E. Traces of a smallpox epidemic in the family of Ramesses V of the Egyptian 20th dynasty. Anthropologie 34, 315–319 (1996).

12. Enard, D., Cai, L., Gwennap, C. & Petrov, D. A. Viruses are a dominant driver of protein adaptation in mammals. Elife 5, e12469 (2016).

13. Jenner, E. An inquiry into the causes and effects of the variolæ vaccinæ, a disease discovered in some of the western counties of England, particularly Gloucestershire, and known by the name of the cow pox. (Sampson Low, 1798).

14. Flexner, S. Some problems in infection and its control. Science 36, 685–702 (1912).

15. Wolbach, S. B. The filterable viruses, a summary. Bost. Med. Surg. J. 167, 419–427 (1912).

16. Stanley, W. M. Isolation of a crystalline protein possessing the properties of tobacco-mosaic virus. Science 81, 644–645 (1935).

17. Bawden, F. C. & Pirie, N. W. The isolation and some properties of liquid crystalline substances from solanaceous plants infected with three strains of tobacco mosaic virus. Proc. R. Soc. London. Ser. B - Biol. Sci. 123, 274–320 (1937).

18. Kausche, G. A., Pfankuch, E. & Ruska, H. Die sichtbarmachung von pflanzlichem virus im übermikroskop. Naturwissenschaften 27, 292–299 (1939).

19. Bernal, J. D. & Fankuchen, I. X-ray and crystallographic studies of plant virus preparations. J. Gen. Physiol. 25, 111–165 (1941).

20. Noguchi, H. Pure cultivation in vivo of vaccine virus free from bacteria. J. Exp. Med. 21, 539–570 (1915).

21. Twort, F. W. An investigation on the nature of ultra-microscopic viruses. Lancet 186, 1241–1243 (1915).

22. D'Hérelle, F. Bacteriophage as a treatment in acute medical and surgical infections. Bull. N. Y. Acad. Med. 7, 329–348 (1931).

23. Hematian, A. et al. Traditional and modern cell culture in virus diagnosis. Osong Public Heal. Res. Perspect. 7, 77–82 (2016).

24. Enders, J. F., Weller, T. H. & Robbins, F. C. Cultivation of the lansing strain of poliomyelitis virus in cultures of various human embryonic tissues. Science 109, 85–87 (1949).

25. Hsiung, G. D. Diagnostic virology: From animals to automation. Yale J. Biol. Med. 57, 727–733 (1984).

26. Rowe, W. P., Huebner, R. J., Hartley, J. W., Ward, T. G. & Parrott, R. H. Studies of the adenoidal-pharyngeal-conjunctival (APC) group of viruses. Am. J. Epidemiol. 61, 197–218 (1955).

27. Mir, M. A., Mehraj, U., Nisar, S. & Qayoom, H. Quantitation of specific antibodies by enzyme-labeled anti-immunoglobulin in antigen-coated tubes. J. Immunol. 109, 129–135 (1972).

28. Linn, S. & Arber, W. Host specificity of DNA produced by Escherichia coli, X. In vitro restriction of phage fd replicative form. Proc. Natl. Acad. Sci. 59, 1300–1306 (1968).

29. Nathans, D. & Smith, H. O. Restriction endonucleases in the analysis and restructuring of DNA molecules. Annu. Rev. Biochem. 44, 273–293 (1975).

30. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. 74, 5463–5467 (1977).

31. Sanger, F. et al. Nucleotide sequence of bacteriophage φX174 DNA. Nature 265, 687–695 (1977).

32. Saiki, R. K. et al. Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230, 1350–1354 (1985).

33. Mullis, K. B. & Faloona, F. A. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol. 155, 335–350 (1987).

34. Lane, D. J. et al. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc. Natl. Acad. Sci. 82, 6955–6959 (1985).

35. Zhu, N. et al. A novel coronavirus from patients with pneumonia in China, 2019. N. Engl. J. Med. 382, 727–733 (2020).

36. Breitbart, M. et al. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. 99, 14250–14255 (2002).

37. Rondon, M. R. et al. Cloning the soil metagenome: A strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl. Environ. Microbiol. 66, 2541–2547 (2000).

38. Nishizawa, T. et al. A novel DNA virus (TTV) associated with elevated transaminase levels in posttransfusion hepatitis of unknown etiology. Biochem. Biophys. Res. Commun. 241, 92–97 (1997).

39. Hoek, L. van der et al. Identification of a new human coronavirus. Nat. Med. 10, 368 (2004).

40. Allander, T., Emerson, S. U., Engle, R. E., Purcell, R. H. & Bukh, J. A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. Proc. Natl. Acad. Sci. 98, 11609–11614 (2001).

41. Endoh, D. et al. Species-independent detection of RNA virus by representational difference analysis using non-ribosomal hexanucleotides for reverse transcription. Nucleic Acids Res. 33, e65 (2005).

42. de Vries, M. et al. A sensitive assay for virus discovery in respiratory clinical samples. PLoS One 6, e16118 (2011).

43. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380 (2005).

44. Myers, E. W. et al. A whole-genome assembly of Drosophila. Science 287, 2196–2204 (2000).

45. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. J. Mol. Biol. 215, 403–410 (1990).

46. Edwards, R. A. et al. Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics 7, 1–13 (2006).

47. Angly, F. E. et al. The marine viromes of four oceanic regions. PLOS Biol. 4, e368 (2006).

48. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461 (2010).

49. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat. Methods 18, 366–368 (2021).

50. Karplus, K., Barrett, C. & Hughey, R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 14, 846–856 (1998).

51. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 33, W244–W248 (2005).

52. Wylezich, C., Papa, A., Beer, M. & Höper, D. A versatile sample processing workflow for metagenomic pathogen detection. Sci. Rep. 8, 13108 (2018).

53. Conceição-Neto, N. et al. Modular approach to customise sample preparation procedures for viral metagenomics: A reproducible protocol for virome analysis. Sci. Rep. 5, 16532 (2015).

54. Tisza, M. J. et al. Discovery of several thousand highly diverse circular DNA viruses. Elife 9, e51971 (2020).

55. Shi, M. et al. Redefining the invertebrate RNA virosphere. Nature 540, 539–543 (2016).

56. Tisza, M. J. & Buck, C. B. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. Proc. Natl. Acad. Sci. 118, e2023202118 (2021).

57. Edgar, R. C. et al. Petabase-scale sequence alignment catalyses viral discovery. Nature 602, 142–147 (2022).

58. Simmonds, P. et al. Virus taxonomy in the age of metagenomics. Nat. Rev. Microbiol. 15, 161–168 (2017).

59. Simmonds, P. et al. Four principles to establish a universal virus taxonomy. PLOS Biol. 21, e3001922 (2023).

60. Wolf, Y. I. et al. Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. Nat. Microbiol. 5, 1262–1270 (2020).

61. Sleator, R. D. The human superorganism – of microbes and men. Med. Hypotheses 74, 214–215 (2010).

62. Patterson, Q. M. et al. Circoviruses and cycloviruses identified in Weddell seal fecal samples from McMurdo Sound, Antarctica. Infect. Genet. Evol. 95, 105070 (2021).

63. Victoria, J. G. et al. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. J. Virol. 83, 4642–4651 (2009).

64. Keeler, E. L. et al. Widespread, human-associated redondoviruses infect the commensal protozoan Entamoeba gingivalis. Cell Host Microbe 31, 58-68.e5 (2023).

65. Yoon, H. S. et al. Single-cell genomics reveals organismal interactions in uncultivated marine protists. Science 332, 714–717 (2011).

66. Marbouty, M., Baudry, L., Cournac, A. & Koszul, R. Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay. Sci. Adv. 3, e1602105 (2017).

67. Ignacio-Espinoza, J. C. et al. Ribosome-linked mRNA-rRNA chimeras reveal active novel virus host associations. bioRxiv (2020).

68. Aiewsakun, P. & Katzourakis, A. Marine origin of retroviruses in the early Palaeozoic Era. Nat. Commun. 8, 1–12 (2017).

69. Kapoor, A., Simmonds, P., Lipkin, W. I., Zaidi, S. & Delwart, E. Use of nucleotide composition analysis to infer hosts for three novel picorna-like viruses. J. Virol. 84, 10322–10328 (2010).

70. Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A. & Sun, F. Alignment-free d2* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Res. 45, 39–53 (2017).

71. Mock, F., Viehweger, A., Barth, E. & Marz, M. VIDHOP, viral host prediction with deep learning. Bioinformatics 37, 318–325 (2021).

72. Eng, C. L. P., Tong, J. C. & Tan, T. W. Predicting host tropism of influenza A virus proteins using random forest. BMC Med. Genomics 7, S1 (2014).

73. Dion, M. B. et al. Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. Nucleic Acids Res. 49, 3127–3138 (2021).

74. Katzourakis, A. & Gifford, R. J. Endogenous viral elements in animal genomes. PLoS Genet. 6, e1001191 (2010).

75. Harrison, B. D. et al. Plant viruses with circular single-stranded DNA. Nature 270, 760–762 (1977).

76. Chu, P. W. G. & Helms, K. Novel virus-like particles containing circular single-stranded DNAs associated with subterranean clover stunt disease. Virology 167, 38–49 (1988).

77. Boevink, P., Chu, P. W. G. & Keese, P. Sequence of subterranean clover stunt virus DNA: Affinities with the geminiviruses. Virology 207, 354–361 (1995).

78. Ritchie, B. W., Niagro, F. D., Lukert, P. D., Steffens, W. L. & Latimer, K. S. Characterization of a new virus from cockatoos with psittacine beak and feather disease. Virology 171, 83–88 (1989).

79. Tischer, I., Rasch, R. & Tochtermann, G. Characterization of papovavirus and picornavirus-like particles in permanent pig kidney cell lines. Zentbl. Bakt. 226, 153–167 (1974).

80. Ellis, J. et al. Isolation of circovirus from lesions of pigs with postweaning multisystemic wasting syndrome. Can. Vet. J. 39, 44–51 (1998).

81. Nagasaki, K. et al. Previously unknown virus infects marine diatom. Appl. Environ. Microbiol. 71, 3528–3535 (2005).

82. Yu, X. et al. A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. Proc. Natl. Acad. Sci. 107, 8387–8392 (2010).

83. Rosario, K. et al. Diverse circular ssDNA viruses discovered in dragonflies (Odonata: Epiprocta). J. Gen. Virol. 93, 2668–2681 (2012).

84. Rosario, K. & Breitbart, M. Exploring the viral world through metagenomics. Curr. Opin. Virol. 1, 289–297 (2011).

85. Rosario, K., Duffy, S. & Breitbart, M. Diverse circovirus-like genome architectures revealed by environmental metagenomics. J. Gen. Virol. 90, 2418–2424 (2009).

86. Siqueira, J. D. et al. Complex virome in feces from Amerindian children in isolated Amazonian villages. Nat. Commun. 9, 4270 (2018).

87. Blinkova, O. et al. Novel circular DNA viruses in stool samples of wild-living chimpanzees. J. Gen. Virol. 91, 74–86 (2010).

88. Breitbart, M. & Rohwer, F. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. Biotechniques 39, 729–736 (2005).

89. Li, L. et al. Exploring the virome of diseased horses. J. Gen. Virol. 96, 2721–2733 (2015).

90. Abbas, A. A. et al. Redondoviridae, a family of small, circular DNA viruses of the human oro-respiratory tract that are associated with periodontitis and critical illness. Cell Host Microbe 25, 719–729 (2019).

91. Phan, T. G. et al. The fecal virome of South and Central American children with diarrhea includes small circular DNA viral genomes of unknown origin. Arch. Virol. 161, 959–966 (2016).

92. Zhao, G. et al. Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. Proc. Natl. Acad. Sci. 114, E6166–E6175 (2017).

93. Varsani, A. & Krupovic, M. Smacoviridae: a new family of animal-associated single-stranded DNA viruses. Arch. Virol. 163, 2005–2015 (2018).

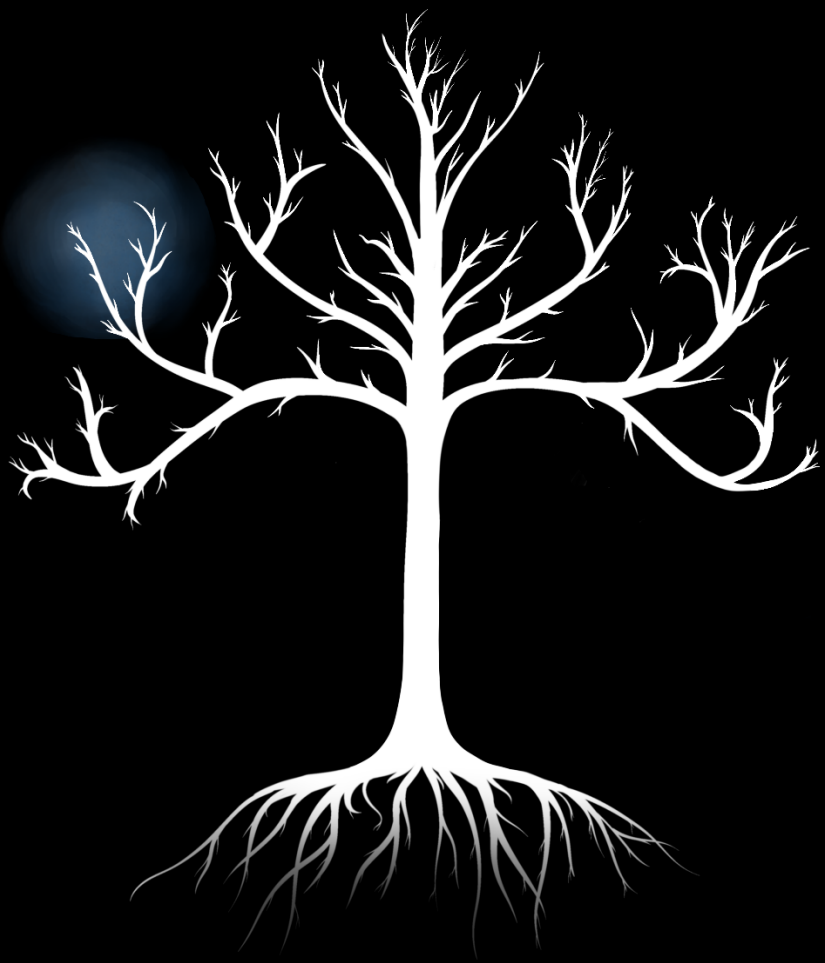
94. Ng, T. F. F. et al. A diverse group of small circular ssDNA viral genomes in human and non-human primate stools. Virus Evol. 1, vev017 (2015).

95. Gronenborn, B., Randles, J., HJ, V. & Thomas, J. Create one new family (Metaxyviridae) with one new genus (Cofodevirus) and one species (Coconut foliar decay virus) moved from the family Nanoviridae (Mulpavirales). Int. Comm. Taxon. Viruses Propos. number 2020.022P (2021).

96. Kazlauskas, D., Varsani, A. & Krupovic, M. Pervasive chimerism in the replication-associated proteins of uncultured single-stranded DNA viruses. Viruses 10, v10040187 (2018).

97. Krupovic, M. et al. Cressdnaviricota: A virus phylum unifying seven families of Rep-encoding viruses with single-stranded, circular DNA genomes. J. Virol. 94, e00582-20 (2020).

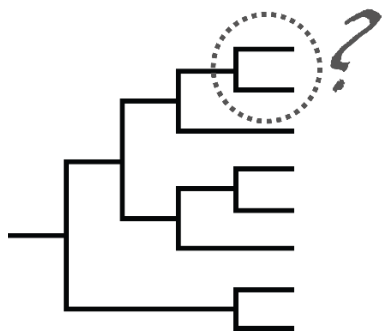
98. Krupovic, M. & Varsani, A. Naryaviridae, Nenyaviridae, and Vilyaviridae: Three new families of single-stranded DNA viruses in the phylum Cressdnaviricota. Arch. Virol. 167, 2907–2921 (2022).



Chapter 2

Enhanced bioinformatic profiling of VIDISCA libraries for virus detection and discovery

Cormac M. Kinsella, Martin Deijs, Lia van der Hoek



Virus Research, 2019

https://doi.org/10.1016/j.virusres.2018.12.010

Abstract

VIDISCA is a next-generation sequencing (NGS) library preparation method designed to enrich viral nucleic acids from samples before highly-multiplexed low depth sequencing. Reliable detection of known viruses and discovery of novel divergent viruses from NGS data require dedicated analysis tools that are both sensitive and accurate. Existing software was utilised to design a new bioinformatic workflow for high-throughput detection and discovery of viruses from VIDISCA data. The workflow leverages the VIDISCA library preparation molecular biology, specifically the use of Mse1 restriction enzyme which produces biological replicate library inserts from identical genomes. The workflow performs total metagenomic analysis for classification of non-viral sequence including parasites and host, and separately carries out virus specific analyses. Ribosomal RNA sequence is removed to increase downstream analysis speed and remaining reads are clustered at 100% identity. Known and novel viruses are sensitively detected via alignment to a virus-only protein database, and false positives are removed. A new cluster-profiling analysis takes advantage of the viral biological replicates produced by Mse1 digestion, using read clustering to flag the presence of short genomes at very high copy number. Importantly, this analysis ensures that highly repeated sequences are identified even if no homology is detected, as is shown here with the detection of a novel gokushovirus genome from human faecal matter. The workflow was validated using read data derived from serum and faeces samples taken from HIV-1 positive adults, and serum samples from pigs that were infected with atypical porcine pestivirus.

Highlights

- A sensitive bioinformatic workflow for virus detection in VIDISCA data.
- Flagging of possible novel viruses in unclassified reads using clustering.
- Cluster-profiling analysis for reproducible sample comparison.
- Multiple analysis approaches provide extra utility to the user.

Introduction

The host range expansion of viral pathogens and emergence of novel species can pose substantial threats to human health (Parrish et al., 2008). Viruses evolve rapidly, possess high molecular diversity, and are found in relatively low concentration alongside host nucleic acids in most sample types. These factors complicate detection of novel viral genetic material and necessitate specific virus discovery methods to achieve sufficient detection sensitivity. Next-generation sequencing (NGS) and metagenomics have greatly accelerated the discovery of novel viruses when contrasted with traditional wet-lab virological techniques such as isolation in cell culture, as they can be performed on any

virus directly from biological or environmental samples, in a high-throughput way (Shi et al., 2018, 2016). Approaches that prioritise an unbiased metagenomic profile require high sequencing depth to ensure pathogen detection, and are therefore relatively expensive per viral nucleotide. The incorporation of virus enrichment techniques prior to sequencing reduces the required depth for detection (Conceição-Neto et al., 2015; de Vries et al., 2011), and may be desirable when processing tens to hundreds of samples.

VIDISCA is a virus discovery NGS library preparation method that enriches viral nucleic acids in samples before low depth Ion Torrent sequencing, allowing processing of 140 samples per week. The wet-lab procedure, described in detail elsewhere (de Vries et al., 2011; Edridge et al., 2018), is summarised here in order to highlight advantages for bioinformatic analysis. First, cells and debris are pelleted, and virus-containing supernatant is DNase treated to reduce residual cellular DNA. Virion proteins are linearised to release nucleic acid, which is extracted using the Boom method (Boom et al., 1990). RNA viruses are reverse transcribed using non-ribosomal RNA (rRNA) hexamer primers (Endoh et al., 2005), which reduce the proportion of rRNA transcribed into DNA. After second-strand synthesis, double-stranded DNA products are digested using the frequent cutting Mse1 restriction enzyme, an important feature unique to VIDISCA library preparation. Sequencing primers are ligated onto the two sticky ends of a restriction fragment, before size selection against both long and short fragments, amplification with PCR, and sequencing with the Ion Torrent PGM platform (Thermo Fisher Scientific, Waltham, MA, USA).

The inclusion of Mse1 digestion during library preparation has advantageous implications for virus discovery bioinformatics. Viral genomes are short compared to their host, and can be at high copy number during infection. Since Mse1 reproducibly cuts homologous restriction fragments from genomes of the same type, high numbers of viral biological replicates with identical start and end sites are expected in library inserts prior to PCR. This is in contrast with a randomly fragmented library in which identical start and end sites are relatively rare. The VIDISCA insert redundancy is not expected from background or host nucleic acid, except that with ‘virus-like’ characteristics, i.e. high copy number, such as mitochondrial DNA. The virus replicates should result in characteristic redundancy in sequencing data, which can be identified via read clustering. Additionally, since Mse1 cuts TTAA sites, it cuts more rarely in GC rich rRNA (de Vries et al., 2011). Viable rRNA VIDISCA fragments are generally longer as a result, and can be disproportionately reduced during size selection, contributing to a high sensitivity that enables lower sequencing depth and analysis time. Recently VIDISCA was used to discover the suspected human pathogen Ntwetwe virus with 2 reads from 6,947, whereas an in-house Illumina workflow optimised for virus detection found only 8 reads among the 2,741,915 obtained (Edridge et al., 2018).

Here we present a new bioinformatic workflow designed to process VIDISCA data. The core task is sensitive virus detection including false positive reduction. The workflow includes metagenomic analysis for identification of host background and non-viral

organisms including parasites, and collects descriptive metrics in order to flag unusual properties of samples, such as high rRNA content. It outputs text and interactive HTML results for detailed investigation of samples, and includes a new cluster-profiling analysis used to flag the presence of sequences at high copy number (e.g. virus infections). This analysis also provides an informative profile of sample content in different classification bins, including known and novel viruses, mitochondrial DNA, and background sequence. Notably, the flagging of highly repetitive reads does not rely on identity searches, ensuring that abundant unknown sequences can be identified. The utility of the workflow is presented with examples.

Materials and methods

2.1. Bioinformatic workflow for VIDISCA next-generation sequencing data

The new bioinformatic workflow for VIDISCA NGS data is summarised graphically (Fig. 1) and described in detail below. As input, the workflow takes FASTA formatted sequences. Eukaryotic and prokaryotic virus protein databases used by the workflow were constructed in advance from respective NCBI Identical Protein Groups datasets, followed by clustering at 95% identity using CD-HIT v4.7 (Fu et al., 2012). First, metagenomic analysis of raw reads is carried out using Centrifuge v1.0.3 (Kim et al., 2016) against the pre-built NCBI non-redundant nucleotide Centrifuge index including known viruses, eukaryotes, and prokaryotes (February 2018). Centrifuge classification tables are visualised as interactive HTML charts using Recentrifuge (Martí, 2018).

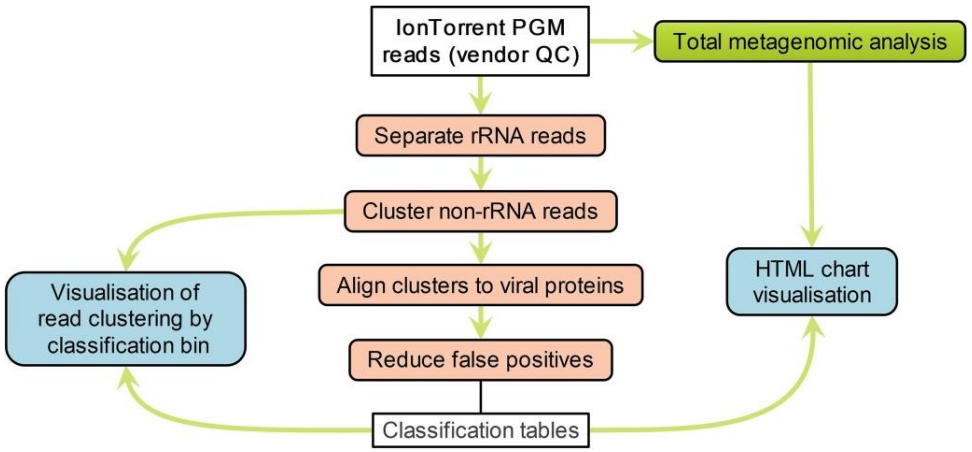


Fig. 1. Schematic overview of the bioinformatic workflow for VIDISCA data, showing the main virus detection and discovery steps (orange), the metagenomic analysis (green), and visualisation processes (blue).

Next, the main virus detection steps are run. Reads from rRNA are separated from raw reads using SortMeRNA v2.1 (Kopylova et al., 2012). Non-rRNA reads are sorted by length and clustered at 100% identity using CD-HIT v4.7, and 'clstr' files are retained for later processing. Clustered non-rRNA reads are queried against the eukaryotic virus protein database using the UBLAST algorithm provided as part of the USEARCH v10 software package, with -mincodons set to 15, -accel to 0.8, and -evalue to 1e-4 (Edgar, 2010). Unmatched reads from this step are queried against the prokaryotic virus protein database, and those remaining unclassified are mapped to human, pig, and chicken mitochondrial DNA sequences using the BWA-MEM algorithm of BWA v0.7.17 (Li, 2013). Reads matching the eukaryotic virus protein database are treated as putatively viral, and are next queried against the NCBI nt. database (April 2018) using BLASTn v2.4.0 (Camacho et al., 2009). Those classified by BLASTn as viral are regarded as confident viral reads (classified as viral twice), those classified as non-viral are regarded as false positives, and those that remain unclassified are regarded as possible unknown viruses (classified as viral once). This information is used to split the UBLAST protein classification tables into the three categories, each of which are visualised separately as interactive HTML charts using KronaTools v2.7 (Ondov et al., 2011). The BLASTn classification of false positives is also visualised for inspection and comparison to the original viral classification.

Cluster-profiling outputs are produced using the CD-HIT 'clstr' files, which are converted into a table reporting the representative sequences, the number of reads clustered per representative, and the proportion of the original non-rRNA that each represents in a sample. The classification bin (such as 'confident virus', or 'mitochondrial DNA') of each representative read is then added to the table, including a bin for unclassified sequences. This output is plotted as a bar chart using ggplot2, with separate bars for classification bins, and representative reads stacked according to proportional amount of clustering (Wickham, 2016). The classification bins are 'Virus (aa + nt)' including reads classified as viral twice, 'Virus (aa)' including reads classified as viral once, 'False pos. (nt)' including reads removed as probable false positives, 'Phage (aa)' including reads aligning to our prokaryotic virus database, 'MitoDNA' including reads mapped to mitochondrial DNA references, 'Centrifuge' including reads identified by the metagenomic tool Centrifuge, and 'No hit' including reads with no assigned classification. The bar chart output provides a visual overview of the proportion of reads from a sample that were classified in a particular bin. Furthermore, reads that represent many other reads are visually identifiable due to their higher relative proportion. This allows the presence of clustering to be identified in each bin separately. Most repetitive non-viral sequences are accounted for via removal of rRNA and binning of mitochondrial DNA, however unclassified sequences putatively from viruses require manual inspection or full-length sequencing in order to establish their likely provenance.

For each classification bin, the 10 representative sequences accounting for the largest proportion of reads are automatically extracted as FASTA files for inspection, for example with BLASTx. All text tables and sample-specific files produced by the analysis are

packaged into sample folders, and descriptive metrics about the run time and classification performance for each sample are reported to a log file for later examination.

2.2. Data selection and workflow testing

Three VIDISCA datasets were selected and analysed using the new bioinformatic workflow, in order to assess specific aspects of workflow performance and utility. First, VIDISCA reads from 194 serum samples collected in 1994–1995 from HIV-1 infected adults were run. The aim was to determine whether the bioinformatic workflow outputs could be used to troubleshoot the likely causes of pathogen detection failure. This was done by comparison of HIV-1 detection by VIDISCA with pre-existing HIV-1 load data obtained using nucleic acid sequence based amplification (NASBA). Outputs from samples in which HIV-1 was unexpectedly not detected were manually inspected to determine the cause of failure.

Second, VIDISCA reads from 194 faecal samples from the above mentioned cohort were run (Oude Munnink et al., 2014). The aim was to test the prediction that cluster-profiling could be used to flag virus-like characteristics in unclassified reads, and therefore identify novel viruses at high load missed by classification algorithms. Cluster-profiling outputs were examined for evidence of clustering among unclassified reads and a single sample (F115) was selected for follow up. Illumina reads from a randomly fragmented library of the sample were downloaded from the European Nucleotide Archive (accession ERR233419), cleaned of adapters, quality trimmed (minimum 50bp, sliding window trim < Q20) with Trimmomatic v0.38 (Bolger et al., 2014), and assembled using SPAdes v3.12 (Bankevich et al., 2012). The 10 unclassified VIDISCA representative sequences accounting for the most clustering were BLAST queried against the contigs, and the most common target sequence was extracted and manually curated.

Third, VIDISCA reads from 13 serum samples taken from sows experimentally infected with atypical porcine pestivirus (APPV) and 16 serum samples taken from the transplacentally-infected piglets of the sows were run (de Groof et al., 2016). In this case, sequencing was carried out on an Ion Proton instrument (Thermo Fisher Scientific, Waltham, MA, USA). The aims were to statistically test support for the assumption that a higher viral load would result in higher clustering among viral reads, and to explore whether such an association was strongly influenced by PCR bias toward abundant templates. Since the dataset included individuals infected with the same virus strain at a large range of viral loads, this was carried out as a reliability test of the main assumption underlying cluster-profiling analysis, that VIDISCA library preparation selects for biological replicates from identical genomes, resulting in read clustering associated with the biological load of a sequence.

3. Results and discussion

3.1. Bioinformatic workflow design

The new VIDISCA bioinformatic workflow has been designed to prioritise sensitivity to viruses, however non-virus metagenomics and the efficiency of analysis have also been considered. *K*-mer based metagenomic tools such as Kraken (Wood and Salzberg, 2014) are commonly used for pathogen detection, since they provide very rapid classification of reads via exact matches of length *k* between reads and reference indexes. Metagenomic samples often contain species with variable nucleotide identity to their most related reference sequence. Since *k* must be set in advance, high *k* decreases classification sensitivity for distantly related species, and low *k* decreases precision to well represented taxa. To circumvent this, the metagenomic software tool Centrifuge was selected for the workflow since it uses FM-indexed reference sequences, allowing *k* to be optimal for each individual read in a metagenomic sample, maximising both sensitivity and precision while simultaneously minimising index size and memory requirements (Kim et al., 2016).

Detection of novel viruses is normally achieved via local alignment of reads to viral proteins, a computationally intensive operation. High speed algorithms are available to decrease analysis time, for example UBLAST (Edgar, 2010), DIAMOND (Buchfink et al., 2015), or Kaiju (Menzel et al., 2016). Minimisation of query reads and database size can provide additional gains. The VIDISCA workflow incorporates several of these speed-ups, including rRNA removal to reduce query reads, and redundancy removal in non-rRNA using clustering. Clustering information is retained for retrospective classification of redundant reads and cluster-profiling analysis. These steps reduced average protein query counts by 31% and 45% in the 194 faecal and 194 serum datasets respectively. A virus-only protein database was constructed and clustered for a size reduction of 81%. Alignment of reads to a taxonomically restricted database raises the likelihood of spurious hits due to chance similarity, therefore false positive removal via BLAST analysis against the NCBI nucleotide database is required. Due to the prior selection steps mentioned above, a minority of reads require this querying, for example an average of 1.5% and 2.4% of reads from the above faecal and serum datasets were queried.

3.2. Assessment of the bioinformatic workflow performance

The VIDISCA bioinformatic workflow was used to identify the causes of HIV-1 detection failure in data generated from archival serum samples collected from HIV-1 positive adults. Bioinformatic analysis detected the pathogen in 128 of 194 samples (66%) with an average of 42,124 total reads per sample. Of the VIDISCA negative samples, 23 (35%) had undetectable HIV-1 loads when specifically tested with NASBA, while 9 (7%) VIDISCA positive samples did. There was a median value of 84 HIV-1 copies/µl in VIDISCA positive samples and 14 in negative (Fig. 2A), suggesting detection failure was mostly attributable to viral load. Viral load was positively associated with the proportion of HIV-1 reads (Spearman's rho = 0.61, $p < .001$), however the variance was poorly described by a

linear regression model (Fig. 2B), showing that sample dependent factors crucially impact the metagenomic profile. Notably, rRNA proportion was weakly but positively associated with HIV-1 proportion (Spearman's rho = 0.34, $p < .001$), while the proportion of non-rRNA identified as human (including residual genomic DNA and cellular RNA) was found to have a weak negative association with the HIV-1 proportion (Spearman's rho = -0.17, $p = .017$). Together these observations imply sample-specific biases against integrity or representation of the RNA fraction. Contributing factors could include higher degradation susceptibility during freeze-thaw cycles, high host DNA content with only partial degradation during DNase treatment, high intrinsic RNase activity in certain samples, or sample-specific inhibition of reverse transcription. An additional explanation could be that rRNA acts as a carrier for low concentrations of viral RNA.

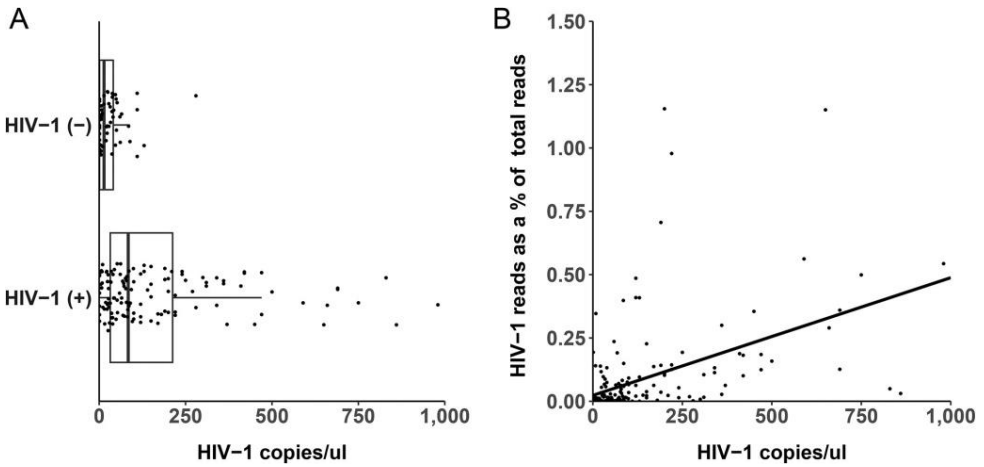


Fig. 2. A: HIV-1 viral RNA load in serum and VIDISCA outcome. HIV-1 detection in sequence reads is indicated with HIV-1 (+), and lack of detection is indication with HIV-1 (-). On the x-axis the HIV-1 RNA load per µl of serum is plotted. B: Linear regression model fitted to HIV-1 viral load against HIV-1 reads as a percentage of total reads, $F(1,192) = 56.68$, $p < .001$, $R^2 = 0.228$. A low 23% of variance in proportion is explained by viral load when assuming a linear relationship.

HIV-1 was not detected in 11 outlier samples with over 50 HIV-1 copies/µl and an average read count of 40,290. In 3 of these, cluster-profiling showed that 78–90% of processed (non-rRNA) reads belonged to Hepatitis B virus, which commonly dominates VIDISCA metagenomic profiles if present. One sample also showed possible competition with Torque Teno virus which represented 30% of processed reads. A further 6 samples had approximately 80–95% of processed reads classified by Centrifuge as host or bacterial sequence with very low read clustering, suggesting a highly diverse library insert distribution probably derived from cell lysis. In the final sample an unusually high 75% of processed reads were not classified by any analysis. Manual BLAST analysis on some of

these unclassified reads gave bacterial hits or weak alignment scores suspected to originate from unknown bacteriophages, suggesting bacterial growth in the stored material.

3.3. Cluster-profiling for virus discovery

A cluster-profiling analysis was incorporated in the workflow based on the prediction that short viral genomes at high load would result in distinctive read clustering characteristics, since VIDISCA library preparation produces homologous library inserts from each genome based on its Mse1 restriction sites. The analysis uses read clustering and classification information generated as part of the workflow to generate a visual output, and therefore does not require significant additional computational time. Importantly, the clustering signal generated by high copy number sequences does not require identity-based classification. This could potentially allow detection of highly divergent viruses with low protein identity to relatives represented in databases.

Cluster-profiling images generated using VIDISCA data from 194 faecal samples were analysed and sample F115 was selected for follow-up due to a high degree of clustering among unclassified reads – 12% of the 16,160 processed reads were clustered into only 100 unclassified representative sequences (Fig. 3), suggesting an unknown entity at high copy number. Available Illumina data from a randomly fragmented library of this sample were assembled into 9157 contigs. Ten unclassified representative VIDISCA sequences accounting for the most reads, which were automatically extracted by the workflow, were aligned to the contigs using BLAST. Of the 10, 8 aligned to a single contig, suggesting that they were part of a genome of a novel virus present at high copy number. Manual curation of this 5 kb sequence showed that it is a novel gokushovirus (circular ssDNA bacteriophage, NCBI accession number MK263179) with 72% nucleotide identity to its closest relative. The sequences of this virus were not identified by the classification components of the workflow since the related viral proteins were not part of the reference set. Mapping of complete read-sets revealed that 6.83% of Illumina read-pairs from the sample were derived from the virus and 17.27% of VIDISCA reads were. The result confirms the expectation that viruses at high load produce characteristic clusters in VIDISCA data, ensuring that those missed by identity searches can still be detected.

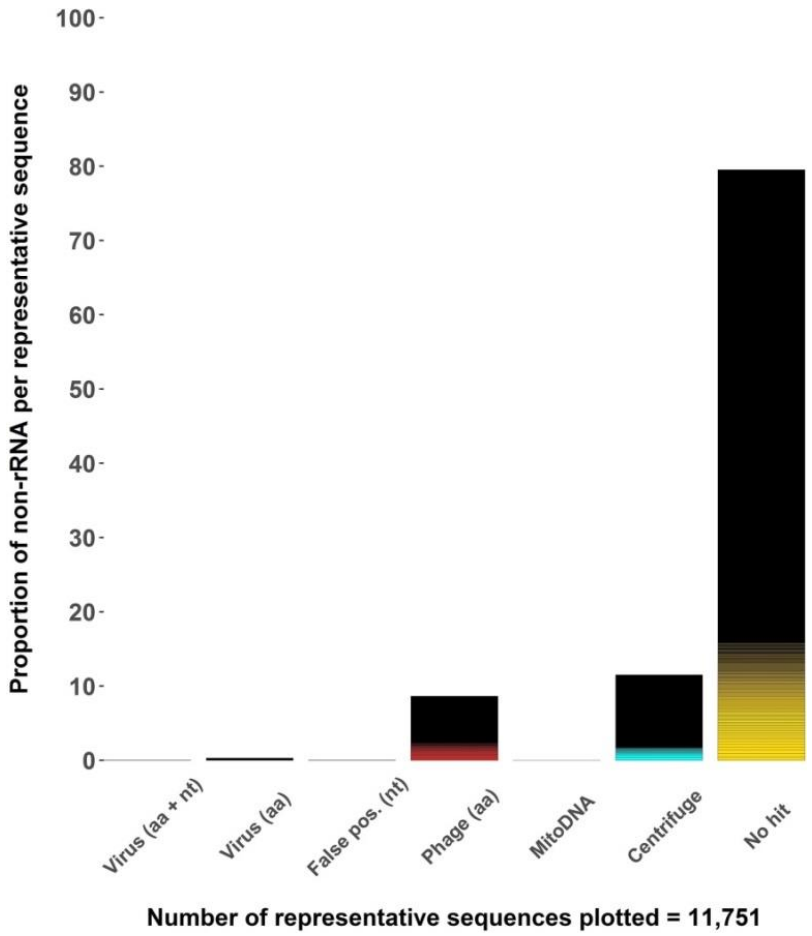


Fig. 3. Cluster-profiling bar chart from sample F115. Representative sequences produced by read clustering are plotted according to their final classification bin (x-axis) and stacked in order of their relative abundance with respect to the original non-rRNA read set (i.e. the proportion of identical reads, y-axis). Coloured bars therefore signify those sequences representing many identical reads, while many singleton reads make up black regions. Classification bins on the x-axis are those described in section 2.1. Read clustering can be seen in the phage ('Phage', red), metagenomically identified ('Centrifuge', blue), and unclassified ('No hit', yellow) read bins.

3.4. Association between viral read clustering and viral load

Cluster-profiling analysis for discovery of viruses, as shown in Fig. 3, relies on a high level of sequence redundancy in order to generate a visible signal that can be investigated. A strong association between viral load and the level of clustering observed in viral reads is expected, an effect that would underlie application of the analysis to the discovery of novel

viruses. To test this assumption VIDISCA reads from 29 serum samples taken from pigs infected with APPV were analysed. The workflow detected APPV reads in 27 of these, and a strong linear association between viral load and the proportion of APPV reads was observed after removal of a single outlier (linear regression, $F(1,26) = 70.57$, $p < .001$, $R^2 = 0.73$). As expected, there was a strong association between viral load and the average number of reads clustered per APPV representative sequence (Spearman's rho = 0.81, $p < .001$). To account for the possibility that this effect was due to stochastic PCR bias disproportionately amplifying abundant templates (Kebschull and Zador, 2015), an association between viral load and the proportion of all APPV reads that were represented by the top APPV sequence cluster was tested for. Since viral load should correspond to the abundance of replicate templates prior to PCR, PCR bias would be expected to occur in samples with the highest loads. No such relationship existed (Spearman's rho = 0.17, $p = 0.41$).

Together the observations show that the degree of clustering among viral reads corresponds well with true biological load, and does not suffer from significant PCR bias toward abundant templates. While the analysis therefore can be applied to detection of novel viruses in unclassified reads, it is important to note that only infections with a high load and a high proportional amount of reads are likely to be observed. For example, it is unlikely that the analysis would have successfully flagged the presence of HIV-1 reads in the human serum samples analysed above, had they not been successfully classified using alignment tools. Nonetheless, it does provide an additional approach to both virus detection and the graphical representation of sample content, which are useful supplements to the more sensitive approaches utilised by the bioinformatic workflow.

3.5. Conclusions

A new bioinformatic workflow for sensitive virus detection and discovery in VIDISCA sequence data has been presented, which includes false positive removal and total metagenomic analysis. The workflow has been validated for virus detection in samples derived from individuals infected with known pathogens. The new cluster-profiling analysis, based on the VIDISCA library preparation molecular biology, has been used to flag a novel virus in unclassified reads, serving as a proof of concept for discovery of more divergent viruses.

Data availability

Code is available upon request. For example outputs from the pipeline, see the GitHub repository at: https://github.com/CormacKinsella/VIDISCA-e.g.-output.

Acknowledgements

This research has received funding from the European Union's Horizon 2020 research and innovation programme, under the Marie Skłodowska-Curie Actions grant agreement no. 721367 (HONOURs). We would like to thank Dr. Ad de Groof of Intervet International BV for sharing APPV RT-qPCR data and Arthur W.D. Edridge for helpful feedback on the manuscript.

References

Bankevich, A. et al., 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–77.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Boom, R. et al., 1990. Rapid and simple method for purification of nucleic acids. J. Clin. Microbiol. 28, 495–503.

Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59–60.

Camacho, C. et al., 2009. BLAST+: Architecture and applications. BMC Bioinformatics 10, 421.

Conceição-Neto, N. et al., 2015. Modular approach to customise sample preparation procedures for viral metagenomics: A reproducible protocol for virome analysis. Sci. Rep. 5, 16532.

de Groof, A. et al., 2016. Atypical porcine pestivirus: A possible cause of congenital tremor type A-II in newborn piglets. Viruses 8, 271.

de Vries, M. et al., 2011. A sensitive assay for virus discovery in respiratory clinical samples. PLoS One 6, e16118.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461.

Edridge, A.W.D. et al., 2018. Novel orthobunyavirus identified in the cerebrospinal fluid of a Ugandan child with severe encephalopathy. Clin. Infect. Dis.

Endoh, D. et al., 2005. Species-independent detection of RNA virus by representational difference analysis using non-ribosomal hexanucleotides for reverse transcription. Nucleic Acids Res. 33, e65.

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152.

Kebschull, J.M., Zador, A.M., 2015. Sources of PCR-induced distortions in high-throughput sequencing data sets. Nucleic Acids Res. 43, e143.

Kim, D., Song, L., Breitwieser, F.P., Salzberg, S.L., 2016. Centrifuge: Rapid and sensitive classification of metagenomic sequences. Genome Res. 26, 1721–1729.

Kopylova, E., Noé, L., Touzet, H., 2012. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28, 3211–3217.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN].

Martí, J.M., 2018. Recentrifuge: Robust comparative analysis and contamination removal for metagenomic data. bioRxiv 190934.

Menzel, P., Ng, K.L., Krogh, A., 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat. Commun. 7, 11257.

Ondov, B.D., Bergman, N.H., Phillippy, A.M., 2011. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 12, 385.

Oude Munnink, B.B. et al., 2014. Unexplained diarrhoea in HIV-1 infected individuals. BMC Infect. Dis. 14, 22.

Parrish, C.R. et al., 2008. Cross-species virus transmission and the emergence of new epidemic diseases. Microbiol. Mol. Biol. Rev. 72, 457–70.

Shi, M. et al., 2018. The evolutionary history of vertebrate RNA viruses. Nature 556, 197–202.

Shi, M. et al., 2016. Redefining the invertebrate RNA virosphere. Nature 540, 1–12.

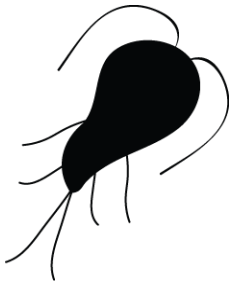
Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

Wood, D.E., Salzberg, S.L., 2014. Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15, R46.

Chapter 3

Entamoeba and *Giardia* parasites implicated as hosts of CRESS viruses

Cormac M. Kinsella, Aldert Bart, Martin Deijs, Patricia Broekhuizen, Joanna Kaczorowska, Maarten F. Jebbink, Tom van Gool, Matthew Cotton, Lia van der Hoek



Nature Communications, 2020

https://doi.org/10.1038/s41467-020-18474-w

Abstract

Metagenomic techniques have enabled genome sequencing of unknown viruses without isolation in cell culture, but information on the virus host is often lacking, preventing viral characterisation. High-throughput methods capable of identifying virus hosts based on genomic data alone would aid evaluation of their medical or biological relevance. Here, we address this by linking metagenomic discovery of three virus families in human stool samples with determination of probable hosts. Recombination between viruses provides evidence of a shared host, in which genetic exchange occurs. We utilise networks of viral recombination to delimit virus-host clusters, which are then anchored to specific hosts using (1) statistical association to a host organism in clinical samples, (2) endogenous viral elements in host genomes, and (3) evidence of host small RNA responses to these elements. This analysis suggests two CRESS virus families (*Naryaviridae* and *Nenyaviridae*) infect *Entamoeba* parasites, while a third (*Vilyaviridae*) infects *Giardia duodenalis*. The trio supplements five CRESS virus families already known to infect eukaryotes, extending the CRESS virus host range to protozoa. Phylogenetic analysis implies CRESS viruses infecting multicellular life have evolved independently on at least three occasions.

Introduction

Determining hosts of viruses is integral to understanding their medical or ecological impact. This is particularly challenging for virus species discovered using metagenomic sequencing, since samples such as stool or environmental matrices contain diverse potential hosts[1,2]. A decade of metagenomic studies have shown that viruses with circular Rep-encoding single-stranded DNA genomes (CRESS viruses) are highly diverse and pervasively distributed[3,4], yet currently, the majority of known CRESS virus genetic diversity falls outside established families with characterised hosts[5]. Five CRESS virus families have experimentally confirmed eukaryotic hosts: *Bacilladnaviridae*, *Circoviridae*, *Geminiviridae*, *Genomoviridae*, and *Nanoviridae*[6], respectively infecting diatoms[7], vertebrates[8,9], plants[10], fungi[11] and plants[12]. Unclassified lineages of metagenomically identified CRESS diversity exist in at least six further clusters labelled CRESSV1 through CRESSV6, and a multitude of chimeric species difficult to place phylogenetically[13].

Unclassified CRESS viruses are frequently found in human and non-human primate stool samples, generating interest into their host specificity and potential impact on health[14,15,16,17]. Classically, virus–host relationships are determined via recognition of host disease, followed by virus isolation in cell culture. Since this is impractical for metagenomically identified viruses, case-control studies are used to reveal associations between viruses and disease. Importantly though, this does not confirm the host; for example, the CRESS virus family *Redondoviridae* is associated with human periodontal disease and critical illness[18], but it remains unknown whether the viruses infect humans or a separate host, itself associated with or causing the observed clinical outcomes.

Genomic evidence of virus–host interactions can directly establish links between species. For instance, the *Smacoviridae*, a CRESS virus family previously assumed to infect eukaryotes, were recently suggested to infect archaea[19] on the basis of CRISPR spacer sequences matching a smacovirus inside the genome of an archaeon. Similarly, virus genomes can integrate into host genomes, leaving endogenous viral elements, identification of which reveals historical infections[20,21]. Searches for endogenous viral elements related to CRESS viruses have revealed integrations into the genomes of eukaryotes, for instance, sequences related to the replication-associated protein (Rep) of *Geminiviridae*, major global crop pathogens, are integrated in the tobacco genome[22].

Rep-like sequences are found in the genomes of the protozoan gut parasites *Entamoeba histolytica* and *Giardia duodenalis*[23], important human pathogens belonging to distantly related genera[24]. The Rep-like elements could imply that the parasites host CRESS viruses, however, the sequences do not belong to a known family[3]. One proposed alternative hypothesis is that that they were gained from bacterial plasmids directly[23], which are thought to be the ancestors of CRESS virus *Rep* genes[25]. Compatible with this, no sequence related to a capsid protein (Cap) has been found integrated in *Entamoeba* or *Giardia* genomes. While several studies have discussed or attempted to identify an association between CRESS viruses and gut parasites[3,26,27,28]—none has been found to date—and indeed no CRESS virus is known to infect any protozoan. Here we provide evidence that the parasite genera *Entamoeba* and *Giardia* are hosts of CRESS viruses, introducing a framework for host determination of metagenomically sequenced viruses that can be widely applied.

Results

Unclassified CRESS viruses are associated to parasites in human stool

Stool samples from 374 individuals (belonging to two independent cohorts, see "Methods") were enriched for viruses using the VIDISCA method, metagenomically sequenced, and bioinformatically analysed to identify unknown CRESS viruses. We used sequence assembly of short reads in combination with inverse PCR and Sanger sequencing to determine 20 full-length CRESS virus coding sequences (accessions MT293410.1–MT293429.1). The 20 sequences included 18 complete genomes covering all untranslated regions, and these had a genome organisation akin to known CRESS viruses, with a conserved nonanucleotide motif at an apparent replication origin, and open reading frames that aligned to viral *Rep* and *Cap* genes (Supplementary Table 1). Using PCR or mapping of sequencing reads to the assembled genomes, we determined that 21 of 374 samples were positive for the viruses.

All 374 samples were also analysed for the presence of *Entamoeba* and *Giardia* parasites using either microscopy, sequencing-based approaches, PCR targeting the 18S ribosomal

RNA, or a combination thereof (see "Methods"). We observed that all 21 of the samples containing one of the CRESS viruses were also positive for either *Entamoeba* or *Giardia* (Table 1 and Supplementary Table 2). Across the 374 samples, presence of any of the 20 viruses was significantly associated with *Entamoeba* or *Giardia* infection using Pearson's chi-squared test ($\chi^2 = 36.77$, $p < 0.001$), therefore we hypothesised that the viruses infected one or both of the parasites. To test the possible host role of other gut protozoa (including *Blastocystis*, *Dientamoeba*, *Cryptosporidium* and *Endolimax* among others), we carried out further parasitological typing on the 21 virus-positive samples (see "Methods"). We found these taxa were absent from all, or a majority of the 21 samples—implying they are not hosts of the viruses (Supplementary Table 2).

Table 1: *Entamoeba* and *Giardia* status of human samples positive for any of the CRESS viruses identified in this study.

Parasite status	Number of samples (N = 374)	Positive for CRESS viruses identified in this study
Entamoeba positive only	130	18
Giardia positive only	3	0
Entamoeba and *Giardia* positive	8	3
Entamoeba and *Giardia* negative	233	0

Whole CRESS virus genomes are integrated into parasite genomes

In order to identify endogenous viral elements related to the identified CRESS viruses, we aligned all 20 coding sequences to GenBank databases, namely the non-redundant nucleotide (BLASTn, Supplementary Table 3), protein (BLASTx, Supplementary Table 4), and whole-genome shotgun contigs of *Entamoeba* and *Giardia* (BLASTn, Supplementary Table 5). Viral queries aligned with high identity and coverage to nucleotides and predicted proteins from parasite genomes, suggesting the presence of CRESS virus-derived endogenous viral elements. The 20 viruses were not uniform in their database hits, showing genetic variation among them; each virus strongly aligned to sequences from either *Entamoeba* or *Giardia*, but not both, suggesting the presence of distinct viral lineages with independent virus–host relationships. Among viruses aligning to sequences from the *Entamoeba* genus, variability was also observed in the parasite species—queries either hit *E. histolytica*, *E. dispar*, *E. nuttalli*, or *E. invadens*. Among viruses aligning to sequences from *Giardia duodenalis*, alignments were found against major genotypes infecting humans, specifically A2 and B. Importantly, alignment to parasite genomes revealed

evidence of whole virus genome integrations. For example, one virus genome (accession MT293413.1) aligned inside an 11.6 kilobase (kb) contig from *E. dispar* (AANV02000527.1) with 100% query coverage and 84% nucleotide identity (Fig. 1a), while another (accession MT293421.1) aligned inside a 15.2 kb contig from *G. duodenalis* (AHGT01000120.1) with 99% query coverage and 73% nucleotide identity. As the only known examples of parasite endogenous viral elements containing both the *Rep* and *Cap* viral genes, they cast doubt on the hypothesis that Rep-like elements in protozoal genomes were derived from bacteria[23]. Since CRESS virus integration is likely mediated by the Rep protein during viral genome replication in the host nucleus[29], the elements directly implicate *Entamoeba* and *Giardia* as hosts.

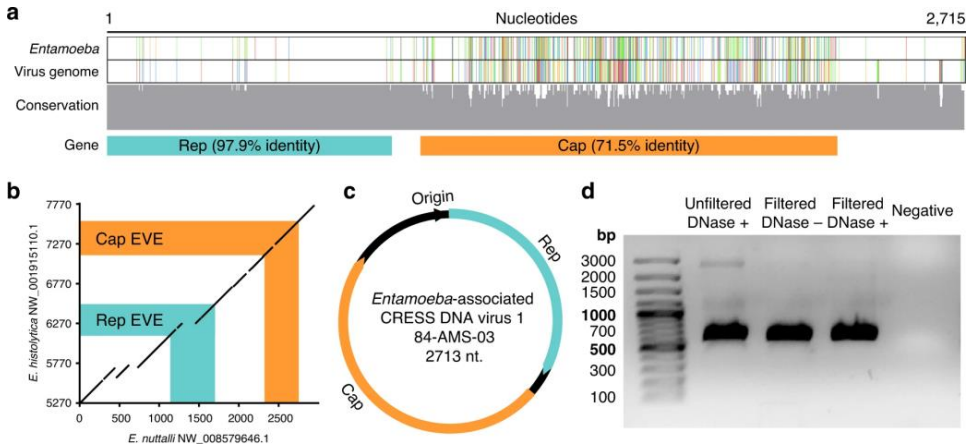


Fig. 1: Whole CRESS virus genomes are integrated in Entamoeba genomes. a Cropped nucleotide alignment between *Entamoeba dispar* contig (AANV02000527.1) containing a complete virus integration and the genome of *Entamoeba*-associated CRESS DNA virus 1, isolate 84-AMS-03 (accession MT293413.1); also see Supplementary Fig. 2. Coloured vertical bars denote single nucleotide variations between the sequences (adenine = green, guanine = red, thymine = blue, cytosine = orange), with conservation across the alignment displayed below. **b** Dotplot of BLAT generated nucleotide alignment between endogenous viral elements and flanking sequence from two closely related *Entamoeba* species (*x*-axis sequence reverse complemented). **c** Example of the circular genome organisation of identified CRESS viruses. **d** Exogenous virus DNA is protected by a viral capsid, as it can be PCR-amplified after filtration and treatment with DNase (one independent experiment).

We next considered and eliminated potential sources of error, firstly, that parasite genomes did not truly contain CRESS endogenous viral elements, but rather that the assemblies were contaminated with virus genome sequences found in the original sample or reagents. To eliminate this possibility, we compared independently generated genome assemblies of *E. histolytica* and *G. duodenalis*, which were derived from parasite stocks in different laboratories or biobanks, and included strains isolated from patients across multiple countries and years. We could identify the same endogenous viral elements in several of the

assemblies, for example an element (EMD43492.1) from *E. histolytica* strain KU27, isolated in Japan in 2001, was also found in strain HM-3:IMSS, isolated in Mexico in 1972 (100% coverage, 100% sequence identity), and three independent assemblies of strain HM-1:IMSS, isolated in Mexico in 1967 (100% coverage, 99.9% sequence identity, Supplementary Tables 6 and 7). Furthermore, in one case an element and its flanking sequence could be aligned between the closely related species *E. histolytica* and *E. nuttalli* (Fig. 1b). This provides evidence of a shared viral integration that must have originated prior to host speciation, although the date of this divergence is currently unknown. Interestingly, *G. duodenalis* elements displayed a lineage-specific distribution, found universally in assemblies of lineages A2 and B, but absent from lineage A1 assemblies and the lone assembly of lineage E (Supplementary Table 7). The results suggest population-level fixation of elements in specific parasite lineages, rather than contamination leading to a misassembly. To rule this out however, for *E. histolytica* HM-1:IMSS we closely examined raw sequencing coverage across a selected endogenous viral element and its flanking sequence, showing that Sanger sequence reads span the element with no coverage aberrations (Supplementary Fig. 1A). We secondarily confirmed this by analysing the raw reads of strain KU27, isolated over thirty years later, with consistent results (Supplementary Fig. 1B). For *G. duodenalis* we examined the elements present in a recent reference quality assembly (GCA_011634595.1, isolate GS, lineage B), since this was generated using a combination of conventional short-reads and nanopore long-reads[30]. The latter technology vastly improves the scaffolding and repeat-resolution of assemblies, and confirmed the presence of endogenous viral elements within host sequence, even resolving a 10 kb-long tandemly repeated element not previously detectable in assemblies relying on short-read technology alone (Supplementary Fig. 1C). For further evidence that the endogenous viral elements were a true genomic feature, we looked for a small RNA response against them in *E. histolytica*, since the parasite silences its own genes post-transcriptionally via the RNA interference pathway[31]. We utilised public data comprising small RNAs immunoprecipitated in association with AGO2-2[32], which is the component of the RNA interference pathway responsible for binding RNA guide strands and target mRNA cleavage, mapping the small RNAs to *E. histolytica* contigs containing endogenous viral elements (Supplementary Fig. 2). We found small RNA coverage peaks coinciding with several endogenous viral elements, including one known to be transcriptionally active[33], suggesting host silencing of the elements. A notable but untested implication is that mRNAs from exogenous CRESS viruses infecting *E. histolytica* may also be silenced by such a response, which may therefore function in antiviral defence, since some small RNA sequences also had exact matches to the CRESS virus sequences of our study (Supplementary Fig. 2).

We secondly confirmed that viral genomes identified in human clinical samples were derived from exogenous viruses, since an alternative possibility is that they represented endogenous viral elements sequenced from parasite chromosomal DNA. The likelihood of this occurrence was minimised by the VIDISCA sequencing library preparation, which included removal of cell debris and degradation of residual chromosomal DNA via DNase

treatment, however, for confirmation, we visually inspected viral reads to verify sequence overlap at the beginning and end of contigs. In this way, we could establish that the majority of viral coding sequences found in human samples were circular whole genomes ($n = 18$, Fig. 1c), and therefore were not from a larger sequence context such as a parasite chromosome. Finally, since exogenous viruses are small in comparison to eukaryotic cells, and their genomes are encapsidated in a protein shell, we experimentally confirmed these features. We filtered supernatant from virus-positive faecal suspension through 1200 and 200 nm pores, and treated the filtrate with DNase to remove unprotected DNA, finding that viral DNA could still be amplified by PCR (Fig. 1d). This shows that the genetic material was protected by a structure, most likely a capsid.

Protozoa-infecting viruses are from previously unknown families

Virus alignments to endogenous viral elements in parasite genomes already suggested that distinct viral lineages with independent virus–host relationships were present among the sequences. We, therefore, resolved the relationships of the exogenous viruses by building a maximum-likelihood phylogenetic tree of the Rep protein. Sequences extracted from Rep-like endogenous viral elements in *Entamoeba* spp. and *G. duodenalis* were included to identify their closest relatives and reveal which virus lineages were the original donors. Known CRESS virus diversity was incorporated by modifying a previously published chimaera-free Rep protein database of CRESS virus families and clusters[13]. We included the *Redondoviridae* in the dataset in addition to our own sequences and the closest viral relatives of our 20 sequences identified by BLAST searches. The viruses belonged to three strongly supported monophyletic Rep lineages, all phylogenetically positioned outside known families (Fig. 2a). Protein sequences from parasite endogenous viral elements clustered within each of the three lineages, and never outside, a firm indication that the exogenous virus lineages were the original donors of the endogenous viral elements. Notably, *Entamoeba* endogenous viral elements clustered exclusively within two of the three lineages, while *Giardia* endogenous viral elements only clustered with the third, indicating their different host specificity. Since the lineages do not belong to a known CRESS virus family, we propose the establishment of three virus families to house them. Following the practice of naming CRESS virus taxa with reference to their circular genomes, we suggest naming the families after three rings from Tolkien's canon: *Naryaviridae* and *Nenyaviridae* for the two *Entamoeba*-infecting virus families and *Vilyaviridae* for the *Giardia* infecting family. The three families are phylogenetically distributed among known CRESS virus diversity, and imply that lineages infecting multicellular life evolved on at least three independent occasions, namely (1) the lineage including *Geminiviridae* and *Genomoviridae*, (2) the *Circoviridae*, and (3) the *Nanoviridae*. The *Nenyaviridae* are nested within the CRESSV2 cluster, suggesting these viruses may also infect protozoa.

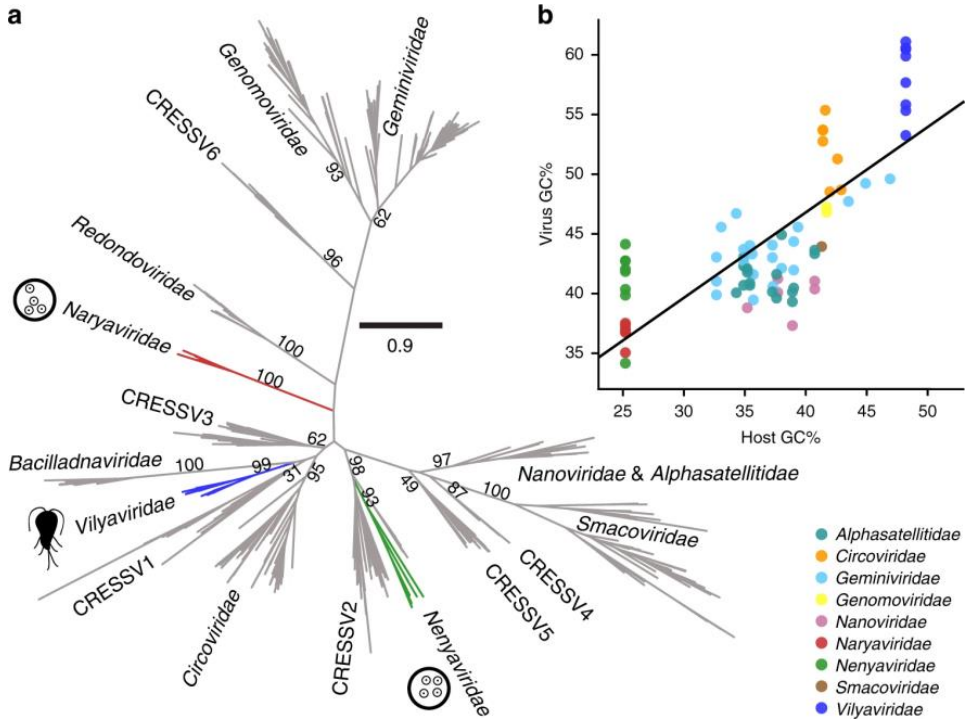


Fig. 2: Parasite-infecting CRESS virus genomes are distinct from known CRESS diversity. a Phylogenetic maximum-likelihood tree of the Rep protein, scale bar refers to amino acid substitutions per site, numerical values represent bootstrap support of major nodes. The *Naryaviridae*, *Nenyaviridae*, and *Vilyaviridae* contain endogenous viral element sequences extracted from host genomes, respective pictograms of *Entamoeba* (tetranucleate cyst stage) and *Giardia* (flagellated trophozoite stage) are shown to indicate this. Five public viral genomes were also found to cluster within these families (MG571899.1, KU043415.1, MH617639.1, KY487991.1 and LC406405.1). **b** Virus GC-content positively correlates with host GC-content (linear regression, $n = 79$ biologically independent viral genome sequences, $r^2 = 0.58$, $p = 0.01$).

We delimited CRESS virus genera using a cutoff of 50% Rep protein identity, following a recent literature example[18]. Genera infecting the same host genus were assigned a Greek number and named with reference to the host (ent for *Entamoeba* and gia for *Giardia*) (Supplementary Table 8). The *Naryaviridae* were thus divided into two genera (*Protoentvirus* and *Deuteroentvirus*), *Nenyaviridae* into two (*Tritoentvirus* and *Tetartoentvirus*), and *Vilyaviridae* into three (*Protogiavirus*, *Deuterogiavirus*, and *Tritogiavirus*). Although the viruses display large intra-family sequence diversity, the families do share distinctive features: *Naryaviridae* and *Nenyaviridae* genomes have sense open reading frames, while *Vilyaviridae* genomes have either ambisense or antisense open reading frames. Nucleotide usage measured by GC-content varies within each of the three

families, but *Naryaviridae* and *Nenyaviridae* have on average 37% and 42% respectively, while the *Vilyaviridae* have a high 59%. The GC-contents of *Naryaviridae* and *Vilyaviridae* respectively represent low and high extremes among eukaryotic CRESS viruses. Since a positive association between host nucleotide usage and virus nucleotide usage has previously been observed among single-stranded DNA bacteriophages[34], we hypothesised that this also underlay the observed distribution. To test this, we modelled the GC-content of CRESS virus lineages against those of known or proposed hosts using linear regression (Fig. 2b and Supplementary Table 9). For *Entamoeba* and *Giardia* we used the GC-content of *E. histolytica* (25.2%, assembly GCA_000365475.1) and *G. duodenalis* (48.2%, assembly GCA_000498735.1), respectively. A positive association was found between virus and host nucleotide usage ($r^2 = 0.58$, $p = 0.01$), consistent with the proposed virus–host relationships. The association may be due to codon usage bias, wherein virus codon usage is constrained by host transfer RNA availability[35]. Despite the positive association, exogenous viruses from the three families did have a higher GC-content than their hosts by an average of 12.6%, suggesting the existence of additional selection pressure on GC-content counter to that of transfer RNA mediated protein translation efficiency. In contrast with exogenous viruses, endogenous representatives of each family had a reduced GC-content, in some cases closely resembling that of the host (Supplementary Fig. 3). We hypothesise that this is due to genetic drift resulting from relaxed selection on elements after integration, wherein the oldest elements may have the lowest GC-content.

Viral recombination networks identify virus–host clusters

During genomic analysis of the CRESS viruses we observed a striking bimodal genome length distribution in both *Naryaviridae* and *Nenyaviridae*, but not in *Vilyaviridae* (Fig. 3a). BLAT alignment between two *Naryaviridae* genomes from the ends of the length distribution showed that the irregularity was caused by *Cap* genes of different lengths (averaging 179 and 439 amino acid residues respectively) with no detectable nucleotide sequence similarity, while the *Rep* genes were closely related (Fig. 3b). The two Cap proteins also had no detectable protein sequence identity upon pairwise BLASTp analysis, suggesting that the smaller of the two is not simply a partial protein, but a protein of different ancestry. To ensure that this was not a result of genome misassembly, we confirmed that Sanger sequencing reads overlapped both the *Rep* and *Cap* genes. Different ancestry of *Cap* genes found in combination with a *Rep* gene strongly suggested recombination of complete genetic modules (i.e. replicative and structural genes). Recombination between viruses occurs during genome replication within the host, and evidently the host range of a virus dictates its potential recombination partners[36]. Detection of recombination between viruses can therefore be used to group together viruses into virus–host clusters.

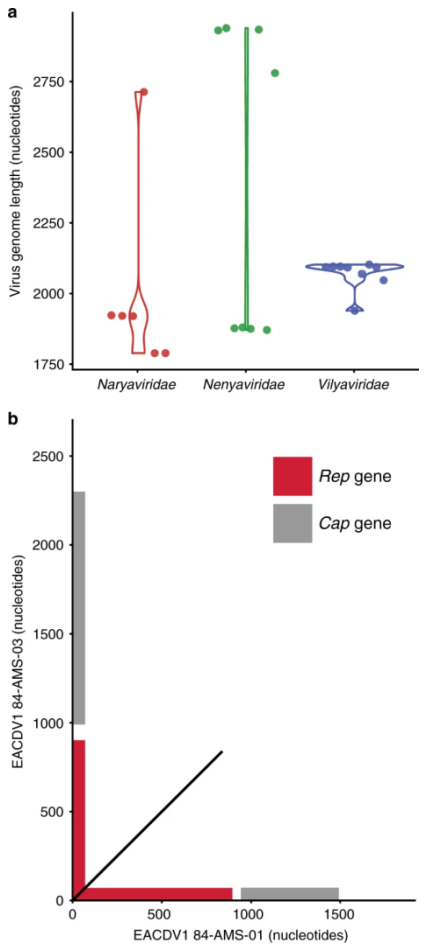


Fig. 3: *Cap* genes of different ancestry in *Naryaviridae* and *Nenyaviridae*. a Genome length variation in *Naryaviridae*, *Nenyaviridae*, and *Vilyaviridae*. Eighteen complete CRESS virus genomes identified in this study were plotted alongside five complete publicly available genomes. **b** Dotplot of BLAT generated nucleotide alignment between a short and a long genome from the *Naryaviridae*, showing no detectable alignment between the *Cap* genes.

To investigate recombination among the identified CRESS viruses, we constructed maximum-likelihood phylogenetic trees of Rep and Cap protein sequences from the three viral families, also including endogenous viral elements if *Rep* and *Cap* genes were found in close proximity in the protozoal genome (Fig. 4). Since *Cap* genes could not be globally aligned together, we first separated them into similar protein clusters which were then aligned and analysed individually. The Rep proteins were resolved into the three groups previously observed, corresponding with the three viral families. The Cap proteins were also divisible into three clusters, and we subsequently refer to these as CRESS virus *Cap*

assemblages (CCAs). We visualised gene swapping between lineages by linking proteins extracted from the same genome across the two phylogenies, and this uncovered clear evidence of recombination of genetic modules between the *Naryaviridae* and *Nenyaviridae*. Members of these *Rep* families possessed either CCA1 (averaging 467 amino acid residues) or CCA2 (averaging 180 amino acid residues), with all four possible *Rep* and *Cap* gene combinations represented. Importantly, while evidence of recombination was also visible within the *Vilyaviridae*, they always possessed CCA3, therefore no evidence for recombination between *Vilyaviridae* and members of either the *Naryaviridae* or *Nenyaviridae* was found. The data strongly support the proposed host-range of the viruses, specifically *Naryaviridae* and *Nenyaviridae* sharing the same host, with *Vilyaviridae* infecting a separate one. Further, they provide a practical framework to identify virus–host clusters in an unbiased way with no a priori knowledge of the potential host required.

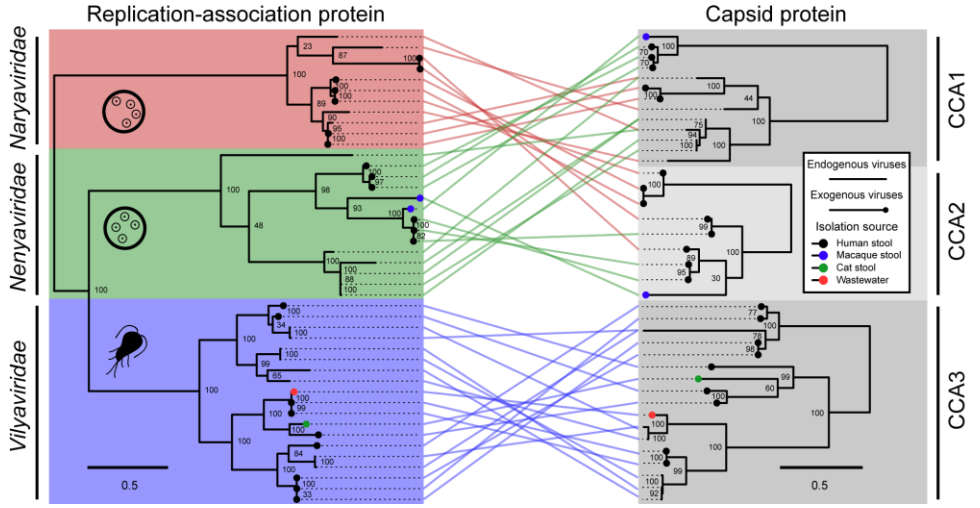


Fig. 4: Recombination of genetic modules between virus families infecting the same host. Phylogenetic maximum-likelihood trees of viral Rep and Cap proteins, scale bars refer to amino acid substitutions per site, numerical values represent bootstrap support. Lines connect genes from the same virus or physically close endogenous viral genes. Pictograms of *Entamoeba* (tetranucleate cyst stage) and *Giardia* (flagellated trophozoite stage) are shown to indicate virus host. CCA = CRESS virus *Cap* assemblage.

Virus families occur alongside specific host genera in human stool

At the outset of investigation, we focused on the association between CRESS viruses and both *Entamoeba* and *Giardia* parasites collectively; however, evidence from endogenous viral elements and patterns of recombination among discovered viruses suggested that *Naryaviridae* and *Nenyaviridae* infect *Entamoeba*, while *Vilyaviridae* infect *Giardia*. We, therefore, tested the statistical associations of the families to their specific proposed host in

human samples using Pearson's chi-squared test, grouping *Naryaviridae* and *Nenyaviridae* together because of recombination between their genomes. Across all 374 study subjects, *Naryaviridae* and *Nenyaviridae* were strongly associated with *Entamoeba* parasites ($\chi^2 = 32.34$, $p < 0.001$), but not with *Giardia* ($\chi^2 = 0.57$, $p = 0.45$), while *Vilyaviridae* were strongly associated with *Giardia* ($\chi^2 = 99.8$, $p < 0.001$). *Vilyaviridae* were also positively associated to *Entamoeba*, however at a greatly reduced significance compared to *Giardia* ($\chi^2 = 5.17$, $p = 0.02$). This result is likely explained by *Entamoeba* coinfections in all 3 *Vilyaviridae* positive samples; indeed, *Entamoeba* coinfection was found in 73% of all *Giardia* positive samples (Table 1 and Supplementary Table 2). Although the cohorts examined here may not be representative of wider parasite populations, the prevalence of *Nenyaviridae* or *Naryaviridae* virus infections was 13% among *Entamoeba* cases (18 of 138), while *Vilyaviridae* had a prevalence of 27% among *Giardia* cases (3 of 11). The observed association between the viruses and their hosts in stool enabled a preliminary investigation into the biogeographic distribution of the three families. We mapped reads from public metagenome datasets derived from faecally polluted wastewater or primate stool to our viral genomes. We found reads from *Naryaviridae*, *Nenyaviridae*, and *Vilyaviridae* were detectable in the datasets examined, sourced from localities across North and South America, Europe, Africa, and Asia (Supplementary Fig. 4). This suggests the virus distributions are large, mirroring those of the hosts.

Discussion

Here we report three CRESS virus families, *Naryaviridae* and *Nenyaviridae* infecting *Entamoeba*, and *Vilyaviridae* infecting *Giardia duodenalis*. Our study expands the number of CRESS families known to infect eukaryotes from five to eight, including the only groups recognised to infect protozoa. The investigation provides the only genome sequences of viruses infecting *Entamoeba*, nearly 50 years after the first of a series of papers studying infectious agents causing cell lysis in axenic *E. histolytica* culture[37]. For *Giardia*, one RNA virus species in the *Totiviridae* (Giardia lamblia virus) was discovered in 1986[38], and the *Vilyaviridae* represent the second group of viruses. The discovery of viruses infecting *Entamoeba* and *Giardia*—collectively responsible for 300 million human disease cases annually[39]—should precipitate investigation of their potential impact on the clinical outcome of parasite infection. It is understood that only a subset of *Entamoeba* and *Giardia* infections result in symptomatic disease[40,41], however, not all the factors underlying case variation are resolved. For example, *E. histolytica* interactions with gut bacteria are thought to play a role in pathogenesis[42], but the effects of viruses are unexplored. As viruses can modulate parasite pathogenicity directly or indirectly via interaction with human immunity, they may result in parasite hypovirulence[11] or hypervirulence[43].

A large proportion of recognised virus genomes are divorced from their biological hosts. Targeted virus discovery from potential host taxa has a vital role to play in resolving this[44],

however, in instances of hosts intractable to culture, high-throughput methods must rely on viral genome sequences alone. Machine-learning algorithms trained on viral sequences with known hosts offer one possible approach[45]; however, due to their reliance on conserved sequence signals between training and test data, they will suffer from increasingly coarse prediction for divergent viruses. As we show, construction of viral recombination networks provides direct and unbiased biological evidence of shared hosts among virus genomes, even when individual genes are highly divergent or non-homologous. Given the highly consequential roles protozoa play in global health and ecosystem processes, deciphering additional unknown virus–host relationships among them is imperative.

Methods

Clinical samples

The 374 human subjects analysed here were from two cohorts. Cohort 1: stool samples of 194 HIV-1 infected individuals not on active antiretroviral therapy, who visited the out-patient clinic at the Amsterdam Medical Center in 1994 and 1995, as part of a study on unexplained diarrhoea[46,47]. Criteria for inclusion in the study were proven HIV-1 infection and being aged 18 years or older. Cohort 2: Stool samples of 85 HIV-1 positive and 95 HIV-1 negative men having sex with men (MSM) as part of the ACS, a prospective cohort study among HIV-positive and HIV-negative MSM, initiated in 1984[48]. Studies were approved by the Medical Ethics Committee of the Amsterdam University Medical Center, the Netherlands (MEC 07/182). Written informed consent of each participant was obtained at enrolment of both cohorts.

VIDISCA library preparation and sequencing of human faecal samples

At collection, faecal samples were suspended 1:3 in broth containing penicillin, streptomycin, and amphotericin B, and stored at −80 °C until processing. Sample suspension (150 µl) was transferred to a reaction tube and centrifuged (10 min at 5000 g) to pellet solid matter and cellular debris. Supernatant was treated with 20 µl TURBO DNase (Thermo Fisher Scientific, Waltham, MA, USA) for 30 min at 37 °C (to remove naked DNA). Nucleic acids were extracted using the Boom method[49] and reverse transcription was done using non-ribosomal hexamer primers designed to avoid mammal rRNA sequences[50]. This was followed by second strand synthesis and a cleanup via phenol/chloroform extraction and ethanol precipitation. Library preparation for the two cohorts varied from this point, since two different sequencing technologies were used. For cohort 1 standard VIDISCA library preparation was carried out[51]. Briefly, double-stranded DNA was digested with Mse1 restriction enzyme, and sequencing adapters were ligated to sticky ends. Libraries were amplified before size selection of fragments between 200 and 600 bp, quantification, and pooling. Sequencing was then done on an IonTorrent PGM instrument. For cohort 2, double-stranded DNA was fragmented to an average length of

400–500 bp, sequencing adapters were ligated, and libraries were amplified before sequencing with Illumina MiSeq instruments (150 bp paired end)[52]. Sequence reads associated with this study have been deposited in the European Nucleotide Archive (ENA) under study accession PRJEB35571.

CRESS virus identification and characterisation

Sequence reads from cohort 1 were analysed to discover viruses[53]. Briefly, non-rRNA reads were identified using SortMeRNA v2.1[54] and made non-redundant using CD-HIT v4.7[55]. Non-redundant reads were then aligned to viral proteins using UBLAST[56], and false positives were reduced via BLASTn[57] alignment of putative viral matches to the GenBank non-redundant nucleotides. Outputs were visualised with KronaTools v2.7[58] and inspected to identify candidate CRESS virus reads. Two genomes were amplified via inverse PCR, the sequences of which were determined using Sanger sequencing (accessions MT293412.1 and MT293415.1). All primers are reported in Supplementary Table 10. An iterative search procedure was then carried out to identify additional samples containing related CRESS viruses. Predicted protein sequences were extracted from the two genomes and used as queries against reads from cohort 1 using UBLAST. This was also carried out against contigs assembled from cohort 2 sequencing data using SPAdes v3.5.0[59]. Further putative CRESS virus hits were manually curated or completed with Sanger sequencing, and were then used in subsequent searches. The process resulted in a final count of 20 CRESS virus coding sequences, 18 of which were complete genomes.

To determine a final list of samples regarded as virus positive, sequence reads from each cohort were mapped to the 20 virus coding sequences using BWA-MEM v0.7.17[60]. Reads mapping to multiple references were reassigned to their single most-likely reference using the PathoID module of PathoScope v2.0.7[61]. High-depth Illumina sequencing is prone to barcode swapping within flow cells, which may result in false positives; therefore, for cohort 2 a cutoff was imposed for a sample to be regarded as positive. Specifically, virus reads from *Entamoeba*-infecting or *Giardia*-infecting families had to make up at least 0.05% of sample reads (in instances where samples had received repeat sequencing, only the run receiving the highest number of sequences was analysed). In addition to the sequencing-based approach described, any PCR positive samples were also included.

Virus protein sequences extracted from open reading frames were queried against the Reference Proteome database with pHMMER[62] and best hits were recorded. DNA secondary structure surrounding the putative nonanucleotide origin motif was assessed using MFOLD[63] to confirm it was situated on a predicted stem loop. Circularity of viruses was confirmed by visual inspection of genomes and mapped reads, specifically reads that overlapped with both the beginning and end of genome sequences. To confirm that viral DNA was protected by a capsid, supernatant was first passed through a filter with 1200 nm pores, then 200 nm (GE Healthcare Life Sciences, Chicago, USA), followed by treatment with TURBO DNase (Thermo Fisher Scientific, Waltham, MA, USA). Subsequently viral

nucleic acid was extracted with the Boom method, and PCR was carried out. To compare CRESS virus GC-content with that of their hosts, the Virus–Host DB[64] was used in conjunction with the GenBank genomes resource to compile this information for virus–host pairs.

Parasitological typing

Faecal samples from cohort 1 were examined by light microscopy for the presence of intestinal parasites (with both direct smears and concentrations using the Ridley technique). From both cohorts, sequence reads were mapped using BWA-MEM to parasite ribosomal RNA reference sequences, with aligning sequences then queried against the GenBank non-redundant nucleotide database. Reads with the best hit to a parasite ribosomal RNA reference, and a minimum alignment of 50 nt at over 95% nucleotide identity was retained as hits. Hits were aligned to diagnostic parasite reference sequences to type the parasite species where possible. Sequence reads were also mapped using BWA-MEM to predicted mRNA sequences from parasite genomes, specifically *E. histolytica* (GCF_000208925.1) and *G. duodenalis* (GCF_000002435.1). Predicted mRNA databases were first curated using identity searches to remove sequences derived from endogenous viral elements and ribosomal RNA. Hits were also filtered to allow only those with a minimum alignment of 50 nt at over 95% nucleotide identity to their respective subject sequence. The possibility of barcode swapping in cohort 2 Illumina data led us to impose a cutoff for a sample to be called as positive; specifically, the parasite sequence reads as a percentage of the total reads had to be greater than the lower quartile value. For a selection of samples from cohort 1, confirmatory testing was done with *E. histolytica* and *E. dispar* diagnostic qPCRs, in addition to *Entamoeba* generic PCR combined with Sanger sequencing of amplicons. Due to generally low read counts observed for *Giardia*, all 21 virus-positive samples were subjected to a confirmatory *Giardia* diagnostic qPCR. The prevalence of *Giardia* infection among our cohort participants was 2.94% (11 of 374), and the prevalence of *Entamoeba* infection was 36.90% (138 of 374). Our participants were 93% MSM, and these *Giardia* and *Entamoeba* frequencies are concordant with previously reported data from this demographic (from 1% to 18% for *Giardia* with a median of 5% infection, and from 3% to 33% for *Entamoeba* with a median of 22% infection[65]). To confirm that other protozoa were not the viral hosts, the 21 virus-positive samples were tested for additional parasites: *Dientamoeba*, *Cryptosporidium*, and *Blastocystis* were tested by diagnostic qPCR, while *Endolimax*, *Chilomastix*, *Pentatrichomonas*, and *Retortamonas* 18S rRNA sequences were analysed in the same manner described above.

Endogenous viral element analysis

CRESS virus genomes were aligned to GenBank databases: the non-redundant nucleotide using BLASTn, the non-redundant protein using BLASTx, and the whole-genome shotgun contigs of *Entamoeba* and *Giardia* using BLASTn. Nucleotide and protein sequences of hits were extracted and manually curated to use in subsequent analyses. Comparison

between independent assemblies of *E. histolytica* and *G. duodenalis* (to confirm consistency of endogenous viral element presence) was done using BLASTn of endogenous *Rep* gene elements from each genus against each assembly, recording the best aligning hit. Pairwise comparisons between sequences were all performed using BLAT via the MAFFT online server[66]. Available genome assemblies from relatives of *Entamoeba* and *G. duodenalis* were also analysed for the presence of elements, specifically *Mastigamoeba balamuthi* (GCA_902651635.1), *Spironucleus salmonicida* (GCA_000497125.1), *Trichomonas vaginalis* (GCA_000002825.1), and *G. muris* (GCA_006247105.1); however, none of these assemblies contained elements belonging to the *Naryaviridae*, *Nenyaviridae*, or *Vilyaviridae*. To assess read coverage across *E. histolytica* contig NW_001915013.1, raw sequencing reads were downloaded from the TraceDB (isolate HM1:IMSS, https://ftp.ncbi.nlm.nih.gov/pub/TraceDB/entamoeba_histolytica/) and ENA (isolate KU27, accessions SRR071802 and SRR072203). BWA-MEM was used to map reads to the complete reference contig, followed by visualisation of coverage using CodonCode Aligner v9.0.1. Easyfig v2.2.5[67] was used to visualise pairwise identity between *G. duodenalis* contigs VSRU01000012.1 and AHHH01000265.1. To identify evidence of an RNA interference response against endogenous viral elements, BWA-backtrack[60] was used to map *E. histolytica* AGO2-2 associated small RNAs from ENA project PRJNA187070[32] against contigs containing elements from the *E. histolytica* RefSeq genome assembly (GCA_000208925.2). Prior to mapping, sequencing adapters were trimmed using BBDuk (http://jgi.doe.gov/data-and-tools/bb-tools/), and sequences over 40 nt and under 15 nt were discarded. Reads mapping with zero sequence mismatches were retained, and coverage of contigs was calculated using the SAMtools mpileup utility[68]. Positions of endogenous viral elements and small RNA coverages were visualised for a selection of contigs using Circos v0.69-8[69].

Phylogenetic analysis and pairwise protein comparison

Phylogenetic analysis of the Rep protein utilised a previously compiled chimaera-free dataset[13], with the addition of the *Redondoviridae*[18], five viral sequences found during BLASTn searches of the GenBank non-redundant nucleotide database, and our CRESS virus sequences (both exogenous and endogenous viruses). Rep proteins were aligned using MAFFT v7[66] with the L-INS-i option leaving gappy regions unaligned. The resulting alignment was trimmed using trimAl v1.4[70] set to gappyout. Maximum-likelihood phylogenetic analysis was performed using RaxML v8.2.9[71] with the PROTCATGTR substitution model and automatic bootstopping, which stopped rapid bootstrap searching after 350 replicates. Treefiles were visualised using Figtree v1.4.4 (https://github.com/rambaut/figtree/releases). The same methods were applied for phylogenetic analysis of the three Rep protein families in isolation, as well as their corresponding Cap proteins. For delimitation of Rep genera, pairwise comparison was carried out using the online tool SIAS (available at: http://imed.med.ucm.es/Tools/sias.html), with the denominator set to mean length of sequences.

Public metagenome data

Data to estimate the global distribution of parasite-infecting CRESS viruses was obtained from a number of public metagenomes and mapped using BWA-MEM to virus genomes. Wastewater samples from ENA project PRJNA169010[72] were from Maiduguri (Nigeria), Kathmandu (Nepal), Bangkok (Thailand), and San Francisco (USA); project PRJNA70623[73] samples were from Addis Ababa (Ethiopia), Barcelona (Spain), and Pittsburgh (USA); project PRJNA322301[74] was from Tallahassee (USA); project PRJNA434744[75] was from Cincinnati (USA); and project PRJNA385831[76] was from Sheboygan (USA). Macaque stool from project PRJNA299332[14] was from the California National Primate Research Center (USA). Human stool from project PRJNA418044[26] was from Caracas (Venezuela) and remote villages in South-East Venezuela; and project PRJEB9524[77] was from Uganda. A further site was annotated based on a public virus genome (LC406405.1) which clustered within the *Vilyaviridae*, sampled from a cat in Japan[78].

Data availability

Viral genomes and coding sequences are available under NCBI accessions MT293410.1–MT293429.1. Raw sequencing reads are available under European Nucleotide Archive study accession PRJEB35571. Protein alignments and tree files are available from Figshare (https://figshare.com/projects/Entamoeba_and_Giardia_parasites_implicated_as_hosts_of_CRESS_viruses/84065). GenBank databases are available via NCBI (https://www.ncbi.nlm.nih.gov/), and the Reference Proteome database was integrated with the pHMMER web service (https://www.ebi.ac.uk/Tools/hmmer/search/phmmer).

Acknowledgements

This research was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions grant agreement no. 721367 (HONOURs). We thank Alexander Suh and Arthur Edridge for helpful discussions and feedback throughout the study, and Margreet Bakker for excellent management of the storage and selection of clinical samples. We also gratefully acknowledge the Amsterdam Cohort Studies (ACS) on HIV infection and AIDS, a collaboration between the Public Health Service of Amsterdam, the Amsterdam UMC of the University of Amsterdam, Sanquin Blood Supply Foundation, Medical Center Jan van Goyen, and the HIV Focus Center of the DC-Clinics. The ACS is part of the Netherlands HIV Monitoring Foundation and financially supported by the Center for Infectious Disease Control of the Netherlands National Institute for Public Health and the Environment. The authors thank all ACS participants for their contribution. Supplementary Fig. 4 was created using world border templates available from http://thematicmapping.org/downloads/world_borders.php, provided by Bjørn Sandvik.

References

1. Finkbeiner, S. R. et al. Metagenomic analysis of human diarrhea: Viral detection and discovery. PLoS Pathog. 4, e1000011 (2008).

2. Gregory, A. C. et al. Marine DNA viral macro- and microdiversity from pole to pole. Cell 177, 1109-1123.e14 (2019).

3. Rosario, K. et al. Virus discovery in all three major lineages of terrestrial arthropods highlights the diversity of single-stranded DNA viruses associated with invertebrates. PeerJ 6, e5761 (2018).

4. Zhao, L., Rosario, K., Breitbart, M. & Duffy, S. Eukaryotic circular Rep-encoding single-stranded DNA (CRESS DNA) viruses: Ubiquitous viruses with small genomes and a diverse host range. Adv. Virus Res. 103, 71–133 (2019).

5. Simmonds, P. et al. Virus taxonomy in the age of metagenomics. Nat. Rev. Microbiol. 15, 161–168 (2017).

6. Lefkowitz, E. J. et al. Virus taxonomy: The database of the International Committee on Taxonomy of Viruses (ICTV). Nucleic Acids Res. 46, D708–D717 (2018).

7. Shirai, Y. et al. Isolation and characterization of a single-stranded RNA virus infecting the marine planktonic diatom Chaetoceros tenuissimus Meunier. Appl. Environ. Microbiol. 74, 4022–4027 (2008).

8. Ritchie, B. W., Niagro, F. D., Lukert, P. D., Steffens, W. L. & Latimer, K. S. Characterization of a new virus from cockatoos with psittacine beak and feather disease. Virology 171, 83–88 (1989).

9. Ellis, J. et al. Isolation of circovirus from lesions of pigs with postweaning multisystemic wasting syndrome. Can. Vet. J. 39, 44–51 (1998).

10. Varma, A. & Malathi, V. G. Emerging geminivirus problems: A serious threat to crop production. Ann. Appl. Biol. 142, 145–164 (2003).

11. Yu, X. et al. A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. Proc. Natl. Acad. Sci. 107, 8387–8392 (2010).

12. Chu, P. W. G. & Helms, K. Novel virus-like particles containing circular single-stranded DNAs associated with subterranean clover stunt disease. Virology 167, 38–49 (1988).

13. Kazlauskas, D., Varsani, A. & Krupovic, M. Pervasive chimerism in the replication-associated proteins of uncultured single-stranded DNA viruses. Viruses 10, 187 (2018).

14. Kapusinszky, B., Ardeshir, A., Mulvaney, U., Deng, X. & Delwart, E. Case-control comparison of enteric viromes in captive rhesus macaques with acute or idiopathic chronic diarrhea. J. Virol. 91, e00952-17 (2017).

15. Phan, T. G. et al. Small circular single stranded DNA viral genomes in unexplained cases of human encephalitis, diarrhea, and in untreated sewage. Virology 482, 98–104 (2015).

16. Phan, T. G. et al. The fecal virome of South and Central American children with diarrhea includes small circular DNA viral genomes of unknown origin. Arch. Virol. 161, 959–966 (2016).

17. Altan, E. et al. Small circular Rep-encoding single-stranded DNA genomes in Peruvian diarrhea virome. Genome Announc. 5, e00822-17 (2017).

18. Abbas, A. A. et al. Redondoviridae, a family of small, circular DNA viruses of the human oro-respiratory tract that are associated with periodontitis and critical illness. Cell Host Microbe 25, 719–729 (2019).

19. Díez-Villaseñor, C. & Rodriguez-Valera, F. CRISPR analysis suggests that small circular single-stranded DNA smacoviruses infect Archaea instead of humans. Nat. Commun. 10, 294 (2019).

20. Feschotte, C. & Gilbert, C. Endogenous viruses: Insights into viral evolution and impact on host biology. Nat. Rev. Genet. 13, 283–296 (2012).

21. Dennis, T. P. W. et al. The evolution, distribution and diversity of endogenous circoviral elements in vertebrate genomes. Virus Res. 262, 15–23 (2019).

22. Bejarano, E. R., Khashoggi, A., Witty, M. & Lichtenstein, C. Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. Proc. Natl. Acad. Sci. 93, 759–764 (1996).

23. Gibbs, M. J., Smeianov, V. V., Steele, J. L., Upcroft, P. & Efimov, B. A. Two families of Rep-like genes that probably originated by interspecies recombination are represented in viral, plasmid, bacterial, and parasitic protozoan genomes. Mol. Biol. Evol. 23, 1097–1100 (2006).

24. Baldauf, S. L. The deep roots of eukaryotes. Science 300, 1703–1706 (2003).

25. Kazlauskas, D., Varsani, A., Koonin, E. V. & Krupovic, M. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. Nat. Commun. 10, 3425 (2019).

26. Siqueira, J. D. et al. Complex virome in feces from Amerindian children in isolated Amazonian villages. Nat. Commun. 9, 4270 (2018).

27. Shan, T. et al. The fecal virome of pigs on a high-density farm. J. Virol. 85, 11697–11708 (2011).

28. Beres Castrignano, S. et al. Identification of circo-like virus-Brazil genomic sequences in raw sewage from the metropolitan area of São Paulo: evidence of circulation two and three years after the first detection. Mem Inst Oswaldo Cruz, Rio Janeiro 112, 175–181 (2017).

29. Krupovic, M. & Forterre, P. Single-stranded DNA viruses employ a variety of mechanisms for integration into host genomes. Ann. N. Y. Acad. Sci. 1341, 41–53 (2015).

30. Pollo, S. M. J. et al. Benchmarking hybrid assemblies of Giardia and prediction of widespread intra-isolate structural variation. Parasites and Vectors 13, 108 (2020).

31. Zhang, H., Alramini, H., Tran, V. & Singh, U. Nucleus-localized antisense small RNAs with 5'-polyphosphate termini regulate long term transcriptional gene silencing in Entamoeba histolytica G3 strain. J. Biol. Chem. 286, 44467–44479 (2011).

32. Zhang, H., Ehrenkaufer, G. M., Hall, N. & Singh, U. Small RNA pyrosequencing in the protozoan parasite Entamoeba histolytica reveals strain-specific small RNAs that target virulence genes. BMC Genomics 14, 53 (2013).

33. Liu, H. et al. Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. BMC Evol. Biol. 11, 276 (2011).

34. Cardinale, D. J. & Duffy, S. Single-stranded genomic architecture constrains optimal codon usage. Bacteriophage 1, 219–224 (2011).

35. Ikemura, T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. J. Mol. Biol. 151, 389–409 (1981).

36. Martin, D. P. et al. Recombination in eukaryotic single stranded DNA viruses. Viruses 3, 1699–1738 (2011).

37. Diamond, L. S., Mattern, C. F. & Bartgis, I. L. Viruses of Entamoeba histolytica I. Identification of transmissible virus-like agents. J. Virol. 9, 326–341 (1972).

38. Wang, A. L. & Wang, C. C. Discovery of a specific double-stranded RNA virus in Giardia lamblia. Mol. Biochem. Parasitol. 21, 269–76 (1986).

39. Kirk, M. D. et al. World Health Organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: A data synthesis. PLoS Med. 12, e1001921 (2015).

40. Sanchez, J. L., Rios, C., Hernandez-Fragoso, I. & Ho, C. K. Parasitological evaluation of a foodhandler population cohort in Panama: Risk factors for intestinal parasitism. Mil. Med. 155, 250–255 (1990).

41. Haque, R. et al. Epidemiologic and clinical characteristics of acute diarrhea with emphasis on Entamoeba histolytica infections in preschool children in an urban slum of Dhaka, Bangladesh. Am. J. Trop. Med. Hyg. 69, 398–405 (2003).

42. Burgess, S. L. & Petri, W. A. The intestinal bacterial microbiome and E. histolytica infection. Curr. Trop. Med. Reports 3, 71–74 (2016).

43. Hartley, M. A., Ronet, C., Zangger, H., Beverley, S. M. & Fasel, N. Leishmania RNA virus: when the host pays the toll. Frontiers in Cellular and Infection Microbiology 2, 99 (2012).

44. Dheilly, N. M. et al. Parasite microbiome project: Grand challenges. PLoS Pathog. 15, e1008028 (2019).

45. Babayan, S. A., Orton, R. J. & Streicker, D. G. Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. Science 362, 577–580 (2018).

46. Van der Hoek, L. et al. Genetic differences between human immunodeficiency virus type 1 subpopulations in faeces and serum. J. Gen. Virol. 79, 259–267 (1998).

47. Oude Munnink, B. B. et al. Unexplained diarrhoea in HIV-1 infected individuals. BMC Infect. Dis. 14, 22 (2014).

48. van Bilsen, W. P. H. et al. Diverging trends in incidence of HIV versus other sexually transmitted infections in HIV-negative MSM in Amsterdam. AIDS 34, 301–309 (2020).

49. Boom, R. et al. Rapid and simple method for purification of nucleic acids. J. Clin. Microbiol. 28, 495–503 (1990).

50. Endoh, D. et al. Species-independent detection of RNA virus by representational difference analysis using non-ribosomal hexanucleotides for reverse transcription. Nucleic Acids Res. 33, e65 (2005).

51. Edridge, A. W. D. et al. Novel orthobunyavirus identified in the cerebrospinal fluid of a Ugandan child with severe encephalopathy. Clin. Infect. Dis. 68, 139–142 (2018).

52. Cotten, M. et al. Full genome virus detection in fecal samples using sensitive nucleic acid preparation, deep sequencing, and a novel iterative sequence classification algorithm. PLoS One 9, e93269 (2014).

53. Kinsella, C. M., Deijs, M. & van der Hoek, L. Enhanced bioinformatic profiling of VIDISCA libraries for virus detection and discovery. Virus Res. 263, 21–26 (2019).

54. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28, 3211–3217 (2012).

55. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152 (2012).

56. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461 (2010).

57. Camacho, C. et al. BLAST+: Architecture and applications. BMC Bioinformatics 10, 421 (2009).

58. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 12, 385 (2011).

59. Bankevich, A. et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477 (2012).

60. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN] (2013).

61. Hong, C. et al. PathoScope 2.0: A complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome 2, 33 (2014).

62. Potter, S. C. et al. HMMER web server: 2018 update. Nucleic Acids Res. 46, W200–W204 (2018).

63. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31, 3406–3415 (2003).

64. Mihara, T. et al. Linking virus genomes with host taxonomy. Viruses 8, 66 (2016).

65. Hung, C. C., Chang, S. Y. & Ji, D. Der. Entamoeba histolytica infection in men who have sex with men. The Lancet Infectious Diseases 12, 729–736 (2012).

66. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. 20, 1160–1166 (2017).

67. Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: A genome comparison visualizer. Bioinformatics 27, 1009–1010 (2011).

68. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).

69. Krzywinski, M. et al. Circos: An information aesthetic for comparative genomics. Genome Res. 19, 1639–1645 (2009).

70. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973 (2009).

71. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313 (2014).

72. Ng, T. F. F. et al. High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. J. Virol. 86, 12161–12175 (2012).

73. Cantalupo, P. G. et al. Raw sewage harbors diverse viral populations. MBio 2, e00180-11 (2011).

74. Pearson, V. M., Caudle, S. B. & Rokyta, D. R. Viral recombination blurs taxonomic lines: Examination of single-stranded DNA viruses in a wastewater treatment plant. PeerJ 4, e2585 (2016).

75. Brinkman, N. E., Villegas, E. N., Garland, J. L. & Keely, S. P. Reducing inherent biases introduced during DNA viral metagenome analyses of municipal wastewater. PLoS One 13, e0195350 (2018).

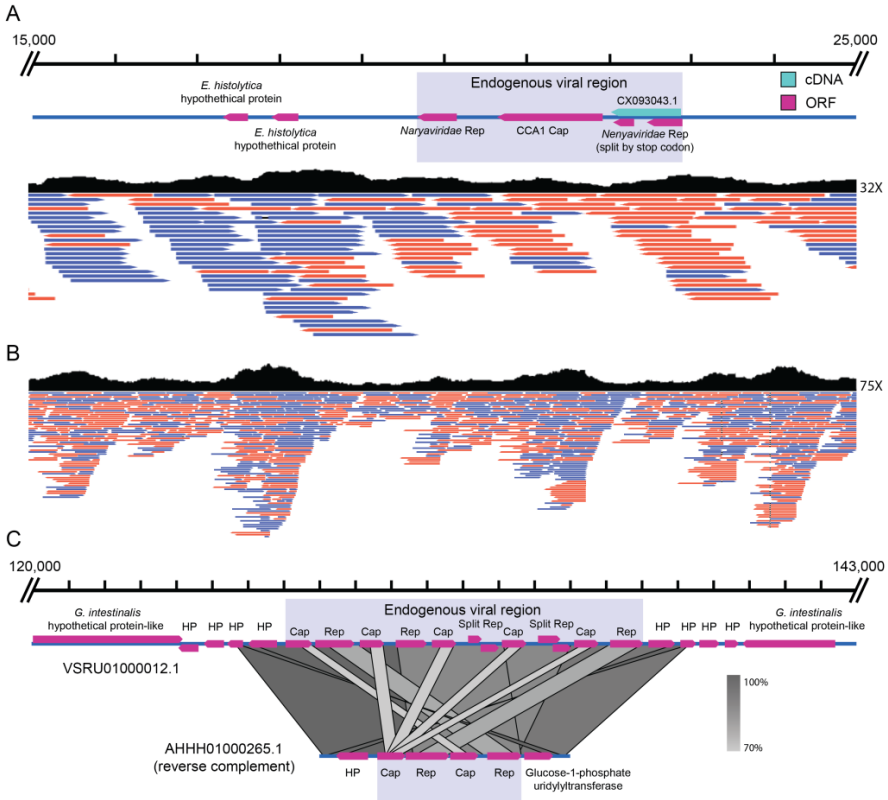
76. Chu, B. T. T. et al. Metagenomics reveals the impact of wastewater treatment plants on the dispersal of microorganisms and genes in aquatic sediments. Appl. Environ. Microbiol. 84, e02168-17 (2018).

77. Monaco, C. L. et al. Altered virome and bacterial microbiome in human immunodeficiency virus-associated acquired immunodeficiency syndrome. Cell Host Microbe 19, 311–322 (2016).

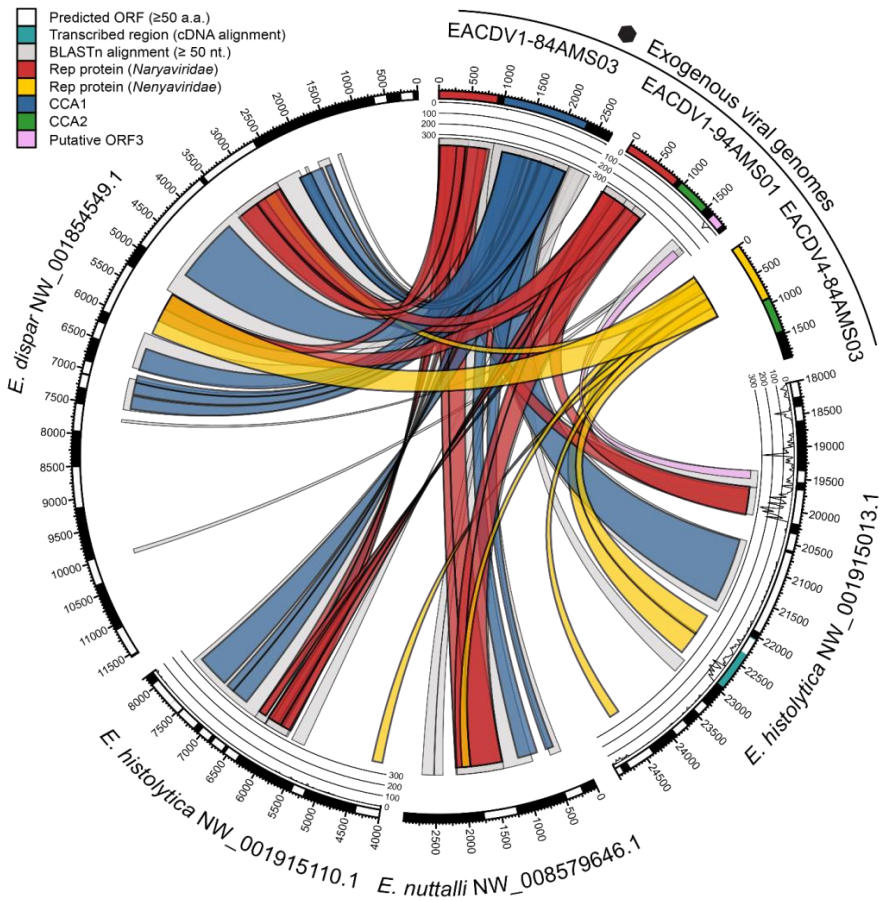
78. Takano, T., Yanai, Y., Hiramatsu, K., Doki, T. & Hohdatsu, T. Novel single-stranded, circular DNA virus identified in cats in Japan. Arch. Virol. 163, 3389–3393 (2018).

Supplementary figures

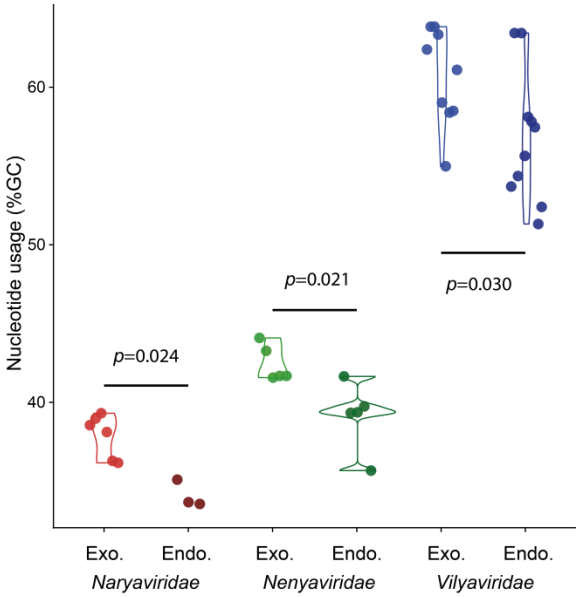
For supplementary tables and references, see the online version: https://doi.org/10.1038/s41467-020-18474-w.



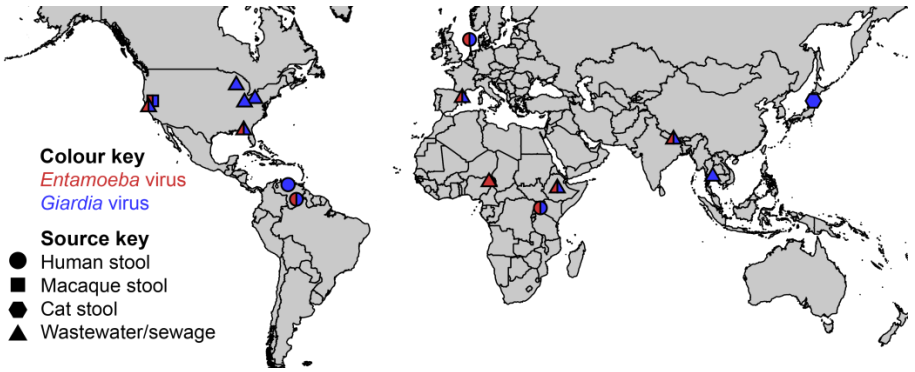
Supplementary Figure 1. The presence of endogenous viral elements is supported by raw-read coverage or long-reads. (A) Sanger sequencing reads from *Entamoeba histolytica* isolate HM1:IMSS (isolated in Mexico, 1967) aligned to genomic contig NW_001915013.1, revealing coverage spanning both the endogenous virus element junctions and flanking host sequence. Maximum coverage depth is shown to the right of the coverage plot. (B) Combined 454 and Illumina sequencing reads from *E. histolytica* isolate KU27 (isolated in Japan, 2001) aligned to contig NW_001915013.1, confirming the element is shared among independent isolates. (C) Long-read technology assists in resolution of repetitive genomic features, as shown by *Giardia duodenalis* genomic contig VSRU01000012.1, which contains a 10 kb region of tandemly repeated integrated viral genomes belonging to the *Vilyaviridae*. This contig belongs to a hybrid assembly built from nanopore long-reads and conventional short-reads[1] (GCA_011634595.1). Host hypothetical proteins are annotated as 'HP'. BLASTn alignment with the short-read based contig AHHH01000265.1 reveals that the two share both the viral integration and neighbouring host sequences; however the latter contig has a shorter endogenous viral region. This is likely a result of assembly software being unable to distinguish between tandem repeat units, leading to assembly collapse. Regions of BLASTn alignment are shown by the grey blocks, with shade corresponding to % identity.



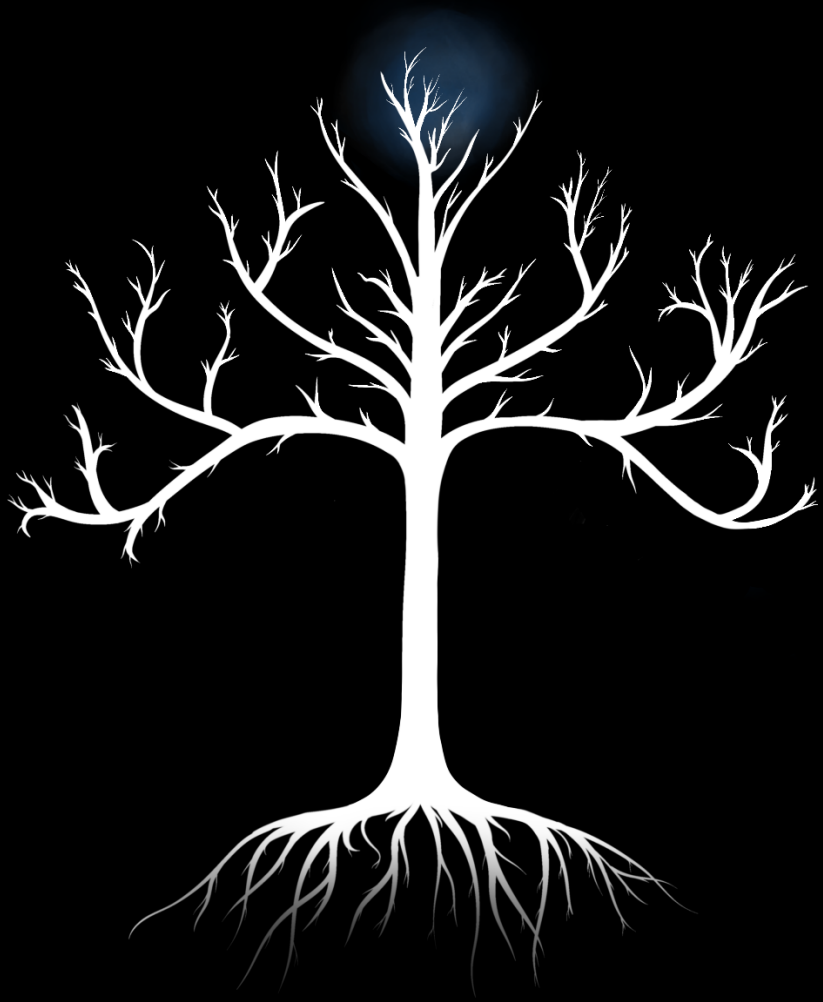
Supplementary Figure 2. Virus genome integrations within *Entamoeba* genomes, and subsequent small RNA control. Alignments were generated using BLASTn (grey bands) and tBLASTn (coloured bands) between three viral genomes (top right, EACDV1-84AMS03 = MT293413, EACDV1-94AMS01 = MT293412, and EACDV4-84AMS03 = MT293420) and contigs from three species of *Entamoeba*. Regions of alignment were visualised with Circos[2]. Open reading frames on parasite contigs were coloured white, while those of exogenous viruses were coloured according to the legend. Alignments reveal a whole viral genome integration (between nucleotides 2,600 and 5,300 of *E. dispar* contig NW_001854549.1), and also multiple unrelated integrations by *Naryaviridae* and *Nenyaviridae* in close physical proximity to each other. Notably, this latter observation suggests CRESS integration is sometimes site-specific, and this is most likely mediated by Rep recognition of previously integrated nonanucleotide origin motifs within the host genome, a model discussed elsewhere[3]. To add support that endogenous viral elements are real features of parasite genomes, rather than artefacts in genome assemblies, *E. histolytica* contigs are annotated with AGO2-2 associated small RNA coverage[4]. Data from other *Entamoeba* species was not available. Coverage peaks (inner ring) often occur across open reading frames derived from virus integrations, suggesting these are real genomic features. Some small RNAs also exactly match exogenous viral genomes, implying a possible indirect role in antiviral defence. CCA = CRESS virus *Cap* assemblage.



Supplementary Figure 3. GC-content of exogenous and endogenous viruses. Points denote GC-content of a viral *Rep* gene sequence, either from the genome of an exogenous virus (Exo.) or extracted from an endogenous viral element within a host genome (Endo.). For each viral family, Mann-Whitney U tests were performed to evaluate whether exogenous and endogenous sequences had significantly different GC-content than expected by chance (*Naryaviridae* n=9, *Nenyaviridae* n=10, *Vilyaviridae* n=19). In each case GC-content was found to be different; exogenous virus genomes have a significantly higher GC-content than related endogenous viral elements.



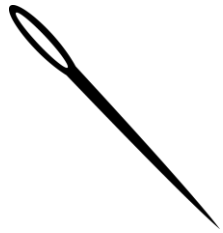
Supplementary Figure 4. Geographical origin of public viral genomes or metagenome samples containing parasite-infecting CRESS viruses. Shape of symbols denotes the sample source type, while colour denotes whether the identified viruses were *Entamoeba*-infecting (*Naryaviridae* and *Nenyaviridae*), *Giardia*-infecting (*Vilyaviridae*), or both (bicolour symbols). The map was adapted from world border templates available from http://thematicmapping.org/downloads/world_borders.php, provided by Bjørn Sandvik under a Creative Commons Attribution-ShareAlike 3.0 license (https://creativecommons.org/licenses/by-sa/3.0/legalcode).



Chapter 4

Host prediction for disease-associated gastrointestinal cressdnaviruses

Cormac M. Kinsella, Martin Deijs, Christin Becker, Patricia Broekhuizen, Tom van Gool, Aldert Bart, Arne S. Schaefer, Lia van der Hoek



Virus Evolution, 2022

https://doi.org/10.1093/ve/veac087

Abstract

Metagenomic techniques have facilitated the discovery of thousands of viruses, yet because samples are often highly biodiverse, fundamental data on the specific cellular hosts are usually missing. Numerous gastrointestinal viruses linked to human or animal diseases are affected by this, preventing research into their medical or veterinary importance. Here, we developed a computational workflow for the prediction of viral hosts from complex metagenomic datasets. We applied it to seven lineages of gastrointestinal cressdnaviruses using 1,124 metagenomic datasets, predicting hosts of four lineages. The *Redondoviridae*, strongly associated to human gum disease (periodontitis), were predicted to infect *Entamoeba gingivalis*, an oral pathogen itself involved in periodontitis. The *Kirkoviridae*, originally linked to fatal equine disease, were predicted to infect a variety of parabasalid protists, including *Dientamoeba fragilis* in humans. Two viral lineages observed in human diarrhoeal disease (CRESSV1 and CRESSV19, i.e. pecoviruses and hudisaviruses) were predicted to infect *Blastocystis* spp. and *Endolimax nana* respectively, protists responsible for millions of annual human infections. Our prediction approach is adaptable to any virus lineage and requires neither training datasets nor host genome assemblies. Two host predictions (for the *Kirkoviridae* and CRESSV1 lineages) could be independently confirmed as virus–host relationships using endogenous viral elements identified inside host genomes, while a further prediction (for the *Redondoviridae*) was strongly supported as a virus–host relationship using a case–control screening experiment of human oral plaques.

Introduction

A defining feature of viruses is their obligate relationship with hosts, yet surprisingly hosts of most newly identified viruses remain unknown (Simmonds et al. 2017; Dolja and Koonin 2018; Greninger 2018). This circumstance is driven by widespread use of high-throughput sequencing for the discovery of viral genomes (Shi et al. 2016; Tisza et al. 2020; Edgar et al. 2022) versus traditional techniques, such as viral isolation in cell culture. In particular, metagenomic sequencing of taxonomically diverse samples obscures virus–host relationships, because of the many potential pairings. Exemplifying this are the cressdnaviruses, a group with small circular ssDNA genomes encoding a replication-associated protein. Now classified under the phylum *Cressdnaviricota* (Krupovic et al. 2020), the vast majority have unknown hosts (Simmonds et al. 2017; Tisza et al. 2020). This even applies to notable disease-associated lineages identified frequently in the gastrointestinal tracts of humans and other animals, referred to hereafter as gastrointestinal cressdnaviruses (Li et al. 2015; Phan et al. 2016; Abbas et al. 2019; Ramos et al. 2021). Among these are the family *Redondoviridae*, residents of the human mouth and lung linked to both periodontitis and critical illness (Abbas et al. 2019; Zhang et al. 2021), and the *Kirkoviridae*, found variously in dead and diseased horses, cows, and pigs, and also in

human stool (Shan et al. 2011; Li et al. 2015; Zhao et al. 2017; Guo et al. 2018; Xie et al. 2020). Because infectious gastrointestinal disease is a leading cause of global mortality and morbidity in humans and livestock (Tam et al. 2012; Kirk et al. 2015; Thumbi et al. 2015), there is a clear need to determine the hosts of gastrointestinal cressdnaviruses, data that will underpin their medical or veterinary relevance.

Historically, no host inference methodology was required for cressdnaviruses, since observation of host disease preceded discovery of the responsible virus. For example, banana bunchy top disease was recognised from approximately 1880 and classified as viral in the 1920s (Magee 1927), before the responsible cressdnavirus of family *Nanoviridae* was characterised later in the 20th century (Harding et al. 1993). Similarly, plant diseases have been linked to the *Geminiviridae* (Varma and Malathi 2003), avian and porcine diseases to the *Circoviridae* (Ritchie et al. 1989; Ellis et al. 1998), fungal debilitation to the *Genomoviridae* (Yu et al. 2010), and diatom lysis to the *Bacilladnaviridae* (Nagasaki et al. 2005). The challenge of the metagenomic age will be the identification of hosts when only the viral genome is known. While promising wet-lab methods, such as single-cell sequencing (Yoon et al. 2011) or proximity ligation (Bickhart et al. 2019; Ignacio-Espinoza et al. 2020) will enable simultaneous virus discovery and linkage to host sequences in future, these techniques are still emerging in the viral metagenomics field, and offer no solution for the thousands of conventionally sequenced viruses. Here, high-throughput computational approaches are required. Indirect solutions such as genome compositional analyses (Kapoor et al. 2010) or machine learning have been suggested; however, the former requires host genome assemblies or validated training datasets and suffers from relatively low accuracy (Ahlgren et al. 2017; Liu et al. 2019), while the latter generally requires validated training datasets, making it most appropriate for host prediction within otherwise well-characterised lineages (Eng, Tong, and Tan 2014; Babayan, Orton, and Streicker 2018).

Viral fossils inside host genomes, such as CRISPR spacers or endogenous viral elements (EVEs), provide direct evidence of virus–host relationships (Liu et al. 2011; Dion et al. 2021; Zhao, Lavington, and Duffy 2021). Among cressdnaviruses, the *Smacoviridae* have been proposed to infect archaea on the basis of matched CRISPR spacers (Díez-Villaseñor and Rodriguez-Valera 2019), while three families (*Naryaviridae*, *Nenyaviridae*, and *Vilyaviridae*) were linked to gut parasites using EVE evidence (Kinsella et al. 2020). Again, however, availability of host genome assemblies limits the range of this approach, and many virus–host relationships are likely unrepresented in the genomic fossil record. To date, no EVEs have been found belonging to the aforementioned redondoviruses or kirkoviruses. For such groups, high-throughput host prediction approaches that do not rely on host assemblies are needed. Here, we developed an analysis workflow for host prediction from metagenomic sequencing datasets, aiming to identify over-represented eukaryotes among virus positive samples for subsequent investigation. Through the analysis of 1,124 gastrointestinal tract samples, we could identify multiple cressdnavirus–eukaryote associations. Host predictions included redondoviruses with the human oral

parasite *Entamoeba gingivalis*, kirkoviruses with parabasalid protists including *Dientamoeba fragilis* in humans, the CRESSV1 lineage (i.e. pecoviruses) with *Blastocystis* spp., and the CRESSV19 lineage (i.e. hudisaviruses) with *Endolimax nana*. Subsequent independent analysis confirmed several of these predictions as virus–host relationships.

Results

Census of gastrointestinal cressdnavirus lineages

Here, we aimed to predict the host of any cressdnavirus lineage displaying an apparently obligate association to the gastrointestinal tracts of vertebrates. Because no study has so far focused collectively on gastrointestinal cressdnaviruses, we first comprehensively censused published cressdnavirus sequences to determine the lineages meeting that definition. Iterative searches of the GenBank protein database collected 15,815 unique cressdnavirus Rep sequences, 2,461 of which remained after clustering. Each taxonomic class (*Arfiviricetes* and *Repensiviricetes*) was phylogenetically analysed separately (1,850 and 611 sequences, respectively). To work with unclassified lineages, clusters of related sequences were assigned a temporary name according to their branch support. We followed the format introduced by Kazlauskas, Varsani, and Krupovic (2018) who named the unclassified lineages CRESSV1 to CRESSV6. We added CRESSV7 to CRESSV33 in the *Arfiviricetes* (Fig. 1) and CRESSV34 to CRESSV39 in the *Repensiviricetes* (Supplementary Fig. S1). All previously named families and lineages were supported by our analysis, with the exception of CRESSV2, whose members remained adjacent but with poor branch support (Supplementary Fig. S2). We suggest that CRESSV2 may be most accurately characterised as multiple distinct lineages, here denoted as CRESSV2.1 to CRESSV2.6. Supporting this, the resulting sublineages showed unique isolation source patterns; for example, most members of CRESSV2.2 came from marine animal tissues and seawater, CRESSV2.3 were found predominately in human or livestock stool and tissue, and CRESSV2.4 members were identified in spiders, insects, and bird anal swabs (Supplementary Table S1).

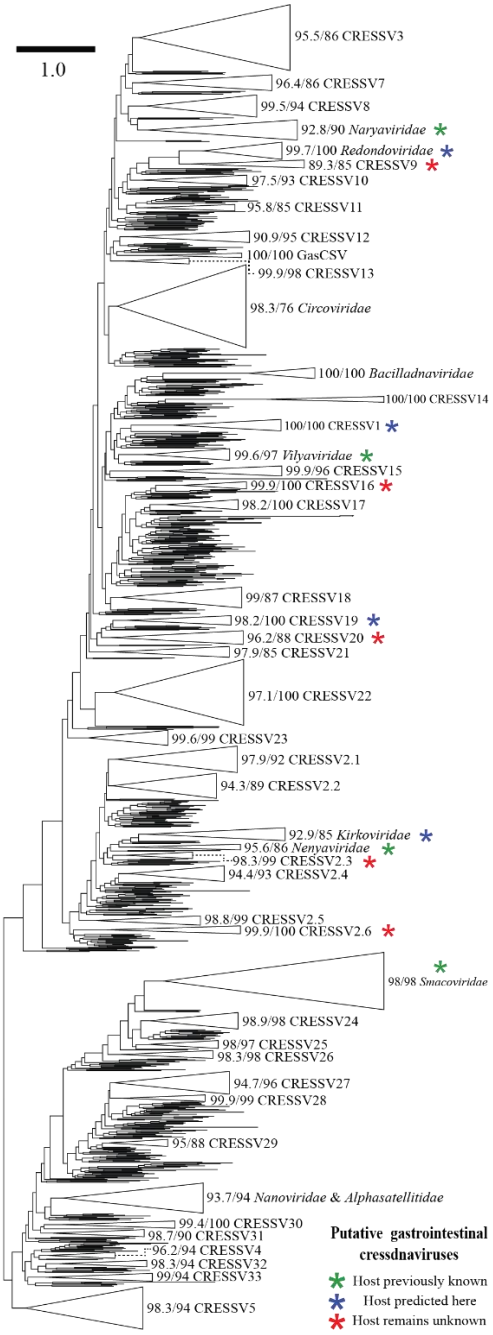


Figure 1. Maximum likelihood phylogenetic tree of the *Arfiviricetes*, rooted at the midpoint. Scale bar denotes amino acid substitutions per site. Branch supports are given for each named lineage, with SH-aLRT scores on the left and ultrafast bootstrap scores on the right. All sequences found outside of collapsed nodes did not meet criteria for naming a lineage.

Of fifty-six named lineages across the *Cressdnaviricota*, we categorised thirteen as putatively gastrointestinal due to their isolation source patterns (see Materials and methods). All were in the *Arfiviricetes* (Fig. 1, Supplementary Table S1). Four of these were excluded immediately because host inferences were already published; these were the *Smacoviridae*, *Naryaviridae, Nenyaviridae*, and *Vilyaviridae* (Díez-Villaseñor and Rodriguez-Valera 2019; Kinsella et al. 2020). Seven of the nine remaining lineages were found mainly in oral, gastrointestinal, or stool samples of various vertebrates, and some wastewater samples. These were the *Redondoviridae, Kirkoviridae*, CRESSV1 (i.e. pecoviruses), CRESSV2.3, CRESSV2.6, CRESSV9, and CRESSV19 (i.e. hudisaviruses). The others (CRESSV16 and CRESSV20) were detected predominately in wastewater, and were included since this source is often stool contaminated. The retained lineages were widely distributed phylogenetically, although notably some neighboured each other. For example, the lineage CRESSV9 was a close relative of the *Redondoviridae*, and together they were related to the *Naryaviridae*, viruses of *Entamoeba* parasites. Meanwhile, CRESSV19 and CRESSV20 clustered together, CRESSV1 was related to the *Giardia*-infecting *Vilyaviridae* and *Kirkoviridae* was related to both the lineage CRESSV2.3 and the *Nenyaviridae*, the latter also infecting *Entamoeba*.

Recombination events and viral distributions reveal host biases

We identified nine gastrointestinal cressdnavirus lineages with unknown hosts. Some lineages were found in multiple vertebrate taxa, for example, CRESSV1 was known from stools of humans, pigs, and a camel, amongst others. This raised the possibility of different host preferences within a given lineage. Because this scenario would affect downstream analysis, we looked for viral recombination within lineages, which serves as evidence of shared host ranges since recombination must occur in the same host cell (Duffy, Burch, and Turner 2007; Kinsella et al. 2020). We first explored the genomic patterns of cressdnavirus recombination, analysing phylogenetic compatibility between nucleotide alignment windows along all available redondovirus genomes. This showed that highest incompatibility is found between genes, not within them (Fig. 2A), suggesting modular recombination of complete genes with different evolutionary histories. This pattern was corroborated by analysing the distribution of breakpoint pair coordinates, showing relative enrichment in two coordinate regions, those linking the start and end of the *Rep* gene, and those linking the start and end of the *Cap* gene (Fig. 2A). This propensity to swap genes as complete units is likely due to a reduced risk of protein structure disruption when compared with intra-gene recombination (Lefeuvre et al. 2007, 2009). We built on the observation by constructing tanglegrams between Rep and Cap protein phylogenies for each lineage (Fig. 2B–F, Supplementary Fig. S3). These provided some insight, for example, the extensive modular recombination among human-associated redondoviruses strongly suggests they share one host (Fig. 2B).

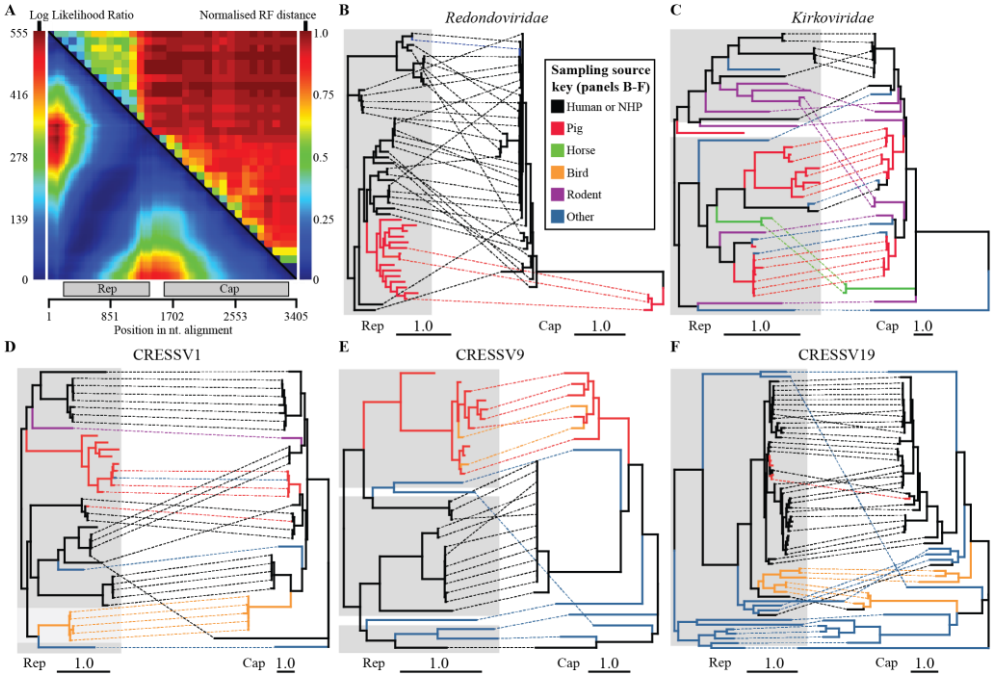


Figure 2. Recombination within gastrointestinal cressdnavirus lineages. (A) Upper right: phylogenetic compatibility matrix (Robinson-Foulds distance) computed on an alignment of redondovirus genomes, lower left: LARD breakpoint matrix computed on the same alignment. (B–F) Rep and Cap protein tanglegrams for five cressdnavirus lineages. Dotted lines connect proteins encoded by the same genome. Branch colour denotes isolation source as listed in the key. Grey blocks denote groups linked by RDP4 detected recombination events, and different shades represent different recombination groups (Panel D only). Scale bars on individual phylograms are in amino acid substitutions per site. NHP: non-human primate.

We annotated tanglegrams with reported isolation sources, finding that related viruses (sublineages) often shared sources. For example, pig-associated sublineages were observed in the *Redondoviridae* (Fig. 2B), the *Kirkoviridae* (Fig. 2C), CRESSV1 (Fig. 2D), and CRESSV9 (Fig. 2E). We hypothesised that such source biases might reflect varying host tropism, and to clarify this we used RDP4 to identify further recombination events within lineages. Interestingly, while most detected events occurred within sublineages, some gene flow was found between them (Fig. 2B–F). Overall, this suggested that members of each lineage overlapped in host range, yet displayed some specialisation at the sublineage level, perhaps to different host subtypes or species. An exception was a ‘reproductively isolated’ CRESSV1 sublineage found in birds (Fig. 2D), that displayed no evidence of recombination outside itself. To explicitly visualise source biases between human and porcine samples, we mapped the distribution of gastrointestinal cressdnaviruses across

seven cohorts comprising 1,124 metagenomic sequencing datasets (Supplementary Tables S2 and S3). These were generated from human stool ($N = 374$), pig stool ($N = 512$), and human oral samples ($N = 238$). The analysis confirmed strongly biased distributions for some viruses, for example, members of CRESSV9 and *Kirkoviridae* were either strictly pig-associated or strictly human-associated across cohorts (Fig. 3). It also showed more flexible viruses, for example, some members of CRESSV1. Consistent with previous literature, we found that human-associated redondoviruses were the only lineage prevalent in the human oral environment, with more sporadic detection in stool (Abbas et al. 2019). Strikingly, previously unrecognised pig-associated redondoviruses (MT135242.1, KJ433989.1, and NC_035476.1) were entirely absent from human oral samples, but highly prevalent in porcine stool. The analysis also revealed that CRESSV16 (previously included for its occurrence in wastewater) was not found in any sample, leading to its exclusion from further analyses. From these analyses, we concluded that members of a viral lineage found in one isolation source (e.g. pig stool) likely shared the same host.

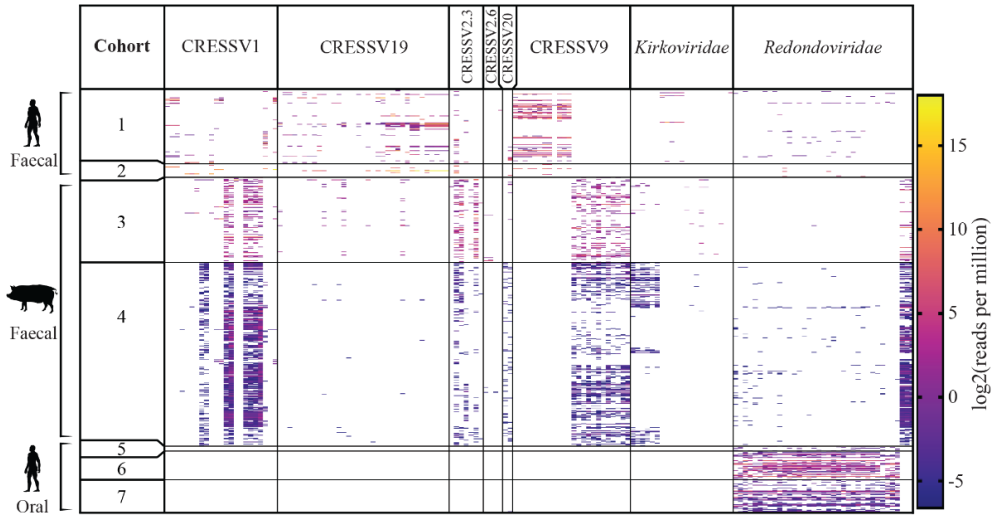


Figure 3. Distribution of gastrointestinal cressdnaviruses across seven sample cohorts. Colour represents normalised read count. Empty columns (viruses not found in any sample) and rows (samples containing no viruses) were removed prior to plotting. Members of the CRESSV16 lineage were not detected. Taxon silhouettes are from phylopic.org (*Homo sapiens* by T. Michael Keesey, *Sus scrofa* by Steven Traver). Sample cohorts and viral reference genomes used are reported in Supplementary Tables S2 and S3.

Viral host prediction

To identify potential hosts of gastrointestinal cressdnaviruses, eukaryotic rRNA content of all 1,124 samples was classified, resulting in taxon lists at the genus level. Individually, for the six cohorts using Illumina deep sequencing, samples highly positive for each virus

lineage were identified and compared to pinpoint prevalent eukaryotic taxa. Thus, shortlists of theoretically possible host candidates were generated for each virus lineage/cohort intersection (Supplementary Table S4). The low number of samples positive for lineage CRESSV2.6 in any cohort excluded it from the analysis at this point, leaving seven lineages (although human-associated and pig-associated redondoviruses were analysed separately). Next, host predictions were made by assessing the statistical associations between viruses and respective host candidates across all samples of all seven cohorts. Human oral cohorts contained only one cressdnavirus lineage, human-associated redondoviruses, and the genus *Entamoeba* was the only host candidate identified. Upon statistical evaluation with Pearson's chi-squared tests, we found that the presence of *Entamoeba* was highly positively associated with the presence of redondoviruses in all three oral cohorts (Supplementary Table S5). Specifically, redondovirus prevalence in subsets of samples positive for *Entamoeba* were 73 per cent, 91 per cent, and 91 per cent, versus 0 per cent, 20 per cent, and 22 per cent in subsets where *Entamoeba* was undetected. In these latter samples, we suspect that if virus was found, non-detection of *Entamoeba* most likely constitutes a false negative. We also found that normalised redondovirus loads were strongly positively correlated with *Entamoeba* loads in the three cohorts, with Spearman's rho values between 0.72 and 0.85 ($P < 0.001$, Supplementary Table S6). At this stage, we therefore predicted *Entamoeba* was the host of redondoviruses. *E. gingivalis* is the only known member of this genus residing in the oral cavity, and examination of BLASTn tables confirmed it was the species identified.

In the two cohorts of human stool samples, presence of the gut protist *Blastocystis* was associated positively with the presence of the CRESSV1 lineage (Supplementary Table S5), with 24 per cent and 9 per cent prevalence in protist positive samples, versus 0 per cent and 1 per cent in negative. Further, CRESSV1 virus loads were positively correlated with *Blastocystis* loads (Supplementary Table S6). The same pattern was observed in both pig stool cohorts; however, in these cases, *Entamoeba* was also associated. While this introduced some uncertainty for host prediction, we noted that prevalences of both protists were extremely high in porcine cohorts, with *Blastocystis* at 76 per cent and 100 per cent prevalence, and *Entamoeba* at 61 per cent and >99 per cent. Normalised loads of both protists were also tightly correlated with each other (cohort 1: rho = 0.72, $P < 0.001$, cohort 2: rho = 0.54, $P < 0.001$), probably due to the shared faecal–oral route of infection, and host factors such as age and health. We suspected the association between CRESSV1 and *Entamoeba* could be driven by this underlying correlation, and we therefore predicted *Blastocystis* (*Blastocystis* spp.) was the likeliest host of CRESSV1, since it was identified and found to be associated in all four stool cohorts, human, and porcine.

Presence of the CRESSV19 lineage was highly positively associated with the presence of *Endolimax* in both human cohorts, and likewise their normalised loads were significantly positively correlated (Supplementary Tables S5 and S6). Importantly, this result was mirrored in both porcine cohorts. Additional associations were found in one of the porcine cohorts, but not both, leading us to predict *E. nana* was the most likely host of CRESSV19.

The result for the kirkoviruses was complex. In both human stool cohorts, presence of kirkoviruses was highly positively associated with the presence of *Dientamoeba*, and likewise their normalised loads were strongly positively correlated. We were therefore surprised when no parabasalid taxa were identified as kirkovirus host candidates in either porcine cohort. Instead, both porcine cohorts showed positive associations to the non-parabasalid genus *Iodamoeba*, in both presence and normalised load. Because evidence from recombination had suggested kirkoviruses at least partly overlapped in host range, we surmised one of the relationships might be incidental. To explore this, we first tested the statistical association between kirkoviruses and *Iodamoeba* in human cohorts, but found none. In the opposite direction, while *Dientamoeba* has been reported in pigs (Cacciò et al. 2012), we found no *Dientamoeba* reads in either porcine cohort. We therefore looked for the presence of other parabasalid taxa. Interestingly, porcine samples highly positive for kirkoviruses did contain a diverse community of parabasalids at high prevalence, including *Trichomitus*, *Tetratrichomonas*, *Hypotrichomonas*, *Trichomonas*, and *Tritrichomonas*. Taken together, at least one parabasalid genus was detected in 10 of 11 samples highly positive for kirkoviruses in cohort 1 and 61 of 62 samples in cohort 2. Since we had previously assumed viruses infected a single genus per cohort type, our host candidate discovery approach would have missed a broader host range. Upon statistical testing, we found significant positive associations between several of the parabasalid genera and kirkoviruses (Supplementary Tables S5 and S6). Despite the lack of clarity in porcine cohorts, due to the strong signal from *Dientamoeba* in human cohorts, we tentatively predicted parabasalids serve as the hosts of kirkoviruses, specifically *D. fragilis* in humans.

Our analyses of CRESSV2.3, CRESSV9, CRESSV20, and pig-associated redondoviruses did not result in host prediction. In the first case, no candidate host taxon was linked to virus presence in human cohorts, although *Iodamoeba* was associated in both porcine cohorts. Testing this genus in human cohorts found no association. Given the high prevalence of parasite infection in porcine cohorts we regarded this as insufficient evidence to predict a host. For both CRESSV9 and CRESSV20, no taxon was identified to be consistently associated with viruses across human and porcine cohorts. In the case of pig-associated redondoviruses, a large set of genera were associated to virus presence in pig stool cohort 1, two of which were also associated in cohort 2 (*Balantioides* and *Balantidium*). Due to the previously mentioned complication of high protist prevalence in porcine samples, we did not make a host prediction.

Confirmation of host–virus relationships

Our computational workflow predicted protist hosts for four viral lineages: *E. gingivalis* for human-associated redondoviruses, *Blastocystis* spp. for CRESSV1, *E. nana* for CRESSV19, and diverse parabasalid genera for kirkoviruses (specifically *D. fragilis* in humans, and a range of genera in pigs). To independently assess the inferred host–virus relationships, we looked for related EVEs in available protist genome assemblies. No assembly was available for *E. gingivalis*, *E. nana*, or *D. fragilis*, but we included close

relatives, and ten *Blastocystis* spp. assemblies (Supplementary Table S7). Notably, four Rep-like EVEs were previously identified in *Blastocystis* spp. (Liu et al. 2011). Our analysis identified thirty-eight cressdnavirus-like EVEs in *Blastocystis* spp., including redetection of the original four. EVEs were distributed across six assemblies from *Blastocystis* spp. subtypes 1, 2, 6, 7, 8, and 9. To confirm their presence in the genome as opposed to assembly contamination, we carried out PCR targeting a subset of six EVEs, using DNA extracted from axenic *Blastocystis* spp. cultures of subtypes 1, 2, 7, and 8. In each case, we could amplify products of the correct size, and two were confirmed by Sanger sequencing. Of the four assemblies in which no EVE was identified, two belonged to subtype 3 and two to subtype 4. Among the thirty-eight EVEs, thirty-seven were Rep-like and one was Cap-like. Clustering of the Rep-like sequences alongside the four of Liu et al. (2011) revealed two distinct clusters and one singleton (Fig. 4A). Cluster 1 included twenty-seven EVEs plus the four previously identified, while cluster 2 contained only newly identified sequences. Phylogenetic analysis confirmed that cluster 2 EVEs belonged to the CRESSV1 virus lineage, validating the prediction that CRESSV1 members infect *Blastocystis* spp. (Fig. 4B, Supplementary Fig. S4A).

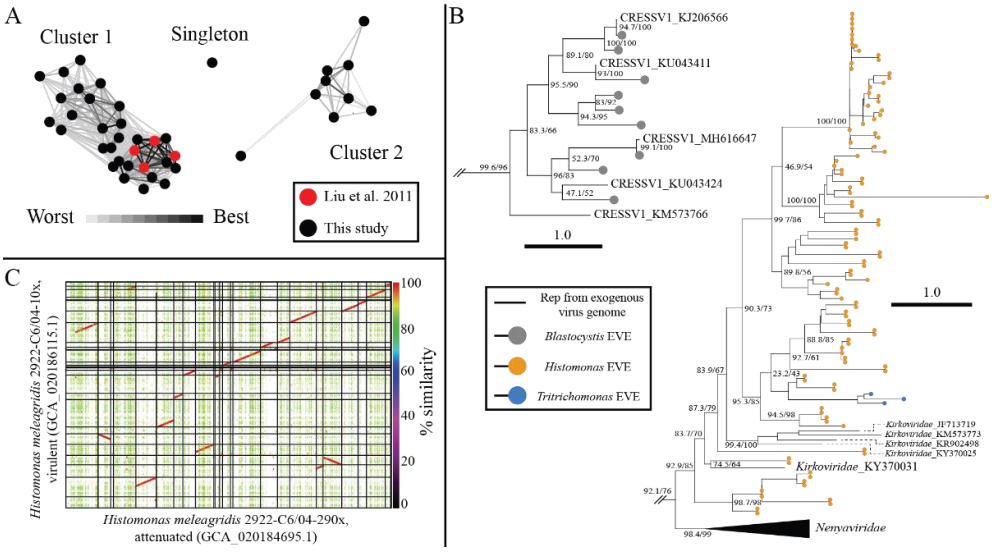


Figure 4. EVEs in protist genomes support host inferences. (A) Clustered Rep-like EVEs from *Blastocystis* spp. assemblies. Connections represent significant BLASTp alignments between EVEs, with shade corresponding to level of significance (maximum/worst e-value = 1e-10). Four EVEs identified by Liu et al. (2011) were clustered alongside all thirty-seven Rep-like EVEs detected here. (B) Regions of interest from a phylogeny of Rep-like EVEs and representatives of cressdnavirus lineages (see also Supplementary Fig. S4). Scale bar represents amino acid substitutions per site. (C) Nucmer alignment dotplot between EVE-containing scaffolds from two *Histomonas meleagridis* genome assemblies. Colour denotes alignment percentage similarity. For the list of aligned scaffolds, see Supplementary Table S8.

Among parabasalids, 145 EVEs were identified in genome assemblies of *Histomonas meleagridis* and 172 were identified in one *Tritrichomonas foetus* assembly. Of the *H. meleagridis* EVEs 104 were Rep-like and forty-one were Cap-like, while *T. foetus* EVEs were all Rep-like. Phylogenetic analysis of Rep-like EVEs revealed 102 *H. meleagridis* sequences and three *T. foetus* sequences belonged to the *Kirkoviridae* (Fig. 4B, Supplementary Fig. S4B). This confirms the prediction that kirkoviruses infect parabasalids, although specific validation for *D. fragilis* is still desirable. Notably, the two *H. meleagridis* assemblies were generated from the same strain, one from a virulent form and the other from an attenuated form. Both were originally cultured from a single micro-manipulated cell, with separate passaging for ten or 290 generations, respectively (Palmieri et al. 2021). We thus predicted that scaffolds containing true EVEs would be homologous between such closely related assemblies. Contrastingly, if the sequences actually derived from assembly contamination and were not shared, we would expect dispersal throughout each assembly, and scaffolds would appear mostly non-homologous. We carried out all-vs.-all alignment between EVE-containing scaffolds from the assemblies, twenty-five for GCA_020184695.1 and twenty-nine for GCA_020186115.1 (Supplementary Table S8). The vast majority of scaffolds were clearly homologous, in line with the expectation for true EVEs (Fig. 4C). Notably, these assemblies were built using Oxford Nanopore Technologies long reads in combination with high accuracy Illumina reads, an approach recognised to result in low misassembly rates and high accuracy assemblies (Wick et al. 2017).

Finally, we assessed our prediction that redondoviruses infect *E. gingivalis*. With no host genome assembly available, we ran a case–control screening experiment on DNA extracted from oral plaques of human subjects with periodontitis ($N = 48$), thirty-one with known *E. gingivalis* infection and seventeen tested negative. Samples were screened using qPCR assays for redondoviruses, *E. gingivalis*, and *Trichomonas tenax. T. tenax* was included because like *E. gingivalis*, it is a protist associated with human periodontitis (Marty et al. 2017; Benabdelkader et al. 2019), and thus represents an appropriate negative control that should have no association to redondoviruses. We found that qPCR detections of redondoviruses and *E. gingivalis* were highly positively associated with each other (Pearson's chi-squared test: $\chi^2 = 36.71$, $P < 0.001$), while results of redondoviruses and *T. tenax* had no association ($\chi^2 = 0.08$, $P = 0.771$). Using linear regression of Ct values, we additionally found that redondovirus loads were positively correlated with *E. gingivalis* loads ($R^2 = 0.24$, $P = 0.013$, Supplementary Fig. S5), but not with *T. tenax* loads ($R^2 = 0.01$, $P = 0.762$). These results lead us to infer that redondoviruses infect *E. gingivalis*, since they are strongly, consistently, and specifically associated.

Discussion

Metagenomics has massively expanded known viral diversity. In recognition of the insurmountable task of characterising 'metagenomic species' using traditional laboratory techniques, official taxonomy can now be applied to virus sequence data, rather than characterised isolates alone (Simmonds et al. 2017). In the metagenomic age, host determination is a comparably large and complex task using traditional techniques, with swathes of eukaryotic and prokaryotic taxa intractable to isolation in culture, which also complicates genome sequencing. Here, we developed a metagenomic analysis approach for host prediction that does not rely on a culture system nor a host genome assembly, improving on our previous method (Kinsella et al. 2020). We applied it to metagenomic sequencing datasets containing seven lineages of gastrointestinal cressdnaviruses, several of which have been linked to human and animal diseases. Host predictions were made for four lineages: human-associated redondoviruses with *E. gingivalis*, kirkoviruses with diverse parabasalid taxa including *D. fragilis* in humans, the CRESSV1 lineage (i.e. pecoviruses) with *Blastocystis* spp., and the CRESSV19 lineage (i.e. hudisaviruses) with *E. nana*. Two of the four predictions (kirkoviruses and lineage CRESSV1) were independently confirmed using EVE evidence, as host genome assemblies were available. For a third prediction (redondoviruses), a case–control experiment was used instead. Our study therefore represents a powerful approach to host identification in the metagenomic age, applicable to any poorly understood virus group found in metagenomic datasets.

Analysis of host presence at the genus level mostly resulted in identification of a single species shared across virus positive samples, yet for kirkovirus hosts in pig stool this resolution was too specific, and was resolved by expanding the taxonomic rank to the Parabasalia. This highlights a complication with utilising taxonomy; equivalent ranks may capture different levels of genetic diversity, and higher ranks may capture the same diversity as lower ones. Illustrating this, the gut-resident amoeba *E. dispar* and *E. histolytica* are closer relatives by rRNA identity than many *Blastocystis* spp. subtypes, and while the former are considered different species, the latter are not (Stensvold et al. 2007). A possible solution for our purpose would be approaching host identity analogously to prokaryotic operational taxonomic units, which apply precise divergence rules to determine taxonomic clusters. Furthermore, while it is broadly true that more closely related viruses are more likely to share hosts, there is no arbitrary genetic divergence cutoff in nature where host switches occur. Purely unsupervised approaches cannot easily address this, and we suggest that the best current solution involves both automated prediction, and expert assessment.

Our findings resolve the possible roles gastrointestinal cressdnaviruses play in human and animal health. Discovered in 2019, the family *Redondoviridae* was found to be strongly associated with human periodontitis and had an observational link to critical illness, but infection of humans has not been demonstrated (Abbas et al. 2019; Zhang et al. 2021). Our finding that the human oral protist *E. gingivalis* is the host of redondoviruses explains their

statistical association to periodontitis, since *E. gingivalis* is also strongly linked to gum disease, possibly causally (Bao et al. 2020, 2021; Badri et al. 2021). It implies that redondoviruses do not cause periodontal disease themselves, although it is unknown if they are commensals, or actively modulate host virulence. Some viruses can cause reduced virulence in their hosts, for example, the genomovirus Sclerotinia sclerotiorum hypovirulence-associated DNA virus 1 (SsHADV-1), which severely impacts its phytopathogenic fungal host *Sclerotinia sclerotiorum*, and may represent a potential biocontrol agent (Yu et al. 2010). Whether redondoviruses represent beneficial (or even potentially therapeutic) viruses remains to be explored. Detection of redondoviruses in respiratory samples (from critically ill patients and others) can be explained either by contamination of samples with oral secretions containing shed virus, or by displacement of oral microbiota and secretions to the lung, a particular problem in critical illness and intubation (Scannapieco 1999; Munro and Grap 2004; Blot, Vandijck, and Labeau 2008). Further, we suggest that the relatively rare gut detections of human-associated redondoviruses must represent swallowed virions rather than a site of viral replication. Notably, the *Redondoviridae* are related to the *Naryaviridae*, a family previously found to infect gut-resident species of *Entamoeba* (Kinsella et al. 2020), adding phylogenetic support to the host inference.

We found that lineages CRESSV1 and CRESSV19 also infected protists (*Blastocystis* spp. and *E. nana*, respectively). Both viral lineages have been observed in cases of human diarrhoeal disease (Phan et al. 2016; Altan et al. 2017; Ramos et al. 2021); however, their role was previously ambiguous. We suggest the viruses do not directly influence human disease, but instead indicate underlying protist infection. Both protists have been linked to diarrhoeal disease previously, yet despite millions of annual infections their pathogenicity remains controversial (Scanlan et al. 2014; Poulsen and Stensvold 2016). Similarly, the finding that kirkoviruses infect parabasalid genera has relevance to both human and veterinary health. Kirkoviruses have been identified in dead and diseased livestock on multiple occasions (Li et al. 2015; Guo et al. 2018; Xie et al. 2020), and have also been found in stools of both humans and pigs (Shan et al. 2011; Zhao et al. 2017). While their impact on health remains unmeasured, any such influence must be via biological modulation of their parasite hosts, and our findings provide the basis for answering this. While intriguing, the role of parabasalid infection in previously reported cases of equine disease and death cannot be determined here.

Our study improves the understanding of cressdnavirus ecology. Five cressdnavirus families were already known to infect eukaryotes including plants, vertebrates, algae, and fungi, and three were found to infect protists (Kinsella et al. 2020). Our findings add *Redondoviridae*, CRESSV1, CRESSV19, and *Kirkoviridae* to the latter group, meaning the majority of known cressdnavirus–eukaryote relationships now involve protists. We expect this reflects a broader pattern for the many undetermined relationships remaining.

Materials and methods

Cressdnavirus lineage inclusion

A database of cressdnavirus Rep sequences was compiled, containing classified and unclassified lineages. This was aligned to the GenBank nr database (April 2021) using BLASTp (Camacho et al. 2009), and non-redundant cressdnavirus hits were incorporated into the query. This process was iterated two further times, achieving a comprehensive set of 15,815 unique cressdnavirus Reps. Of these, 2,461 remained after clustering with CD-HIT v4.7 (Fu et al. 2012) at 70 per cent global amino acid identity. Reps belonging to the *Arfiviricetes* and *Repensiviricetes* classes were separately aligned using the MUSCLE v5.0.1278 super5 algorithm (Edgar 2021), with -perturb set from 0 to 4 to generate five versions. Best-fit amino acid substitution models were assessed to be VT + G4 + F for all alignments using ModelTest-NG v0.1.6 (Darriba et al. 2020). Maximum likelihood phylogenetic analysis was performed using IQ-TREE v2.1.4-beta (Minh et al. 2020), with settings --ninit 200 -bnni --allnni -B 1000 -alrt 1000. Trees were examined for consistency, and one was annotated per class (that with the highest likelihood score). Unclassified lineages were annotated if the cluster had an UFBoot score ≥85 and at least nine sequences (mean 31 and median 16). Isolation source and host records of annotated sequences were downloaded using Entrez Direct tools (Kans 2013), and used to determine which lineages would be included as 'gastrointestinal cressdnaviruses' (Supplementary Table S1). Strict criteria were not applied, but in practice inclusion required ≥70 per cent of source annotations to be gastrointestinal tract, stool, or wastewater. In the case of human-associated redondoviruses, found predominately in the human oral cavity, respiratory sources were accepted because we considered it plausible they were seeded or contaminated by oral secretions.

Viral recombination analyses

All available complete genome assemblies from gastrointestinal cressdnavirus lineages were rotated with MARS (Ayad and Pissis 2017) to ensure concordant start positions. Rotated sequences were aligned using MAFFT v7.487 (Katoh, Rozewicki, and Yamada 2017) with automatic settings, and recombination events were analysed using RDP4 v4.101 (Martin et al. 2015). RDP4 was also used to display phylogenetic compatibility and breakpoint pair distribution for the *Redondoviridae*. To construct tanglegrams for each lineage, Rep and Cap proteins were separately aligned using MAFFT v7.487 with automatic settings, and phylogenetic analysis was done using IQ-TREE v1.6.11. Treefiles were loaded into Dendroscope v.3.7.2 (Huson and Scornavacca 2012), rooted at the midpoint, and analysed with the tanglegram algorithm.

Cressdnavirus distribution across gastrointestinal tract samples

Publically available metagenomic datasets from 1,124 gastrointestinal tract samples belonging to seven cohorts were downloaded (Supplementary Table S2). BWA MEM

v0.7.17-r1188 (Li 2013) was used to map reads to 241 gastrointestinal cressdnavirus genomes (Supplementary Table S3). SAM files were processed using the PathoID module of PathoScope v2.0.7 (Hong et al. 2014). False positive mappings were removed by realigning filtered reads to the same genome database using BLASTn with settings -word_size 11 -gapopen 5 -gapextend 2 -penalty -3 -reward 2 -dust yes, and then removing unaligned reads and any with alignment length <40 from the original SAM file. Where original samples were accessible (human stool cohort 1), samples suspected of being false positive due to proximity to a highly positive sample were also curated with PCR (Supplementary Table S9). A matrix of viral distribution covering all cohorts was generated, with empty rows and columns removed. Counts were normalised to reads per million, log2 transformed, and visualised as a heatmap in GraphPad Prism v9.0 (GraphPad Software, San Diego, CA, USA, www.graphpad.com).

Classification of eukaryotic content in gastrointestinal tract samples

Reads from all 1,124 gastrointestinal samples were mapped to the combined SILVA 138.1 SSU and LSU NR99 databases (Quast et al. 2013). SAM files were processed with PathoScope as above before being filtered to remove bacterial and archaeal hits. Eukaryotic reads were realigned to the GenBank v5 nucleotide database (February 2021) using BLASTn and alignments were filtered with quality cutoffs according to the library preparation method and read length. Specifically, Illumina reads of 100 bp required 100 per cent identity for ≥50 bp, while those ≥150 bp read length required 100 per cent identity for ≥100 bp. VIDISCA IonTorrent reads (Kinsella, Deijs, and van der Hoek 2019) required ≥98 per cent identity for ≥100 bp to allow for possible homopolymer errors. Filtered outputs were processed using Linux command line tools to count occurrences of any specific taxon. Clinically validated qPCRs for *Blastocystis* spp. and *D. fragilis* were run on any sample previously tested for viruses by PCR (Supplementary Table S9), and count tables were updated accordingly.

Host prediction

Initial host prediction was done on the six of seven cohorts with Illumina deep sequencing data available (Supplementary Table S2). For each viral lineage, samples considered ‘highly positive’ were selected per cohort. To accommodate variation between different biological lineages and cohorts, we did not apply identical cutoffs, instead treating samples with normalised viral read counts (reads per million) above the inclusive lower quartile value as highly positive. Eukaryotic NCBI taxonomy ID numbers were extracted from the BLASTn tables of these samples, and converted into non-redundant lists of genera using Linux and Entrez Direct tools. Prevalent eukaryotes in highly virus positive samples were then identified using Linux command line tools. Genera normally resident in the gastrointestinal tract were retained, while transient taxa or otherwise implausible identifications were not. We did not apply strict percentage prevalence cutoffs for inclusion as a host candidate, although the lowest was 87.5 per cent (a genus detected in 7 of 8 highly

virus positive samples). Next, we tested statistical associations between viruses and respective host candidates across all samples in each separate cohort. Two tests were used; Pearson's chi-squared tests were used to determine if an association existed between presence of a host candidate and a respective cressdnavirus lineage (presence scored 1 and absence scored 0), while Spearman's rank correlation tests were used to determine any correlation between normalised loads of a host candidate and a cressdnavirus lineage. Genera with significant associations to a viral lineage were tested across all cohorts of the same sample type to assess reproducibility. For the main workflow code, see: https://github.com/CormacKinsella/Metagenomic-virus-host-prediction.

Endogenous viral element analysis

Selected eukaryotic genome assemblies (Supplementary Table S7) were downloaded and searched for Rep and Cap EVEs using tBLASTn (e-value threshold of 1e-5) and a query including 2,923 Rep and 2,122 Cap sequences. Alignment regions were converted to BED format with ascending coordinate ranges, and overlapping features were merged using BEDTools v2.27.1 (Quinlan and Hall 2010). Features were extracted as FASTA sequences, and open reading frames (ORFs) were predicted and translated using EMBOSS v6.3.1 getorf (Rice, Longden, and Bleasby 2000), with settings -minsize 120 -find 0. Virus-like sequences were separated from others using UBLAST v10 (Edgar 2010) and the same query database as above. Filtered candidate EVEs were then aligned to the GenBank nr database with BLASTp, and outputs were inspected to remove false positives. Sequences were clustered using CLANS (Frickey and Lupas 2004). To assess the phylogenetic affiliations of Rep-like EVE sequences, they were aligned alongside five representatives of each cressdnavirus lineage using MAFFT v7.487 E-INS-i, and analysed with IQ-TREE v1.6.11. Based on the results, alignment and phylogenetic analysis was done including all exogenous and endogenous members of the *Kirkoviridae*, using nenyaviruses as an outgroup, and the same for CRESSV1, using vilyaviruses as an outgroup. To confirm *Blastocystis* spp. EVEs were truly found inside genomes and were not assembly contaminants, we extracted genomic DNA from *Blastocystis* spp. axenic cultures belonging to subtypes 1, 2, 7, and 8 using the Boom method (Boom et al. 1990). We then designed and ran PCR assays on extracted DNA to amplify six selected EVEs, and attempted Sanger sequencing of products. To confirm *H. meleagridis* EVEs were genuine, we instead used a computational approach. We carried out all-vs.-all alignment of EVE-containing scaffolds from the two source genome assemblies (built from combined long and short read technologies) using nucmer --maxmatch --nosimplify, within MUMmer v4.0.0rc1 (Marçais et al. 2018). The delta file was then processed using mummerplot.

Human oral plaque qPCR

Subgingival plaques were collected with curettes from inflamed periodontal pockets of patients with clinically diagnosed periodontitis, at the Department of Periodontology, Oral Medicine and Oral Surgery, Charité—Universitätsmedizin Berlin. Plaque was directly

transferred into lysis buffer and DNA was extracted by the phenol/chloroform method. Samples were PCR screened using *E. gingivalis* specific primers (Bonner et al. 2014), and the human gene *ACTB* (beta-actin) was also amplified as a DNA isolation control (Supplementary Table S9). For this study, forty-eight DNA extractions with sufficient residual material were selected, comprising thirty-one *E. gingivalis* positive and seventeen negative samples. Three TAMRA qPCR assays targeting *Redondoviridae*, *E. gingivalis*, and *T. tenax* were designed (Supplementary Table S9), and tenfold dilutions of each target were used to construct standard curves and determine cycle threshold limits (Ct values ≥37 were considered negative). All forty-eight samples were screened once for the three targets alongside standards and negative controls. Association between test outcomes (positive scored 1 and negative scored 0) was assessed using Pearson's chi-squared test, and correlation between Ct values was explored using linear regression.

Ethical approval

Work with clinical samples from human subjects was approved by a vote of the local ethics committee (Campus Charité Mitte, application number EA1/169/20).

Data availability

All sequence datasets and genome assemblies utilised here are available in public databases; see Supplementary Tables S2, S3 and S7. For workflow code, see: https://github.com/CormacKinsella/Metagenomic-virus-host-prediction.

Acknowledgements

Computational work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

Funding

This work was supported by a grant from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie agreement No. 721367 (HONOURs), awarded to Lia van der Hoek.

References

Abbas, A.A. et al., 2019. Redondoviridae, a family of small, circular DNA viruses of the human oro-respiratory tract that are associated with periodontitis and critical illness. Cell Host Microbe 25, 719–729.

Ahlgren, N.A., Ren, J., Lu, Y.Y., Fuhrman, J.A., and Sun, F. 2017. Alignment-free d2* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Res. 45, 39–53.

Altan, E., Del Valle Mendoza, J., Deng, X., Phan, T.G., Sadeghi, M., and Delwart, E.L. 2017. Small circular Rep-encoding single-stranded DNA genomes in Peruvian diarrhea virome. Genome Announc. 5, e00822-17.

Ayad, L.A.K., and Pissis, S.P. 2017. MARS: Improving multiple circular sequence alignment using refined sequences. BMC Genomics 18, 1–10.

Babayan, S.A., Orton, R.J., and Streicker, D.G. (2018). Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. Science 362, 577–580.

Badri, M., Olfatifar, M., Abdoli, A., Houshmand, E., Zarabadipour, M., Abadi, P.A., Johkool, M.G., Ghorbani, A., and Eslahi, A.V. (2021). Current global status and the epidemiology of Entamoeba gingivalis in humans: A systematic review and meta-analysis. Acta Parasitol. 66, 1102–1113.

Bao, X., Wiehe, R., Dommisch, H., and Schaefer, A.S. (2020). Entamoeba gingivalis causes oral inflammation and tissue destruction. J. Dent. Res. 99, 561–567.

Bao, X., Weiner, J., Meckes, O., Dommisch, H., and Schaefer, A.S. (2021). Entamoeba gingivalis exerts severe pathogenic effects on the oral mucosa. J. Dent. Res. 100, 771–776.

Benabdelkader, S., Andreani, J., Gillet, A., Terrer, E., Pignoly, M., Chaudet, H., Aboudharam, G., and Scola, B. La (2019). Specific clones of Trichomonas tenax are associated with periodontitis. PLoS One 14, e0213338.

Bickhart, D.M., Watson, M., Koren, S., Panke-Buisse, K., Cersosimo, L.M., Press, M.O., Van Tassell, C.P., Van Kessel, J.A.S., Haley, B.J., Kim, S.W., et al. (2019). Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation. Genome Biol. 20.

Blot, S., Vandijck, D., and Labeau, S. (2008). Oral care of intubated patients. Clin. Pulm. Med. 15, 153–160.

Bonner, M., Amard, V., Bar-Pinatel, C., Charpentier, F., Chatard, J.M., Desmuyck, Y., Ihler, S., Rochet, J.P., De La Tribouille, V.R., Saladin, L., et al. (2014). Detection of the amoeba Entamoeba gingivalis in periodontal pockets. Parasite 21.

Boom, R., Sol, C.J., Salimans, M.M., Jansen, C.L., Wertheim-Van Dillen, P.M., and van der Noordaa, J. (1990). Rapid and simple method for purification of nucleic acids. J. Clin. Microbiol. 28, 495–503.

Cacciò, S.M., Sannella, A.R., Manuali, E., Tosini, F., Sensi, M., Crotti, D., and Pozio, E. (2012). Pigs as natural hosts of Dientamoeba fragilis genotypes found in humans. Emerg. Infect. Dis. 18, 838–841.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: Architecture and applications. BMC Bioinformatics 10, 421.

Darriba, Di., Posada, D., Kozlov, A.M., Stamatakis, A., Morel, B., and Flouri, T. (2020). ModelTest-NG: A new and scalable tool for the selection of DNA and protein evolutionary models. Mol. Biol. Evol. 37, 291–294.

Díez-Villaseñor, C., and Rodriguez-Valera, F. (2019). CRISPR analysis suggests that small circular single-stranded DNA smacoviruses infect Archaea instead of humans. Nat. Commun. 10.

Dion, M.B., Plante, P.L., Zufferey, E., Shah, S.A., Corbeil, J., and Moineau, S. (2021). Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. Nucleic Acids Res. 49, 3127–3138.

Dolja, V. V., and Koonin, E. V. (2018). Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. Virus Res. 244, 36–52.

Duffy, S., Burch, C.L., and Turner, P.E. (2007). Evolution of host specificity drives reproductive isolation among RNA viruses. Evolution 61, 2614–2622.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461.

Edgar, R.C. (2021). MUSCLE v5 enables improved estimates of phylogenetic tree confidence by ensemble bootstrapping. BioRxiv.

Edgar, R.C., Taylor, J., Lin, V., Altman, T., Barbera, P., Meleshko, D., Lohr, D., Novakovsky, G., Buchfink, B., Al-Shayeb, B., et al. (2022). Petabase-scale sequence alignment catalyses viral discovery. Nature 602, 142–147.

Ellis, J., Hassard, L., Clark, E., Harding, J., Allan, G., Willson, P., Strokappe, J., Martin, K., McNeilly, F., Meehan, B., et al. (1998). Isolation of circovirus from lesions of pigs with postweaning multisystemic wasting syndrome. Can. Vet. J. 39, 44–51.

Eng, C.L.P., Tong, J.C., and Tan, T.W. (2014). Predicting host tropism of influenza A virus proteins using random forest. BMC Med. Genomics 7.

Frickey, T., and Lupas, A. (2004). CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics 20, 3702–3704.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152.

Greninger, A.L. (2018). A decade of RNA virus metagenomics is (not) enough. Virus Res. 244, 218–229.

Guo, Z., He, Q., Tang, C., Zhang, B., and Yue, H. (2018). Identification and genomic characterization of a novel CRESS DNA virus from a calf with severe hemorrhagic enteritis in China. Virus Res. 255, 141–146.

Harding, R.M., Burns, T.M., Hafner, G., Dietzgen, R.G., and Dale, J.L. (1993). Nucleotide sequence of one component of the banana bunchy top virus genome contains a putative replicase gene. J. Gen. Virol. 74, 323–328.

Hong, C., Manimaran, S., Shen, Y., Perez-Rogers, J.F., Byrd, A.L., Castro-Nallar, E., Crandall, K.A., and Johnson, W.E. (2014). PathoScope 2.0: A complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome 2.

Huson, D.H., and Scornavacca, C. (2012). Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Syst. Biol. 61, 1061–1067.

Ignacio-Espinoza, J.C., Laperriere, S.M., Yeh, Y.-C., Weissman, J., Hou, S., Long, A.M., and Fuhrman, J.A. (2020). Ribosome-linked mRNA-rRNA chimeras reveal active novel virus host associations. BioRxiv.

Kans, J. (2013). Entrez Direct: E-utilities on the Unix Command Line. https://www.ncbi.nlm.nih.gov/books/NBK179288/.

Kapoor, A., Simmonds, P., Lipkin, W.I., Zaidi, S., and Delwart, E. (2010). Use of nucleotide composition analysis to infer hosts for three novel picorna-like viruses. J. Virol. 84, 10322–10328.

Katoh, K., Rozewicki, J., and Yamada, K.D. (2017). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. 20, 1160–1166.

Kazlauskas, D., Varsani, A., and Krupovic, M. (2018). Pervasive chimerism in the replication-associated proteins of uncultured single-stranded DNA viruses. Viruses 10.

Kinsella, C.M., Deijs, M., and van der Hoek, L. (2019). Enhanced bioinformatic profiling of VIDISCA libraries for virus detection and discovery. Virus Res. 263, 21–26.

Kinsella, C.M., Bart, A., Deijs, M., Broekhuizen, P., Kaczorowska, J., Jebbink, M.F., van Gool, T., Cotten, M., and van der Hoek, L. (2020). Entamoeba and Giardia parasites implicated as hosts of CRESS viruses. Nat. Commun. 11, 1–10.

Kirk, M.D., Pires, S.M., Black, R.E., Caipo, M., Crump, J.A., Devleesschauwer, B., Döpfer, D., Fazil, A., Fischer-Walker, C.L., Hald, T., et al. (2015). World Health Organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: A data synthesis. PLoS Med. 12, e1001921.

Krupovic, M., Varsani, A., Kazlauskas, D., Breitbart, M., Delwart, E., Rosario, K., Yutin, N., Wolf, Y.I., Harrach, B., Zerbini, F.M., et al. (2020). Cressdnaviricota: A virus phylum unifying seven families of Rep-encoding viruses with single-stranded, circular DNA genomes. J. Virol. 94.

Lefeuvre, P., Lett, J.M., Reynaud, B., and Martin, D.P. (2007). Avoidance of protein fold disruption in natural virus recombinants. PLoS Pathog. 3, e181.

Lefeuvre, P., Lett, J.M., Varsani, A., and Martin, D.P. (2009). Widely conserved recombination patterns among single-stranded DNA viruses. J. Virol. 83, 2697–2707.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv:1303.3997v1 [q-Bio.GN].

Li, L., Giannitti, F., Low, J., Keyes, C., Ullmann, L.S., Deng, X., Aleman, M., Pesavento, P.A., Pusterla, N., and Delwart, E. (2015). Exploring the virome of diseased horses. J. Gen. Virol. 96, 2721–2733.

Liu, D., Ma, Y., Jiang, X., and He, T. (2019). Predicting virus-host association by kernelized logistic matrix factorization and similarity network fusion. BMC Bioinformatics 20.

Liu, H., Fu, Y., Li, B., Yu, X., Xie, J., Cheng, J., Ghabrial, S.A., Li, G., Yi, X., and Jiang, D. (2011). Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. BMC Evol. Biol. 11.

Magee, C.J. (1927). Investigation on the bunchy top disease of the banana. Bull. Counc. Sci. Ind. Res. 30.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. PLoS Comput. Biol. 14, e1005944.

Martin, D.P., Murrell, B., Golden, M., Khoosal, A., and Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol. 1.

Marty, M., Lemaitre, M., Kémoun, P., Morrier, J.-J., and Monsarrat, P. (2017). Trichomonas tenax and periodontal diseases: A concise review. Parasitology 144, 1417–1425.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37, 1530–1534.

Munro, C.L., and Grap, M.J. (2004). Oral health and care in the intensive care unit: State of the science. Am. J. Crit. Care 13, 25–34.

Nagasaki, K., Tomaru, Y., Takao, Y., Nishida, K., Shirai, Y., Suzuki, H., and Nagumo, T. (2005). Previously unknown virus infects marine diatom. Appl. Environ. Microbiol. 71, 3528–3535.

Palmieri, N., de Jesus Ramires, M., Hess, M., and Bilic, I. (2021). Complete genomes of the eukaryotic poultry parasite Histomonas meleagridis: Linking sequence analysis with virulence / attenuation. BMC Genomics 22.

Phan, T.G., Costa, A.C. da, Mendoza, J. del V., Bucardo-Rivera, F., Nordgren, J., O'Ryan, M., Deng, X., and Delwart, E. (2016). The fecal virome of South and Central American children with diarrhea includes small circular DNA viral genomes of unknown origin. Arch. Virol. 161, 959–966.

Poulsen, C.S., and Stensvold, C.R. (2016). Systematic review on Endolimax nana: A less well studied intestinal ameba. Trop. Parasitol. 6, 8.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 41, D590–D596.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842.

Ramos, E.D.S.F., Ribeiro, G. de O., Villanova, F., Milagres, F.A. de P., Brustulin, R., Araújo, E.L.L., Pandey, R.P., Raj, V.S., Deng, X., Delwart, E., et al. (2021). Composition of eukaryotic viruses and bacteriophages in individuals with acute gastroenteritis. Viruses 13.

Rice, P., Longden, L., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16, 276–277.

Ritchie, B.W., Niagro, F.D., Lukert, P.D., Steffens, W.L., and Latimer, K.S. (1989). Characterization of a new virus from cockatoos with psittacine beak and feather disease. Virology 171, 83–88.

Scanlan, P.D., Stensvold, C.R., Rajilić-Stojanović, M., Heilig, H.G.H.J., De Vos, W.M., O'Toole, P.W., and Cotter, P.D. (2014). The microbial eukaryote Blastocystis is a prevalent and diverse member of the healthy human gut microbiota. FEMS Microbiol. Ecol. 90, 326–330.

Scannapieco, F.A. (1999). Role of oral bacteria in respiratory infection. J. Periodontol. 70, 793–802.

Shan, T., Li, L., Simmonds, P., Wang, C., Moeser, A., and Delwart, E. (2011). The fecal virome of pigs on a high-density farm. J. Virol. 85, 11697–11708.

Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., Qin, X.-C., Li, J., Cao, J.-P., Eden, J.-S., et al. (2016). Redefining the invertebrate RNA virosphere. Nature 540, 539–543.

Simmonds, P., Adams, M.J., Benkő, M., Breitbart, M., Brister, J.R., Carstens, E.B., Davison, A.J., Delwart, E., Gorbalenya, A.E., Harrach, B., et al. (2017). Virus taxonomy in the age of metagenomics. Nat. Rev. Microbiol. 15, 161–168.

Stensvold, C.R., Suresh, G.K., Tan, K.S.W., Thompson, R.C.A., Traub, R.J., Viscogliosi, E., Yoshikawa, H., and Clark, C.G. (2007). Terminology for Blastocystis subtypes - a consensus. Trends Parasitol. 23, 93–96.

Tam, C.C., O'Brien, S.J., Tompkins, D.S., Bolton, F.J., Berry, L., Dodds, J., Choudhury, D., Halstead, F., Iturriza-Gómara, M., Mather, K., et al. (2012). Changes in causes of acute gastroenteritis in the United Kingdom over 15 Years: Microbiologic findings from 2 prospective, population-based studies of infectious intestinal disease. Clin. Infect. Dis. 54, 1275–1286.

Thumbi, S.M., Njenga, M.K., Marsh, T.L., Noh, S., Otiang, E., Munyua, P., Ochieng, L., Ogola, E., Yoder, J., Audi, A., et al. (2015). Linking human health and livestock health: A "one-health" platform for integrated analysis of human health, livestock health, and economic welfare in livestock dependent communities. PLoS One 10, e0120761.

Tisza, M.J., Pastrana, D. V., Welch, N.L., Stewart, B., Peretti, A., Starrett, G.J., Pang, Y.Y.S., Krishnamurthy, S.R., Pesavento, P.A., McDermott, D.H., et al. (2020). Discovery of several thousand highly diverse circular DNA viruses. Elife 9, e51971.

Varma, A., and Malathi, V.G. (2003). Emerging geminivirus problems: A serious threat to crop production. Ann. Appl. Biol. 142, 145–164.

Wick, R.R., Judd, L.M., Gorrie, C.L., and Holt, K.E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput. Biol. 13, e1005595.

Xie, J., Tong, P., Zhang, A., Yan, Y., Zhang, L., Song, X., Chen, J., Zhai, S., Shaya, N., Wang, D., et al. (2020). First detection and genetic characterization of a novel kirkovirus from a dead thoroughbred mare in northern Xinjiang, China, in 2018. Arch. Virol. 165, 403–406.

Yoon, H.S., Price, D.C., Stepanauskas, R., Rajah, V.D., Sieracki, M.E., Wilson, W.H., Yang, E.C., Duffy, S., and Bhattacharya, D. (2011). Single-cell genomics reveals organismal interactions in uncultivated marine protists. Science 332, 714–717.

Yu, X., Li, B., Fu, Y., Jiang, D., Ghabrial, S.A., Li, G., Peng, Y., Xie, J., Cheng, J., Huang, J., et al. (2010). A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. Proc. Natl. Acad. Sci. 107, 8387–8392.

Zhang, Y., Wang, C., Feng, X., Chen, X., and Zhang, W. (2021). Redondoviridae and periodontitis: a case–control study and identification of five novel redondoviruses from periodontal tissues. Virus Evol. 7, 33.

Zhao, G., Vatanen, T., Droit, L., Park, A., Kostic, A.D., Poon, T.W., Vlamakis, H., Siljander, H., Härkönen, T., Hämäläinen, A.M., et al. (2017). Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. Proc. Natl. Acad. Sci. 114, E6166–E6175.

Zhao, L., Lavington, E., and Duffy, S. (2021). Truly ubiquitous CRESS DNA viruses scattered across the eukaryotic tree of life. J. Evol. Biol. 34, 1901–1916.

Supplementary figures

For supplementary tables, see the online version: https://doi.org/10.1093/ve/veac087.

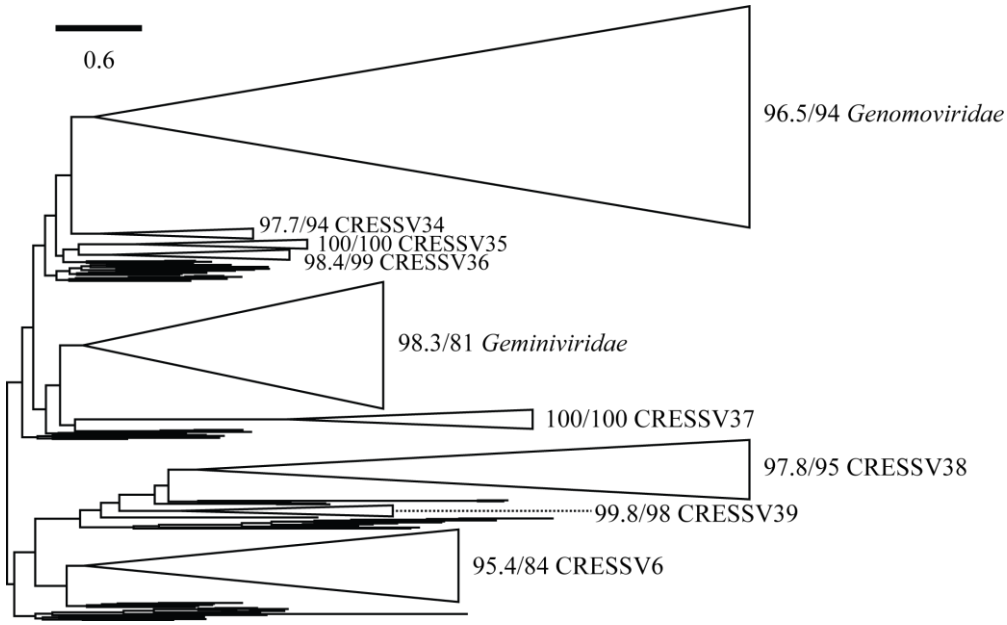


Figure S1. Maximum likelihood phylogenetic tree of the *Repensiviricetes*, rooted at the midpoint. Scale bar denotes amino acid substitutions per site. Branch supports are given for each named lineage, with SH-aLRT scores on the left and ultrafast bootstrap scores on the right. All sequences found outside of collapsed nodes did not meet criteria for naming a lineage.

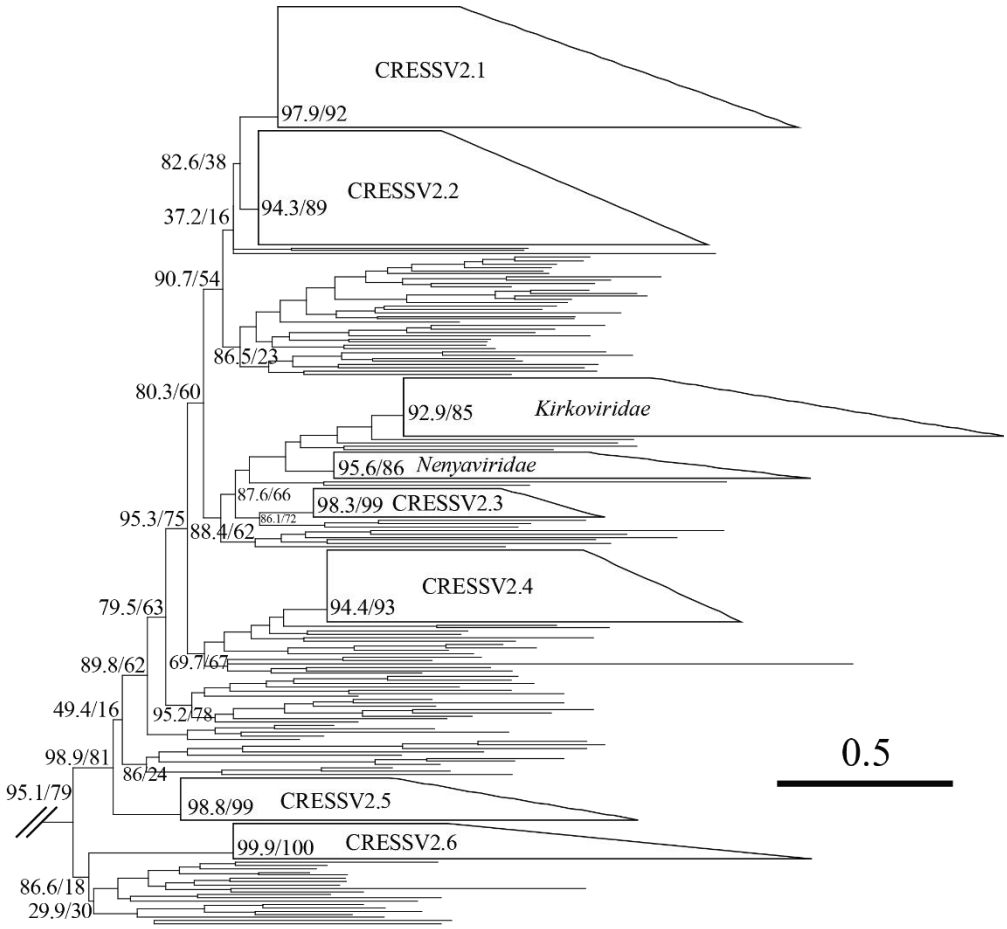


Figure S2. Part of a maximum likelihood phylogenetic tree of the *Arfiviricetes*, focused on the CRESSV2-like sublineages. Scale bar denotes amino acid substitutions per site. Branch supports have SH-aLRT scores on the left and ultrafast bootstrap scores on the right. All sequences found outside of collapsed nodes were CRESSV2-like, but did not meet criteria for naming a lineage.

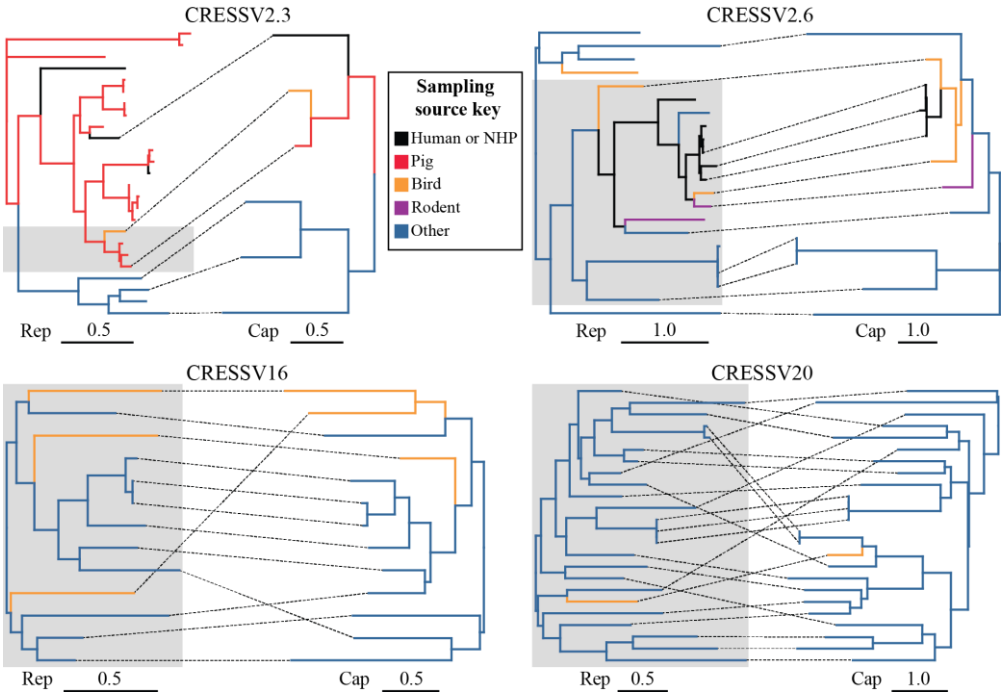


Figure S3. Recombination within gastrointestinal cressdnavirus lineages. Rep and Cap protein tanglegrams for four cressdnavirus lineages. Dotted lines connect proteins encoded by the same genome. Branch colour denotes isolation source as listed in the key. Grey blocks denote groups linked by RDP4 detected recombination events. Scale bars on individual phylograms are in amino acid substitutions per site. NHP: non-human primate.

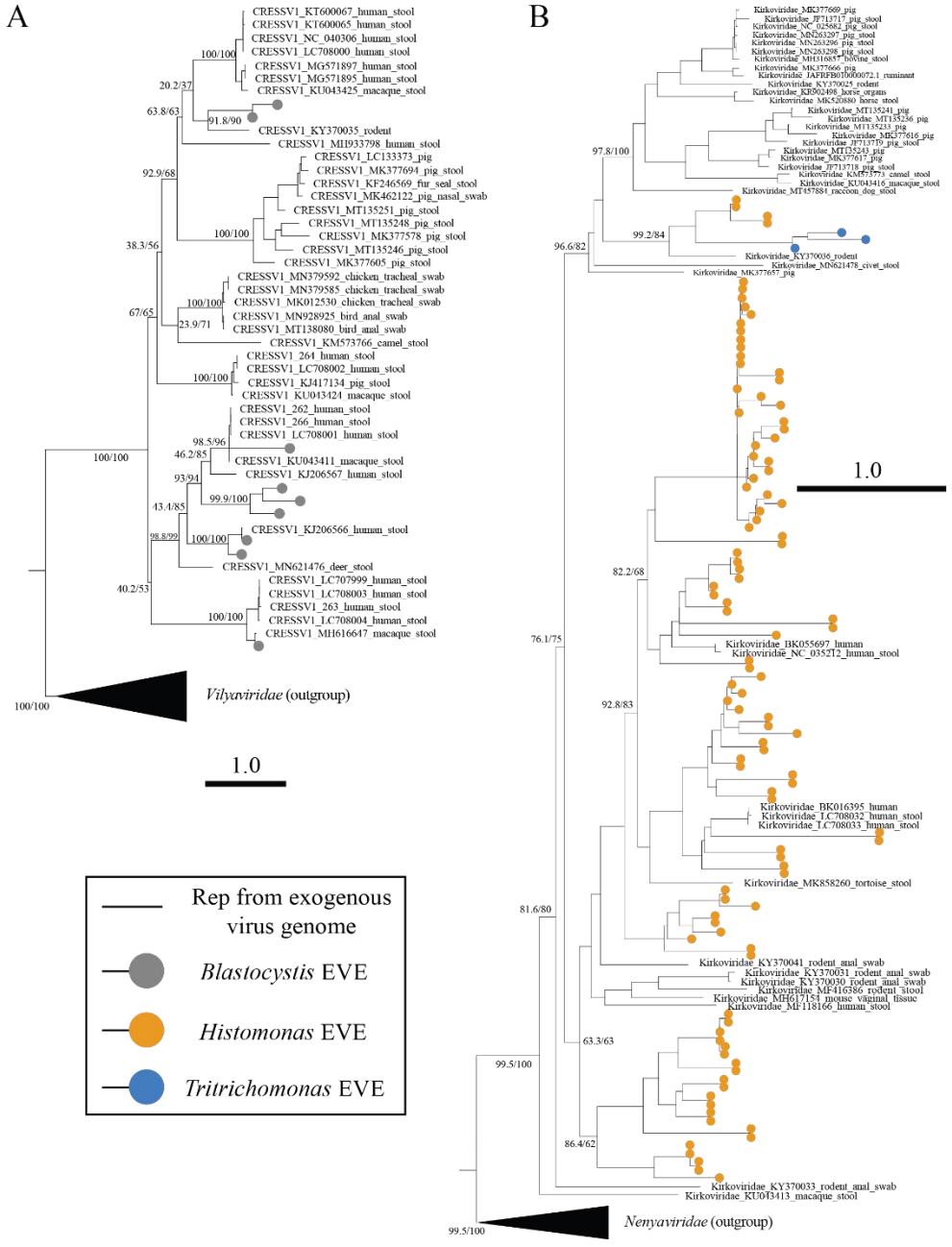


Figure S4. Endogenous viral elements (EVEs) belong to classified cressdnavirus lineages. A) Maximum likelihood phylogenetic tree of CRESSV1 members, showing some *Blastocystis* spp. EVEs belong to the lineage. B) Maximum likelihood phylogenetic tree of the *Kirkoviridae*, showing some EVEs of parabasalid taxa belong to the family. Branch supports for both trees report SH-aLRT scores on the left and ultrafast bootstrap scores on the right. Scale bars refer to amino acid substitutions per site.

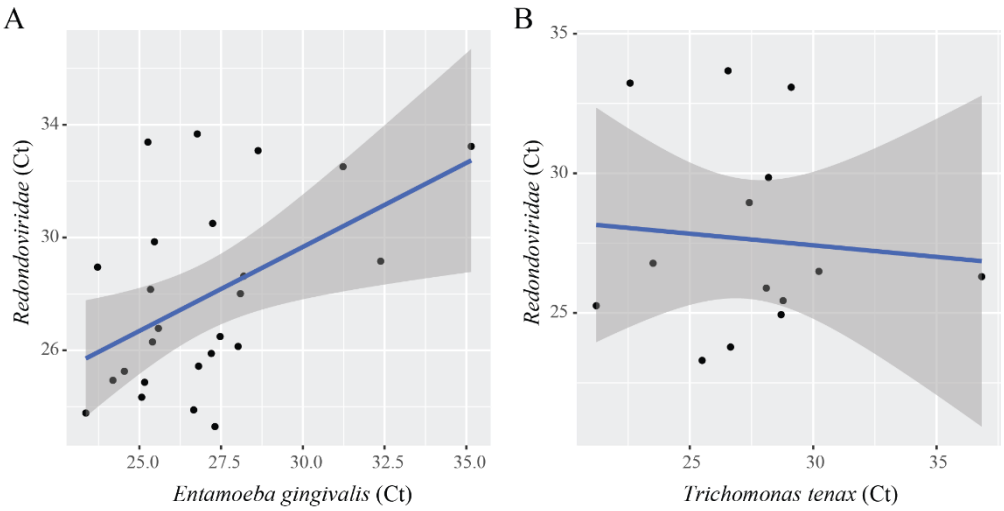
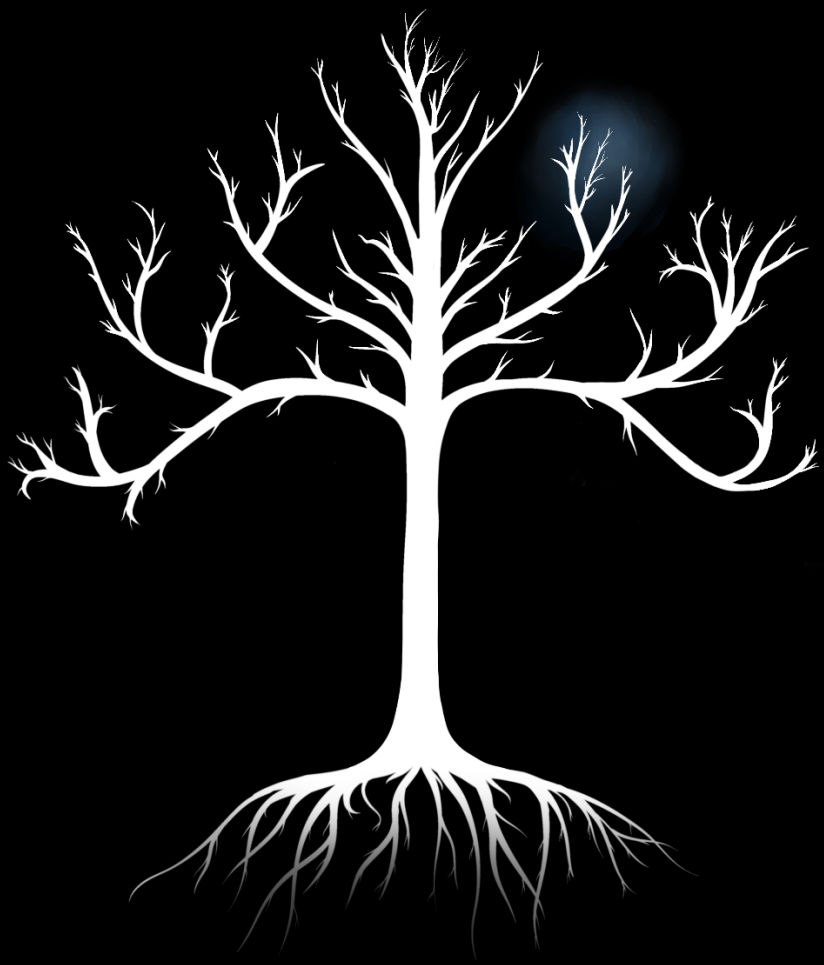


Figure S5. Relationships between qPCR measured loads of protists and redondoviruses in human oral plaques. A) qPCR cycle threshold (Ct) values of *Entamoeba gingivalis* versus *Redondoviridae*. A linear regression model is plotted with 95% confidence intervals denoted in grey, and a positive association was found ($R^2 = 0.24$, $p = 0.013$). B) No association was found between loads of *Trichomonas tenax* and *Redondoviridae* ($R^2 = 0.01$, $p = 0.762$). The assays were run once on all samples (N = 48) alongside standards and negative controls, with Ct values ≥37 considered negative.



Chapter 5

Vertebrate-tropism of a cressdnavirus lineage implicated by poxvirus gene capture

Cormac M. Kinsella, Lia van der Hoek

PNAS, 2023

https://doi.org/10.1073/pnas.2303844120

Abstract

Among cressdnaviruses, only the family *Circoviridae* is recognized to infect vertebrates, while many others have unknown hosts. Detection of virus-to-host horizontal gene transfer is useful for solving such virus–host relationships. Here, we extend this utility to an unusual case of virus-to-virus horizontal transfer, showing multiple ancient captures of cressdnavirus *Rep* genes by avipoxviruses—large dsDNA pathogens of birds and other saurians. As gene transfers must have occurred during virus coinfections, saurian hosts were implied for the cressdnavirus donor lineage. Surprisingly, phylogenetic analysis revealed that donors were not members of the vertebrate-infecting *Circoviridae*, instead belonging to a previously unclassified family that we name *Draupnirviridae*. While draupnirviruses still circulate today, we show that those in the genus *Krikovirus* infected saurian vertebrates at least 114 Mya, leaving endogenous viral elements inside snake, lizard, and turtle genomes throughout the Cretaceous Period. Endogenous krikovirus elements in some insect genomes and frequent detection in mosquitoes imply that spillover to vertebrates was arthropod mediated, while ancestral draupnirviruses likely infected protists before their emergence in animals. A modern krikovirus sampled from an avipoxvirus-induced lesion shows that their interaction with poxviruses is ongoing. Captured *Rep* genes in poxvirus genomes often have inactivated catalytic motifs, yet near-total presence across the *Avipoxvirus* genus, and evidence of both expression and purifying selection on them suggests currently unknown functions.

Significance

A single family of cressdnaviruses is known to infect vertebrates, the *Circoviridae*. Here, we identified a second that has historically infected saurians, naming them the *Draupnirviridae*. The initial clue was that some draupnirviruses donated their *Rep* gene to poxviruses exclusively infecting birds and their relatives. Since this implied the donors also infected vertebrates, a search for draupnirvirus-derived endogenous viral elements was done in ~25,000 eukaryotic genome assemblies. This confirmed that some draupnirviruses infected saurian vertebrates as long ago as the Cretaceous Period, over 100 million years before present, and they still circulate today. We propose that their evolutionary path likely began with protist-infecting ancestors, followed by emergence in insects, and eventual transfer to vertebrates by blood feeders.

Introduction

The majority of newly identified viral genomes are from “stray viruses,” which we define as those with known genome sequences but with unknown hosts. This situation arose due to widespread application of metagenomic high-throughput sequencing, an efficient culture-

independent virus discovery method (1, 2). When applied to samples containing diverse lifeforms, linkage of viruses to specific hosts is challenging. Identifying hosts is essential to understanding virus evolution and their medical or ecological roles (3). Viruses of eukaryotes with circular single-stranded DNA (ssDNA) genomes and a homologous replication-associated protein (Rep) are classified in the phylum *Cressdnaviricota* (4), currently containing 11 official families (5) and 46 unclassified lineages (3, 6, 7). Across the phylum, only some members of the family *Circoviridae* are recognized to infect vertebrates, most notably the pathogens porcine circovirus 2 (PCV2) and beak and feather disease virus (BFDV), infecting pigs and birds, respectively (8–10).

Fossils in eukaryotic genomes known as endogenous viral elements (EVEs) are the product of horizontal gene transfer (HGT) from viruses to host germline genomes (11), and analysis of EVE genetic relationships has helped identify hosts of some stray viruses (3, 12, 13). *Circoviridae*-derived EVEs have previously been identified in vertebrate genomes including snakes and mammals, some dating back as far as 65 to 68 million years, accounting for recent host divergence estimates (11, 14, 15). Though infrequent, analogous HGT between groups of unrelated viruses has also been observed (16–18). For example, some dsDNA herpesviruses and adenoviruses possess *Rep* genes gained from ssDNA parvoviruses (19–21), the *U94* gene of human herpesvirus 6 (HHV-6) being a well-characterized example (16, 22). Since virus-to-virus HGT requires coinfection by donor and recipient, if the host of one is known, inference of the second is theoretically possible.

Upon genome sequencing of canarypox virus (CNPV, family *Poxviridae*, genus *Avipoxvirus*), sequence similarity was observed between genes *CNPV153*, *CNPV200*, and circovirus *Reps*, suggesting HGT between ssDNA cressdnaviruses and a dsDNA avipoxvirus (23). Avipoxviruses primarily infect birds (24), though other saurians including turtles and lizards are also hosts (25, 26). If *CNPV153* and *CNPV200* truly represent HGT from cressdnaviruses, this suggests the donor viruses also infected saurians. Since discovery, the CNPV *Rep*-like genes have not been further researched. With over 1,000 *Poxviridae* genome assemblies now available, detailed comparative analysis is possible. Among the *Cressdnaviricota*, multiple thousands of genomes and a revised taxonomy have also facilitated research into their diversity (4). Here, we investigated possible HGT between viral realms, showing *Rep* genes were indeed horizontally transferred to an ancestor of extant avipoxviruses. Surprisingly, we found that donor viruses belonged to the unclassified lineage CRESSV3, which we propose be officially named as the family *Draupnirviridae*. Confirming our hypothesis, we found that draupnirviruses of the genus *Krikovirus* first infected saurian hosts at least 114 Mya.

Results

Rep Was Donated by a Cressdnavirus to an Ancestor of Extant Avipoxviruses.

To detect HGT from cressdnaviruses to poxviruses, we screened 1,090 poxvirus genomes using a phylogenetically broad protein database comprising cressdnavirus Reps and Caps. We found 89 sequences with high sequence identity to cressdnavirus Reps within 51 poxvirus genomes, including that of CNPV (*SI Appendix*, Tables S1 and S2). All the 51 genomes belonged to the genus *Avipoxvirus*. Other genera were not detected, including during manual examination of *Macropopoxvirus*, the closest relative of *Avipoxvirus* (*SI Appendix*, Fig. S1*A*), and *Crocodylidpoxvirus*, the other genus infecting archosaurs (crocodiles). The 51 *Rep*-like+ genomes represent all but one sequenced avipoxviruses, as teiidaepox virus 1 (TePV-1) contained none (Fig. 1*A*). Rather than ancestral absence, this likely reflects gene loss, as TePV-1 phylogenetically nests within *Rep*-like+ avipoxviruses (*SI Appendix*, Fig. S1*B*). Our result confirms that the observations of Tulman et al. (23) represent HGT, extending this to extant avipoxviruses rather than CNPV specifically. Hereafter, we refer to HGT-derived *Rep* genes as *apvRep* genes (for avipoxvirus *Rep*).

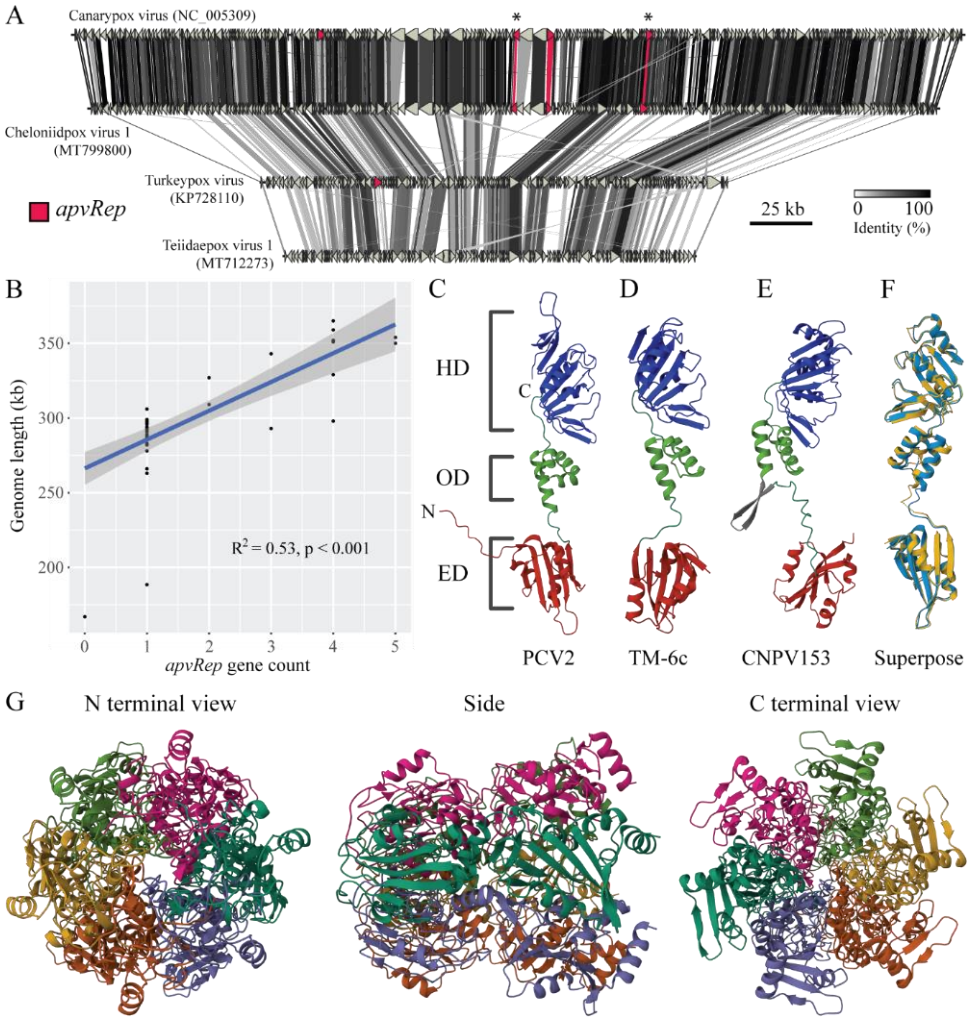


Fig. 1. Cressdnaviruses horizontally transferred *Rep* to members of the genus *Avipoxvirus* (*Poxviridae*). (*A*) Synteny map of four avipoxvirus genomes. Red-highlighted *apvReps* have been enlarged for visibility. The two asterisked *apvReps* denote *CNPV153* and *CNPV200* (left to right respectively), discussed by Tulman et al. (23). (*B*) Scatterplot and linear regression showing the relationship between *apvRep* gene count and genome size. (*C*) AlphaFold predicted structure of the complete PCV2 Rep protein (*Circoviridae*, NP_937956.1). Domains are coloured and annotated. ED = endonuclease domain, OD = oligomerisation domain, HD = helicase domain. N and C denote the respective termini. (*D*) AlphaFold predicted structure of Rep from the prototypical krikovirus TM-6c (27, 28) (CRESSV3, ADI48253.1). (*E*) AlphaFold predicted structure of CNPV153 (*apvRep-2*, NP_955176.1). The predicted pair of antiparallel beta strands are coloured grey. (*F*) Superposed structure alignment between TM-6c Rep (cyan) and CNPV153 (gold) performed with the jFATCAT flexible algorithm; rmsd = 1.02, TM-score = 0.32, query (TM-6c) coverage 99%, target coverage 84%. (*G*) Orthogonal views of the predicted hexamer structure of CNPV153.

The *apvRep* gene count varied from one to five in avipoxvirus genomes, excepting TePV-1 (*SI Appendix*, Fig. S1*B*). We observed that specific genes maintained their relative genomic positions across the genus; for example, three of the four *apvReps* in CNPV are syntenic with those in cheloniidpox virus 1 (ChePV-1, Fig. 1*A*), while the fourth in CNPV is syntenic with the lone *apvRep* in turkeypox virus HU1124/2011 (TKPV HU1124). This suggests that discrete homologous *apvReps* are found across avipoxvirus genomes. Using gene synteny analyses, we could group the 89 *apvRep* sequences into five separate genes (*apvRep-1* to *apvRep-5*). We found *apvRep-1* was distinct in having alleles across 48 of 51 *apvRep+* genomes, covering the breadth of recognized avipoxvirus diversity. This suggests that *apvRep-1* is the oldest surviving, gained in an ancestor of the genus. Notably, *apvRep* gene count follows a phylogenetic pattern; early-branching avipoxviruses contain *apvRep-1* only, while subsequent branches gained genes stepwise over time (*SI Appendix*, Fig. S1*B*). Increasing genome size may have driven or facilitated this, as early-branching avipoxviruses have short genomes (e.g., TKPV HU1124 and TePV-1, 189 and 167 kb, respectively) compared with late-branching species (e.g., CNPV and ChePV-1, 365 and 343 kb, respectively). Small genome size is likely an ancestral trait of avipoxviruses, given the genome lengths of closely related macropopoxviruses (167 to 170 kb). As predicted, *apvRep* gene count positively correlates with genome length (Fig. 1*B*), suggesting larger genomes can house and benefit from increased gene dosage.

Large variability was found in the predicted protein lengths encoded by different *apvRep* genes, for example apvRep-1 sequences ranged from 100 to 111 amino acids (aa), while some apvRep-3 sequences reached 1,006 aa (*SI Appendix*, Table S2). Sequences of apvRep-2 were predicted at lengths comparable to exogenous cressdnavirus Reps, 310 to 312 aa long. One of these was encoded by *CNPV153*, originally noted as a cressdnavirus-like gene by Tulman et al. (23). To explore structural conservation of these full-length apvReps in comparison to cressdnavirus proteins, we predicted the Rep structures of PCV2 (*Circoviridae*), TM-6c (27) (CRESSV3), and CNPV153 using AlphaFold. The PCV2 predicted structure was highly consistent with experimental solutions of all the three expected domains, the endonuclease (29, 30), oligomerization, and helicase domains (31) (Fig. 1*C*). The TM-6c prediction resembled that of PCV2 (Fig. 1*D*), though no experimental solutions from CRESSV3 are published. Overall, the predicted structure of CNPV153 matched the cressdnavirus Reps; for example, the endonuclease domains all shared a five-stranded beta sheet with two alpha helices on one side and a third on the other (Fig. 1*E*). Relaxed structure alignment between TM-6c and CNPV153 revealed broad concordance (Fig. 1*F*). However, AlphaFold predicted a pair of antiparallel beta strands 14 residues long, just N-terminal of the CNPV153 oligomerization domain, which were not seen in the cressdnaviruses (Fig. 1*E*). This prevented full TM-6c and CNPV153 alignment coverage using a rigid structure method (*SI Appendix*, Fig. S1*C*). Whether this represents prediction error or true biology is unclear, though if the latter—any disruptive effect on oligomerization would be notable. To explore this, we predicted the structure of hexameric CNPV153, the subunit count found in PCV2 Rep complexes (31). AlphaFold predicted

hexameric CNPV153 to form a torus (Fig. 1*G*) with similarity to the solved PCV2 hexamer (which covers the oligomerization and helicase domains) (31). Unlike with monomer prediction, the antiparallel beta strands adjacent to the oligomerization domain were not observed, suggesting that oligomerization of CNPV153 is possible.

After finding that apvRep-2 structure is broadly conserved with cressdnavirus Reps, we investigated why other *apvRep* genes displayed high variability in predicted protein length. Most extreme among these were intact *apvRep-3* alleles, which had a bimodal length distribution, encoding either 176 to 179 aa or 827 to 1,006 aa, far longer than any canonical cressdnavirus Rep. Using a comparative genomics approach, we found that long alleles came about by gene fusion, subsequently inherited by four species (*SI Appendix*, Table S2). Illustrating this, the long allele in finchpox virus shares synteny and sequence identity with three separate open-reading frames in ChePV-1, the last of which is a short *apvRep-3* allele (*SI Appendix*, Fig. S1*D*). We did not observe a genome with only two fused genes, and thus cannot determine whether fusion occurred in one step or two. The functions of the two genes sometimes fused to *apvRep-3* are unknown. Searches for conserved domains in the unfused homologs of ChePV-1 revealed none for the first (*ChPV157*, QRI42875.1), though the second (*ChPV158*, QRI42876.1) contained domain similarity to accession cl31759 (ring-infected erythrocyte surface antigen domain, known from *Plasmodium falciparum*). An equivalent search in the fused allele of CNPV (*CNPV156*, NP_955179.1) found the same, plus a hit to cl38662 (domain of unknown function found in the roundworm class *Chromadorea*). A search of the finchpox virus-fused allele (UOX38671.1) additionally detected similarity to cl27103 (secretion system effector C-like domain, found in some bacterial pathogens). Though the biological significance of these findings is uncertain, it is notable that *CNPV156* is found in a virulence and host range–related genomic region (23).

After accounting for the high length of some alleles, we investigated those shorter than typical Rep proteins. Seven of nine collective *apvRep* alleles from crowpox virus, magpiepox virus 1, and magpiepox virus 2 (ON408417.1, MK903864.1, and MW485973.1) were either fragmented by stop codons, truncated relative to homologous alleles in other species, or missed start codons, evidence of pseudogenization. The three genomes are closely related (*SI Appendix*, Fig. S1*B*), suggesting that a lineage-specific loss of several *apvRep* genes is ongoing. Similar sequence degradation was not observed in *apvReps* of other species, though we noted that apart from the full-length *apvRep-2* already discussed, all genes were simplified in terms of their domain architecture. They encoded either the endonuclease (*apvRep-1*, *apvRep-3*, and *apvRep-4*) or the helicase (*apvRep-5*) (Fig. 2 and *SI Appendix*, Fig. S2), explaining alleles with low predicted length. The finding implies the five *apvRep* genes originated from at least two gene capture events, as *apvRep-1* lost its helicase domain in an ancestor of all extant avipoxviruses, and thus is not a paralog of either *apvRep-2* or *apvRep-5*, which appeared later in avipoxvirus evolution and possess a helicase. The majority of *apvRep* alleles lack the oligomerization domain (*SI Appendix*, Fig. S2), suggesting normal Rep functionality is altered or absent. Cressdnavirus Rep proteins

require nuclear import for normal replicative functions, and for PCV2 Rep, the nuclear localization signal is within the 20 N-terminal residues (32). The equivalent residues are absent or contain a sizable deletion in all *apvRep* alleles, consistent with the cytoplasmic localization of poxviruses. Furthermore, some key functional motifs of Rep proteins are not conserved in apvRep sequences. The endonuclease active site HUH motif crucial for ssDNA cleavage activity (HUQ in *Circoviridae* (33) and several other cressdnavirus lineages) is inactivated in all *apvRep* genes with endonuclease domains, while the key tyrosine residue that covalently binds the 5′ phosphate of cleaved ssDNA is missing in all but *apvRep-2* (Fig. 2) (34). The helicase motifs of *apvRep-2* and *apvRep-5* appear more conserved overall, including the key GK residues used in nucleotide binding of the Walker A motif, and Walker B residues involved in ATPase activity (35). However, arginine finger sequences were not fully conserved, with potential impacts on nucleotide hydrolysis. Overall, full canonical enzymatic activity by apvRep proteins appears unlikely, though the potential for some nucleotide interaction may remain. These findings are remarkably similar to the U94 protein of HHV-6, which has inactivated endonuclease motifs but a conserved Walker A and B, and partial conservation of other helicase motifs (*SI Appendix*, Fig. S3), compatible with experimental results showing endonuclease inactivity yet retention of several helicase functions (22).

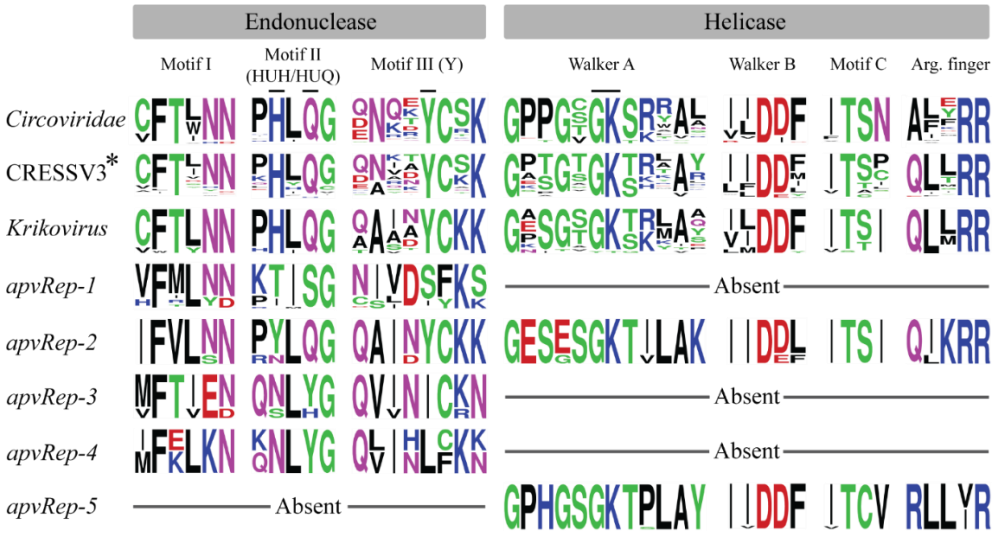


Fig. 2. Rep protein sequence motifs in cressdnaviruses and avipoxviruses. Asterisked lineage CRESSV3 (also referred to as *Draupnirviridae* in this study) includes members except for *Krikovirus*, which is shown separately. Arg. = arginine. Residue colours: hydrophobic = black, polar = green, basic = blue, acidic = red, neutral = purple. Key residues discussed in the main text are marked.

That apvRep sequences vary dramatically in their represented domains and often have canonical functional motifs inactivated may appear contradictory with their near-total

presence across the genus *Avipoxvirus*, and conservation since the common ancestor, which together suggest *apvRep* presence enhances viral fitness. To examine whether they could be pseudogenes (except those seven already discussed), we looked for evidence of their expression. Using published RNA sequencing data (PRJNA524335) from cell cultures infected for 16 h with either CNPV or fowlpox virus (FPV) (36), we confirmed that at least *apvRep-2*, *apvRep-3*, and *apvRep-5* are expressed by CNPV, while *apvRep-1* was silent (*SI Appendix*, Fig. S4). Interestingly, *apvRep-1* was expressed in FPV, which lacks other *apvReps*, and this difference may be due to the genetic redundancy found in CNPV. While no proteomic data was available to analyze *apvRep* translation, we found indirect evidence via purifying selection on *apvRep* coding sequences—including those encoding only the endonuclease (*apvRep-1*), all three domains (*apvRep-2*), or only the helicase (*apvRep-5*). Global d_N/d_S ratios were estimated at 0.23, 0.33, and 0.17, respectively, indicating purifying selection has acted to conserve each coding sequence and implying translation occurs. At the site level, the majority of *apvRep-1* codons had an estimated $d_N/d_S < 1$ and many were statistically significant for purifying selection (Fig. 3*A*). These tended to cluster on the endonuclease domain itself, and also the codons C-terminal of it. We noted that *apvRep-1* also had comparatively few invariant sites (Fig. 3 *A*–*C*), though alleles were available from most avipoxvirus species, and this likely introduced more variation. Both factors probably reflect its relatively old age. Sites of *apvRep-2* were mostly invariant (Fig. 3*B*), though it is found in relatively few species and is the only fully intact *apvRep*, suggesting evolutionary youth. Despite this, we still observed evidence of purifying selection on some sites across the gene. Notably, the first residue of what was once the endonuclease HUH/HUQ motif may have experienced diversifying selection, inactivating it; however, the *p*-value did not reach significance (0.059). One residue just C-terminal of the Walker B motif also reached significance for diversifying selection, though the possible impact on helicase activity is unclear. The *apvRep-5* gene displayed strong evidence of purifying selection targeting the helicase domain, which was dense with sites significantly below d_N/d_S of 1 (Fig. 3*C*). Overall, the evidence suggests *apvReps* of each domain architecture experience purifying selection on the peptide sequence and are not pseudogenes. Given the tendency for selection to target the functional domains themselves, it appears likely that tertiary structure is being maintained. With evidence suggesting canonical enzymatic Rep functionality is disrupted, *apvRep* genes may instead provide nonenzymatic Rep functions or have undergone exaptation.

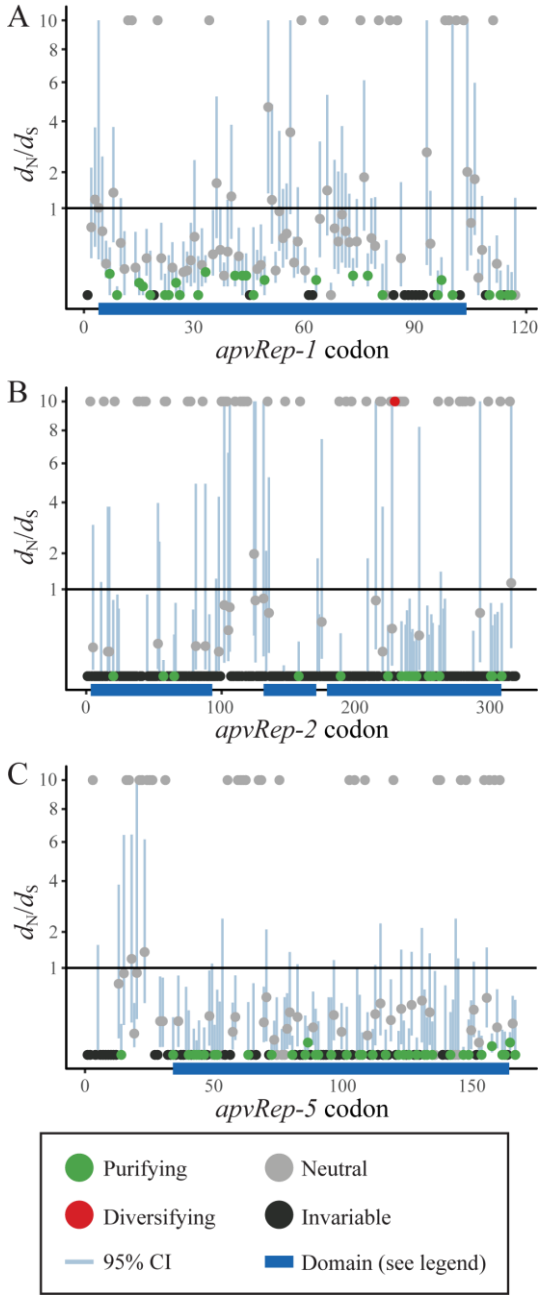


Fig. 3. Purifying selection has acted on *apvReps*. (*A*) Site-by-site d_N/d_S estimates for *apvRep-1*. The endonuclease domain is annotated. Statistically significant difference from null (neutrality, i.e., $d_N/d_S = 1$) required $P \leq 0.05$. (*B*) Site-by-site d_N/d_S estimates for *apvRep-2*. From left to right, the endonuclease, oligomerisation, and helicase domains are annotated. (*C*) Site-by-site d_N/d_S estimates for *apvRep-5*. The helicase domain is annotated.

Cressdnavirus *Rep* Donors Belong to a New Cressdnavirus Family, *Draupnirviridae*.

Since avipoxviruses infect saurian hosts and cressdnaviruses have donated *Reps* to them, we predicted the donor lineage would also infect saurians. We thus expected them to fall within the vertebrate-infecting *Circoviridae*, yet phylogenetic analysis alongside all cressdnavirus lineages resolved apvRep sequences within the unclassified lineage CRESSV3 (7) (*SI Appendix*, Fig. S5). A second more focused phylogeny confirmed that apvRep proteins belong to CRESSV3, specifically within the previously proposed genus *Krikovirus* (28) (Fig. 4). Our analyses thus robustly support *Krikovirus* as a monophyletic lineage within CRESSV3 (Fig. 4 and *SI Appendix*, Fig. S6), and as the lineage that donated *Reps* to avipoxviruses.

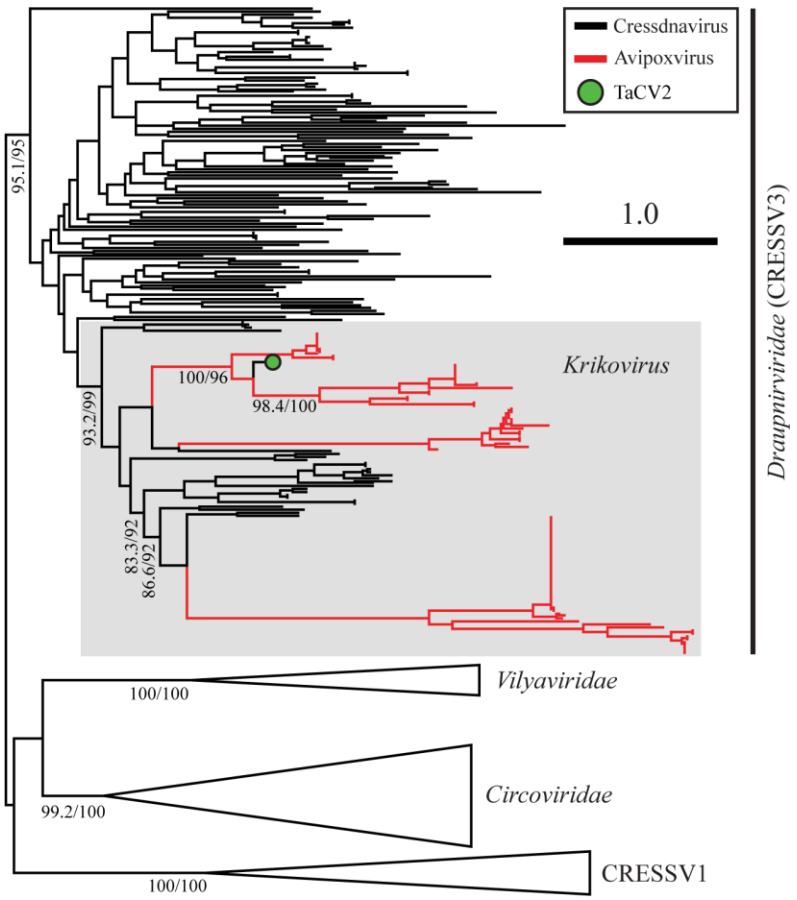


Fig. 4. Maximum-likelihood phylogeny of selected Rep lineages. All apvRep sequences belong to the genus *Krikovirus* (grey box), within the *Draupnirviridae* (i.e., CRESSV3). TaCV2 = Tanager-associated CRESS DNA virus 2 (MF804498.1). Scale bar is in amino acid substitutions per site. Branch supports report Shimodaira Hasegawa approximate likelihood-ratio test (SH-aLRT) scores on the left and ultrafast bootstrap scores on the right.

In line with the step-wise gain of *apvRep* genes already discussed, phylogenetic analysis hinted at multiple independent captures of *Rep* genes by avipoxviruses (Fig. 4). Notably, two Reps from exogenous krikoviruses appeared more closely related to some apvReps than to other krikovirus Reps, suggesting proximity to original donor viruses. These came from Tanager-associated CRESS DNA virus 2 (TaCV2, MF804498.1) and bat circovirus isolate BtPa-CV-3/NX2013 (KJ641729.1). Strikingly, TaCV2 was sampled directly from an avipoxvirus-induced cutaneous lesion on the foot of a bird, *Thraupis episcopus* (37), showing that an interaction between avipoxviruses and krikoviruses is ongoing. The isolate BtPa-CV-3/NX2013 was sequenced from an insectivorous bat, probably from stool. Across the *Krikovirus* genus, a distinct isolation source pattern was apparent; of 28 available genomes, 14 were bat associated (stool or undescribed sample type), six were bird associated (stool and the avipoxvirus-induced lesion), and the remaining eight were insect associated, six from mosquitoes (*SI Appendix*, Table S3). This contrasted with other members of CRESSV3, 76% of which were associated with water or aquatic life. We hypothesize that bat stool-associated krikoviruses represent ingested insect-associated viruses. Krikoviruses in bird stool may also have been ingested, or instead were shed virus. Based on evidence that krikoviruses have an ancient yet ongoing relationship with avipoxviruses, we hypothesized that they share saurian hosts, since HGT would require coinfection of the same host. Alternatively, mosquitoes may represent another setting where HGT could have occurred, as they carry both krikoviruses and avipoxviruses (28, 38, 39).

In light of our phylogenetic analyses and investigations into the hosts of both krikoviruses and the wider CRESSV3 lineage (see below), we propose that a cressdnavirus family be created to replace the temporary name CRESSV3. We suggest the name *Draupnirviridae*. The name comes from the ring Draupnir of Norse mythology, said to have multiplied itself every ninth night. It alludes to both the circular viral genome and genome replication. Furthermore, we support the proposal of Garigliany et al. (28), insofar as *Krikovirus* should be an official cressdnavirus genus, and we propose this should be within the *Draupnirviridae*. In this report, we use the name *Draupnirviridae* instead of CRESSV3 hereafter.

Draupnirviruses Have Infected Saurians for Millions of Years.

We hypothesized that krikoviruses (family *Draupnirviridae*) infect saurian hosts, based on evidence that they donated *Reps* to saurian-infecting avipoxviruses. To explore this, we carried out a detailed search for draupnirvirus-derived EVEs in nearly all available eukaryotic genome assemblies. After quality curation, a total of 145 *Rep*-like and 38 *Cap*-like EVEs were identified, often in long scaffolds or chromosome-resolved assemblies. In line with expectations for EVE sequences (12), the guanine-cytosine (GC) contents of endogenous elements were lower than those of homologous exogenous viruses (*SI Appendix*, Fig. S7). Of the *Rep*-like sequences, 133 belonged to *Krikovirus*, 11 to *Draupnirviridae* (but were not assignable to *Krikovirus*), and the last was discarded due to inconsistent phylogenetic placement (Fig. 5*A*). The krikovirus-like sequences were found

in the genome assemblies of 47 species, all either saurians or insects. These included 37 snakes, 3 lizards, 2 turtles, 4 beetles, and 1 earwig (*SI Appendix*, Table S4). Together, they suggest a saurian and insect host range for krikoviruses, in line with both theoretical prediction and observed isolation sources. The 11 draupnirvirus-like EVE sequences outside of *Krikovirus* were resolved in two sections of the tree. Two short (≤48 aa) sequences came from the snake *Anilios bituberculatus*, and may be phylogenetically misplaced krikoviruses or even circoviruses. The other nine sequences came from *Chromera velia* and *Polymyxa betae*, both of the stramenopiles, alveolates, and rhizarians (SAR) supergroup of protists. This suggests that draupnirviruses outside *Krikovirus* infect various protists, with the ancestral krikoviruses spilling over into animals.

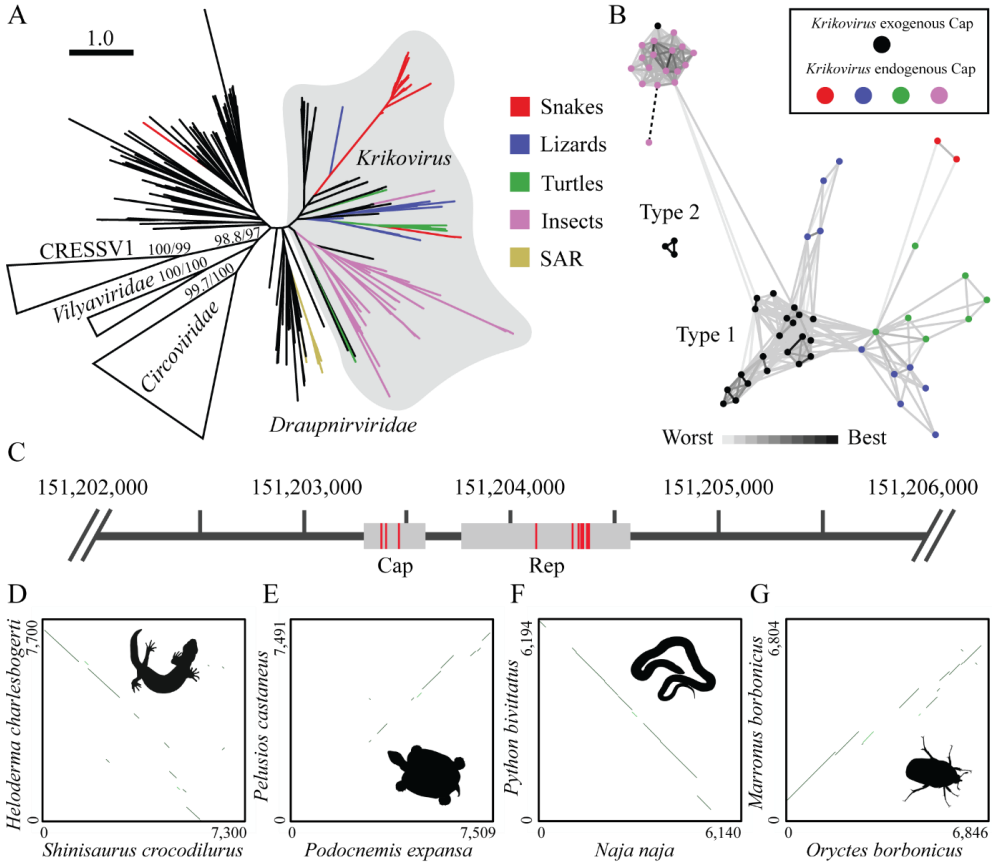


Fig. 5. Saurian genomes contain ancient endogenous krikovirus-derived elements. (*A*) Maximum-likelihood phylogeny of selected Rep protein lineages within *Cressdnaviricota*. CRESSV1, *Vilyaviridae*, and *Circoviridae* serve as outgroups for the *Draupnirviridae*. EVEs extracted from eukaryotic genome assemblies are shown as coloured branches. SAR refers to stramenopiles, alveolates, and rhizarians. Scale bar is in amino acid substitutions per site. Branch supports report SH-aLRT scores on the left and

ultrafast bootstrap scores on the right. (*B*) Clustered krikovirus Caps including exogenous and endogenous sequences. Connections represent BLASTp alignments, with shade denoting significance level (maximum/worst e-value = $1e^{-10}$). The dotted line connects a Cap to known relatives despite no significant BLASTp alignment using CLuster ANalysis of Sequences (CLANS). (*C*) Integration of a complete krikovirus genome into chromosome 2 of *S. crocodilurus* (CM037877.1). Regions with alignment to query krikovirus proteins are shown in grey. Red bars indicate stop codons. (*D*) LAST alignment dotplot of shared krikovirus EVE and sequence context in *S. crocodilurus* (151,200,291-151,207,590 in CM037877.1) and *H. charlesbogerti* (1,842,167-1,849,871 in JANEZZ010002294.1). EVE sequence was masked prior to alignment. (*E*) Shared krikovirus EVE and sequence context in *P. expansa* (26,222,225-26,229,734 in ML681998.1) and *P. castaneus* (5,808,247-5,815,738 in ML685784.1). (*F*) Shared krikovirus EVE and sequence context in *N. naja* (297,176,533-297,182,673 in CM019148.1) and *P. bivittatus* (36,161-42,355 in NW_006537177.1). (*G*) Shared krikovirus EVE and sequence context in *O. borbonicus* (2,594-9,440 in LJIG01000918.1) and *M. borbonicus* (18,407,007-18,413,811 in LR737382.1). Animal silhouettes were retrieved from phylopic.org (*Heloderma* representing *Anguimorpha* by Nicolas Mongiardino Koch, *Stupendemys* representing *Pleurodira* by Roberto Díaz Sibaja, *Rhabdophis* (i.e., *Macropisthodon*) representing *Serpentes* by V. Deepak, and *Cotinis* representing *Dynastinae* by C. Camilo Julián-Caballero). Silhouettes were available in the public domain or under a creative commons license (https://creativecommons.org/licenses/by/3.0).

Cap proteins of sequenced exogenous krikoviruses belong to two lineages; the lineage found in the majority of genomes is *Circoviridae*-like based on HHpred analysis (40) (type 1, e.g., ARO38299.1), while the other has ambiguous ancestry and is found in three genomes (type 2, e.g., QKN88852.1). Of the 38 *Cap*-like EVEs identified, 36 were related to type 1 krikovirus Caps and were found in genomes of snakes, lizards, turtles, and insects (Fig. 5*B*). None were related to type 2 Caps. While related, *Cap*-like EVEs in insect and vertebrates separated during cluster analysis, suggesting some Cap-mediated host tropism may occur, or that there were phylogenetic biases in the progenitor viruses depositing EVEs in respective host lineages. The last two *Cap*-like EVEs belonged to lineages sometimes found with draupnirvirus Reps and were again found in *C. velia* and *P. betae. Krikovirus Cap*-like EVEs were often found paired with *Krikovirus Rep*-like EVEs (17 out of 36 cases), showing they were integrated as whole virus genomes (Fig. 5*C*). EVEs of both gene types regularly contained numerous stop codons, suggestive of ancient origin; for example, one integration in *Shinisaurus crocodilurus* chromosome 2 had three stops in the predicted *Cap* sequence and seven in the *Rep* (Fig. 5*C*).

The indication of ancient krikovirus-derived EVEs led us to explore their possible homology between host species. After masking individual EVE sequences, we carried out all-versus-all alignment of sequence contexts around each integration site, finding many were homologous to each other (*SI Appendix*, Table S4). Time-calibrated host phylogenies allowed some to be temporally constrained. We found that an EVE in the neoanguimorph lizard *Heloderma charlesbogerti* is shared with the paleoanguimorph *S. crocodilurus* (Fig.

5*D*), indicating that integration occurred prior to species divergence during the Cretaceous, ~114 Mya (41). A separate integration observed in the two pleurodiran turtles *Pelusios castaneus* (*Pelomedusidae*) and *Podocnemis expansa* (*Podocnemididae*) (Fig. 5*E*) predated species divergence ~112.5 Mya (42). An integration found shared between many snake species including *Python bivittatus* and *Naja naja* (Fig. 5*F*) can be dated to at least ~65 Mya, before the rapid expansion in snake diversity precipitated by the Cretaceous-Tertiary mass extinction (15, 43). While no time-calibrated phylogeny was available for the *Dynastinae* subfamily of beetles within the *Scarabaeidae*, we did observe homologous integrations, for example between *Marronus borbonicus* and *Oryctes borbonicus* (Fig. 5*G*).

Discussion

Of 57 named cressdnavirus lineages, 14 have known or proposed hosts for some representatives (3, 44), while the rest contain stray viruses with unknown hosts. One of these is the *Circoviridae*, members of which cause severe disease in some vertebrates. Here, we presented the likely hosts of a 15th lineage, which we named as the family *Draupnirviridae*, only the second recognized to have infected vertebrates. Among draupnirviruses, the genus *Krikovirus* is linked by HGT to saurian vertebrates, insects, and strikingly avipoxviruses. Modern infection of saurians remains to be directly confirmed, though identification of TaCV2 in a bird lesion could represent viral shedding during a coinfection with an avipoxvirus (37). The remainder of the *Draupnirviridae* remains poorly characterized, though some detected EVEs suggest protistan hosts. A possible scenario for krikovirus evolution is emergence of protistan viruses into animals such as insects, followed by spillover into vertebrates. We show krikoviruses infected saurians over 100 Mya; therefore, any such spillover was ancient. Hematophagous mosquitoes are at least 100 million years old (45, 46) and can precipitate spillover by vectoring viruses (47). The possibility that mosquitoes spread krikoviruses to vertebrates is concordant with their occurrence in modern mosquitoes, as well as the donation of krikovirus *Reps* to avipoxviruses, also vectored by mosquitoes (38, 39). However, it remains uncertain whether HGT occurred in an insect or a vertebrate host, and mosquito detection may also be derived from bloodmeals. Future work should test vector competency of mosquitoes for krikoviruses and establish whether genome replication occurs in their midgut and salivary glands.

We observed krikovirus-derived *apvReps* across the genus *Avipoxvirus*, large dsDNA pathogens of conservation and animal welfare concern. At least one *apvRep* is found in all sequenced members of the genus except TePV-1, which has likely experienced gene loss. Three additional species closely related to each other are currently undergoing pseudogenization of most of their *apvRep* alleles, and together this shows gene loss is nonlethal. However, near-total *apvRep* presence across avipoxviruses, evidence of RNA expression, and purifying selection on coding sequences all suggest they are active and

generally enhance viral fitness. The observed temporal expansion in *apvRep* gene count alongside genome size is also notable, and it is possible they have contributed to poxvirus adaptive evolution (48). Four of five *apvRep* genes have a simplified domain architecture, a process often seen in horizontally acquired virus genes, hypothesized to mimic or interfere with canonical functions of intact homologs (49). Given that different domains are represented, exact functions are difficult to interpret and likely vary by gene. While not directly comparable, the *Rep* gene *U94* gained by HHV-6 from a parvovirus has diverse co-opted functions, for example in viral latency (50, 51). We observed that in both *apvReps* and *U94*, sequence motifs involved in endonuclease activity are inactivated, while helicase motifs are partly conserved. Instead of canonical cressdnavirus Rep functions in rolling circle replication, it is possible that apvReps have undergone exaptation, or retain nonenzymatic Rep functions. If krikovirus Reps are inhibitory to avipoxvirus replication in the same way the homologous (33) Reps of adeno-associated viruses are to both adenoviruses and herpesviruses (52, 53), then this function may have been co-opted by avipoxviruses for regulation of genome replication, perhaps explaining higher *apvRep* count in longer genomes. Alternatively, apvReps may serve in antiviral defense against coinfecting krikoviruses. If they can disrupt krikovirus Rep complexes via a dominant negative effect, then avipoxviruses could limit krikovirus genome replication and protein production, which might inhibit their own. Devaluing this hypothesis, we found apvRep proteins are likely localized to the cytoplasm, while cressdnavirus Reps localize to the nucleus (32, 54). Krikovirus-derived EVEs in animal genomes show nuclear replication has occurred historically, as would be expected. Our study complements examples of *Rep* capture by dsDNA viruses, yet no study has shown similar evidence for *Cap* transfer. Rather than a mechanistic bias, we suspect this reflects survivorship bias—in that *Rep* is more often advantageous to recipients, and thus maintained during evolution. Given the different localization of genome replication for cressdnaviruses and poxviruses, we suspect HGT is mediated by retrotransposition of single-gene transcripts, recently confirmed to occur in poxviruses (55, 56).

Using the known host range of unrelated viruses, we used interrealm HGT to infer hosts of some stray viruses. We presume such long distance HGT is uncommon; however, we have not exhaustively characterized virus-to-virus HGT events, and thus additional virus–host relationships may be solved by doing so. While we cannot infer whether the first *apvRep* originated in a saurian host or a mosquito, it was apparently gained by an ancestor of all avipoxviruses circulating today. Estimates of ancient divergence times for viral lineages are prone to underestimation due to substitution saturation effects over long timeframes (57, 58), and high levels of genetic saturation have been observed in core poxvirus genes (59). Published estimates of divergence times between CNPV and FPV fall within the last 100 thousand years (60, 61); however, these estimates are liable to increase upon correction for substitution saturation and inclusion of early-branching avipoxvirus genomes such as TePV-1 and TKPV HU1124 (26, 62). Even so, given that krikoviruses first infected saurians over 100 Mya, their interaction with avipoxviruses appears comparatively young.

Further research on this system is warranted to determine any functions of *apvReps*, the nature of krikovirus-avipoxvirus relationships, and the pathogenic potential of krikoviruses.

Materials and Methods

Detection of HGT to the *Poxviridae*.

A computational HGT detection workflow was designed, available from: https://github.com/CormacKinsella/HGT_finder (63). It requires a protein query; we used a phylogenetically broad cressdnavirus database of 2,923 Rep and 2,122 Cap sequences. The other input is a list of genome assemblies to process; ours contained 1,090 poxvirus assemblies available in July 2022 from GenBank and RefSeq databases. The workflow iteratively processes assemblies, handling assembly download, corruption testing, and replacement if necessary. Features aligning to query proteins are identified with tBLASTn (64) with e-value set to $1e^{-5}$, and alignment coordinates are converted to BED format with ascending ranges. Strictly overlapping alignments are merged with BEDTools (65) to generate a minimum–maximum coordinate range for each feature, which is extracted as a nucleotide FASTA. For each feature, the predicted protein sequence is recorded using the single best tBLASTn alignment, which can align past frameshifting mutations and retains stop codons as asterisks. For analysis of proximal or tandem features, a further BED file is generated merging features ≤1 kb apart, while for analysis of feature context, a BED and corresponding FASTA is created allowing ≤1 kb between features and appending 3 kb of sequence context to each end, scaffold length permitting. Finally, the workflow removes unrequired files and proceeds to the next assembly. Potential HGT-derived features were aligned to the GenBank nr database using DIAMOND BLASTp v2.0.15 (66) set to "--ultra-sensitive --max-target-seqs 50" to ensure reciprocal cressdnavirus alignment. Manual curation ensured all avipoxvirus *apvRep* elements were detected and complete, utilizing the NCBI tBLASTn tool and GenBank assembly annotations.

Comparative Genomics and Characterization of *apvRep* Genes.

Comparative genomic analyses used clinker (67). Protein structures were predicted using AlphaFold v2.1.1 (68), aligned using the Protein Data Bank (PDB) pairwise structure alignment tool (69), and visualized using Mol* (70). Presence of apvRep functional domains was assessed at the structural level using AlphaFold predictions, and visually in Jalview (71) after alignment of sequences to the Reps of BFDV (ADN80874.1) and TaCV2 (AVH76405.1). For assessment of sequence motif conservation, sequence logos were generated using WebLogo (72). Possible domains in gene fusion partners were assessed using CD-Search (73). For phylogenetic analyses, regions of apvRep proteins gained by gene fusion were manually trimmed prior to alignment with cressdnavirus references using MAFFT v7.487 (74), and analysis with IQ-TREE v2.2.0 (75). Sequence GC contents were calculated using the geecee tool within EMBOSS v6.6.0.0 (76). To test for selection,

apvRep alleles per gene were first deduplicated at the species level, degraded sequences were discarded, and sequences were aligned using MAFFT as above. Alignments were analyzed using the fixed-effects likelihood method (77), allowing synonymous rate variation and performing bootstrap resampling 100 times. Expression of the *apvRep* genes of CNPV and FPV was assessed using publicly available data (PRJNA524335) (36). RNA-Seq reads were mapped to respective reference genomes (NC_005309 and AJ581527) using Burrows-Wheeler Aligner (BWA) (78), and coverages across *apvRep* genes were plotted.

Phylogenetic Analysis of the *Poxviridae* and *Draupnirviridae*.

Protein sequences of nine conserved genes (major core protein 4a, major core protein 4b, DNA polymerase, RNA polymerase subunit RPO132, RNA polymerase subunit RPO147, messenger RNA capping enzyme catalytic subunit, RNA polymerase-associated protein RAP94, early transcription factor large subunit, and primase D5) were extracted from representatives of all official or proposed genera in the *Poxviridae*. These were concatenated, aligned with MAFFT, and analyzed with IQ-TREE as above. We produced a second tree (not shown) using four core genes (RNA polymerase subunit RPO132, RNA polymerase subunit RPO147, early transcription factor large subunit, and RNA polymerase-associated protein RAP94), previously identified to produce phylogenies consistent with whole genome analyses (59), and this allowed inclusion of additional partial genomes for which all the nine genes were not available. Branch order was observed to be highly consistent between the trees. The same methods were applied to phylogenetic analysis of cressdnaviruses.

Detection and Analysis of EVEs.

Two rounds of EVE discovery were performed using the HGT detection workflow described above. Round one used the same protein query and targeted 24,764 eukaryotic genome assemblies available in GenBank and RefSeq databases in July 2022. Curation of hits began with DIAMOND analysis, this time after removing stop codons. Putative EVEs in contigs <4 kb were discarded, as were those in assemblies apparently containing numerous cressdnaviral lineages, which raised suspicion of contamination. Round two used a protein query comprehensively covering draupnirviruses and any draupnirvirus-related EVEs from the first round. We now targeted 6,639 assemblies, with an inclusive focus on apparent host lineages of *Draupnirviridae* (e.g., Sauropsida, Insecta, and SAR). Quality control of candidate EVEs was as above. Cluster analysis of Cap sequences was done using CLANS (79). To determine homology between curated elements, sequence contexts were masked for EVE sequences using maskfasta within BEDTools v2.27.1 (65) and all-versus-all aligned using LAST v1422 (80). Self-alignments were removed, along with any alignment below 300 bp in length (query strand) or with e-value over $1e^{-50}$. To ensure cutoff suitability, manual curation of selected low scoring alignments was done using D-GENIES v1.4 (81), which was also used to produce alignment plots.

Data, Materials, and Software Availability

All genome assemblies and datasets analyzed here are available in public databases. The computational workflow for HGT detection is available at: https://github.com/CormacKinsella/HGT_finder (63). Predicted protein structures described here are available at: https://figshare.com/projects/RepStructures/158462 (82). All other data are included in the manuscript and/or *SI Appendix*.

Acknowledgments

Computational work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. We thank Tim C. Passchier and Jacopo Martelossi for helpful discussions. This work was supported by a grant from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie agreement No. 721367, host switching pathogens, infectious outbreaks and zoonosis (HONOURs), awarded to L.v.d.H.

References

1. Tisza, M. J. et al. Discovery of several thousand highly diverse circular DNA viruses. Elife 9, e51971 (2020).

2. Edgar, R. C. et al. Petabase-scale sequence alignment catalyses viral discovery. Nature 602, 142–147 (2022).

3. Kinsella, C. M. et al. Host prediction for disease-associated gastrointestinal cressdnaviruses. Virus Evol. 8, 1–11 (2022).

4. Krupovic, M. et al. Cressdnaviricota: A virus phylum unifying seven families of Rep-encoding viruses with single-stranded, circular DNA genomes. J. Virol. 94, e00582-20 (2020).

5. Walker, P. J. et al. Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). Arch. Virol. 167, 2429–2440 (2022).

6. Li, L. et al. Exploring the virome of diseased horses. J. Gen. Virol. 96, 2721–2733 (2015).

7. Kazlauskas, D., Varsani, A. & Krupovic, M. Pervasive chimerism in the replication-associated proteins of uncultured single-stranded DNA viruses. Viruses 10, v10040187 (2018).

8. Ritchie, B. W., Niagro, F. D., Lukert, P. D., Steffens, W. L. & Latimer, K. S. Characterization of a new virus from cockatoos with psittacine beak and feather disease. Virology 171, 83–88 (1989).

9. Ellis, J. et al. Isolation of circovirus from lesions of pigs with postweaning multisystemic wasting syndrome. Can. Vet. J. 39, 44–51 (1998).

10. Tischer, I., Rasch, R. & Tochtermann, G. Characterization of papovavirus and picornavirus-like particles in permanent pig kidney cell lines. Zentbl. Bakt. 226, 153–167 (1974).

11. Katzourakis, A. & Gifford, R. J. Endogenous viral elements in animal genomes. PLoS Genet. 6, e1001191 (2010).

12. Kinsella, C. M. et al. Entamoeba and Giardia parasites implicated as hosts of CRESS viruses. Nat. Commun. 11, 1–10 (2020).

13. Patterson, Q. M. et al. Circoviruses and cycloviruses identified in Weddell seal fecal samples from McMurdo Sound, Antarctica. Infect. Genet. Evol. 95, 105070 (2021).

14. Dennis, T. P. W. et al. The evolution, distribution and diversity of endogenous circoviral elements in vertebrate genomes. Virus Res. 262, 15–23 (2019).

15. Klein, C. G. et al. Evolution and dispersal of snakes across the Cretaceous-Paleogene mass extinction. Nat. Commun. 12, 1–9 (2021).

16. Thomson, B. J., Efstathiou, S. & Honess, R. W. Acquisition of the human adeno-associated virus type-2 rep gene by human herpesvirus type-6. Nature 351, 78–80 (1991).

17. Diemer, G. S. & Stedman, K. M. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. Biol. Direct 7, 1–14 (2012).

18. La Scola, B. et al. The virophage as a unique parasite of the giant mimivirus. Nature 455, 100–104 (2008).

19. Zhang, H. et al. A novel bat herpesvirus encodes homologues of major histocompatibility complex classes I and II, C-type lectin, and a unique family of immune-related genes. J. Virol. 86, 8014–8030 (2012).

20. Vink, C., Beuken, E. & Bruggeman, C. A. Complete DNA sequence of the rat cytomegalovirus genome. J. Virol. 74, 7656–7665 (2000).

21. Caprari, S., Metzler, S., Lengauer, T. & Kalinina, O. V. Sequence and structure analysis of distantly-related viruses reveals extensive gene transfer between viruses and hosts and among viruses. Viruses 7, 5388–5409 (2015).

22. Trempe, F. et al. Characterization of human herpesvirus 6A/B U94 as ATPase, helicase, exonuclease and DNA-binding proteins. Nucleic Acids Res. 43, 6084–6098 (2015).

23. Tulman, E. R. et al. The genome of canarypox virus. J. Virol. 78, 353–366 (2004).

24. Boyle, D. B. Genus Avipoxvirus. in Poxviruses (eds. Mercer, A. A., Schmidt, A. & Weber, O.) 217–251 (Birkhäuser, 2007).

25. Sarker, S., Hannon, C., Athukorala, A. & Bielefeldt-Ohmann, H. Emergence of a novel pathogenic poxvirus infection in the endangered green sea turtle (Chelonia mydas) highlights a key threatening process. Viruses 13, v13020219 (2021).

26. Seitz, K. et al. Discovery of a phylogenetically distinct poxvirus in diseased Crocodilurus amazonicus (family Teiidae). Arch. Virol. 166, 1183–1191 (2021).

27. Li, L. et al. Bat guano virome: Predominance of dietary viruses from insects and plants plus novel mammalian viruses. J. Virol. 84, 6955–6965 (2010).

28. Garigliany, M. M. et al. Characterization of a novel circo-like virus in Aedes vexans mosquitoes from germany: Evidence for a new genus within the family Circoviridae. J. Gen. Virol. 96, 915–920 (2015).

29. Luo, G. et al. Crystal structure of the dimerized N terminus of porcine circovirus type 2 replicase protein reveals a novel antiviral interface. J. Virol. 92, e00724-18 (2018).

30. Vega-Rocha, S., Byeon, I. J. L., Gronenborn, B., Gronenborn, A. M. & Campos-Olivas, R. Solution structure, divalent metal and DNA binding of the endonuclease domain from the replication initiation protein from porcine circovirus 2. J. Mol. Biol. 367, 473–487 (2007).

31. Tarasova, E., Dhindwal, S., Popp, M., Hussain, S. & Khayat, R. Mechanism of dna interaction and translocation by the replicase of a circular Rep-encoding single-stranded dna virus. MBio 12, e00763-21 (2021).

32. Lin, W. L., Chien, M. S., Du, Y. W., Wu, P. C. & Huang, C. The N-terminus of porcine circovirus type 2 replication protein is required for nuclear localization and ori binding activities. Biochem. Biophys. Res. Commun. 379, 1066–1071 (2009).

33. Kazlauskas, D., Varsani, A., Koonin, E. V. & Krupovic, M. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. Nat. Commun. 10, 1–12 (2019).

34. Chandler, M. et al. Breaking and joining single-stranded DNA: The HUH endonuclease superfamily. Nature Reviews Microbiology 11, 525–538 (2013).

35. Hanson, P. I. & Whiteheart, S. W. AAA+ proteins: Have engine, will work. Nat. Rev. Mol. Cell Biol. 6, 519–529 (2005).

36. Giotis, E. S., Montillet, G., Pain, B. & Skinner, M. A. Chicken embryonic-stem cells are permissive to poxvirus recombinant vaccine vectors. Genes 10, 237 (2019).

37. Moens, M. A. J., Pérez-Tris, J., Cortey, M. & Benítez, L. Identification of two novel CRESS DNA viruses associated with an Avipoxvirus lesion of a blue-and-gray Tanager (Thraupis episcopus). Infect. Genet. Evol. 60, 89–96 (2018).

38. Kligler, I. J., Muckenfuss, R. S. & Rivers, T. M. Transmission of fowl-pox by mosquitoes. J. Exp. Med. 49, 649–660 (1929).

39. Akey, B. L., Nayar, J. K. & Forrester, D. J. Avian pox in Florida wild turkeys: Culex nigripalpus and Wyeomyia vanduzeei as experimental vectors. J. Wildl. Dis. 17, 597–599 (1981).

40. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 33, W244–W248 (2005).

41. Zheng, Y. & Wiens, J. J. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. Mol. Phylogenet. Evol. 94, 537–547 (2016).

42. Shaffer, H. B., McCartney-Melstad, E., Near, T. J., Mount, G. G. & Spinks, P. Q. Phylogenomic analyses of 539 highly informative loci dates a fully resolved time tree for the major clades of living turtles (Testudines). Mol. Phylogenet. Evol. 115, 7–15 (2017).

43. Irisarri, I. et al. Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nat. Ecol. Evol. 1, 1370–1378 (2017).

44. Gronenborn, B., Randles, J., HJ, V. & Thomas, J. Create one new family (Metaxyviridae) with one new genus (Cofodevirus) and one species (Coconut foliar decay virus) moved from the family Nanoviridae (Mulpavirales). Int. Comm. Taxon. Viruses Propos. number 2020.022P (2021).

45. Borkent, A. & Grimaldi, D. A. The earliest fossil mosquito (Diptera: Culicidae), in mid-Cretaceous Burmese amber. Ann. Entomol. Soc. Am. 97, 882–888 (2004).

46. Labandeira, C. C. & Li, L. The history of insect parasitism and the mid-Mesozoic parasitoid revolution. in The Evolution and Fossil Record of Parasitism (eds. De Baets, K. & Huntley, J. W.) 377–533 (Springer, 2021).

47. Rückert, C. & Ebel, G. D. How do virus–mosquito interactions lead to viral emergence? Trends Parasitol. 34, 310–321 (2018).

48. McLysaght, A., Baldi, P. F. & Gaut, B. S. Extensive gene gain associated with adaptive evolution of poxviruses. Proc. Natl. Acad. Sci. 100, 15655–15660 (2003).

49. Rappoport, N. & Linial, M. Viral proteins acquired from a host converge to simplified domain architectures. PLOS Comput. Biol. 8, e1002364 (2012).

50. Caselli, E. et al. The U94 gene of human herpesvirus 6: A narrative review of its role and potential functions. Cells 9, cells9122608 (2020).

51. Caselli, E. et al. Human herpesvirus 6 (HHV-6) U94/REP protein inhibits betaherpesvirus replication. Virology 346, 402–414 (2006).

52. Glauser, D. L. et al. Inhibition of herpes simplex virus type 1 replication by adeno-associated virus rep proteins depends on their combined DNA-binding and ATPase/helicase activities. J. Virol. 84, 3808–3824 (2010).

53. Needham, P. G. et al. Adeno-associated virus Rep protein-mediated inhibition of transcription of the adenovirus major late promoter in vitro. J. Virol. 80, 6207–6217 (2006).

54. Meerts, P., Misinzo, G., McNeilly, F. & Nauwynck, H. J. Replication kinetics of different porcine circovirus 2 strains in PK-15 cells, fetal cardiomyocytes and macrophages. Arch. Virol. 150, 427–441 (2005).

55. Rahman, M. J. et al. LINE-1 retrotransposons facilitate horizontal gene transfer into poxviruses. Elife 11, 63327 (2022).

56. Fixsen, S. M. et al. Poxviruses capture host genes by LINE-1 retrotransposition. Elife 11, 63332 (2022).

57. Aiewsakun, P. & Katzourakis, A. Time-dependent rate phenomenon in viruses. J. Virol. 90, 7184–7195 (2016).

58. Ghafari, M., Simmonds, P., Pybus, O. G. & Katzourakis, A. A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses. Curr. Biol. 31, 4689-4696.e5 (2021).

59. Yu, Z. et al. Genomic analysis of Poxviridae and exploring qualified gene sequences for phylogenetics. Comput. Struct. Biotechnol. J. 19, 5479–5486 (2021).

60. Babkin, I. V. & Babkina, I. N. Molecular dating in the evolution of vertebrate poxviruses. Intervirology 54, 253–260 (2011).

61. Le Loc'h, G., Bertagnoli, S. & Ducatez, M. F. Time scale evolution of avipoxviruses. Infect. Genet. Evol. 35, 75–81 (2015).

62. Bányai, K. et al. Unique genomic organization of a novel Avipoxvirus detected in turkey (Meleagris gallopavo). Infect. Genet. Evol. 35, 221–229 (2015).

63. C. M. Kinsella, Horizontal gene transfer finder. GitHub. https://github.com/CormacKinsella/HGT_finder. Deposited 11 January 2023.

64. Camacho, C. et al. BLAST+: Architecture and applications. BMC Bioinformatics 10, 421 (2009).

65. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).

66. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat. Methods 18, 366–368 (2021).

67. Gilchrist, C. L. M. & Chooi, Y. H. Clinker & clustermap.js: Automatic generation of gene cluster comparison figures. Bioinformatics 37, 2473–2475 (2021).

68. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).

69. Prlić, A. et al. Pre-calculated protein structure alignments at the RCSB PDB website. Bioinformatics 26, 2983–2985 (2010).

70. Sehnal, D. et al. Mol* Viewer: Modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Res. 49, W431–W437 (2021).

71. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189–1191 (2009).

72. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: A sequence logo generator. Genome Res. 14, 1188–1190 (2004).

73. Marchler-Bauer, A. & Bryant, S. H. CD-Search: Protein domain annotations on the fly. Nucleic Acids Res. 32, W327-331 (2004).

74. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. 20, 1160–1166 (2017).

75. Minh, B. Q. et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37, 1530–1534 (2020).

76. Rice, P., Longden, L. & Bleasby, A. EMBOSS: The European molecular biology open software suite. Trends in Genetics 16, 276–277 (2000).

77. Pond, S. L. K. & Frost, S. D. W. Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 22, 1208–1222 (2005).

78. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN] (2013).

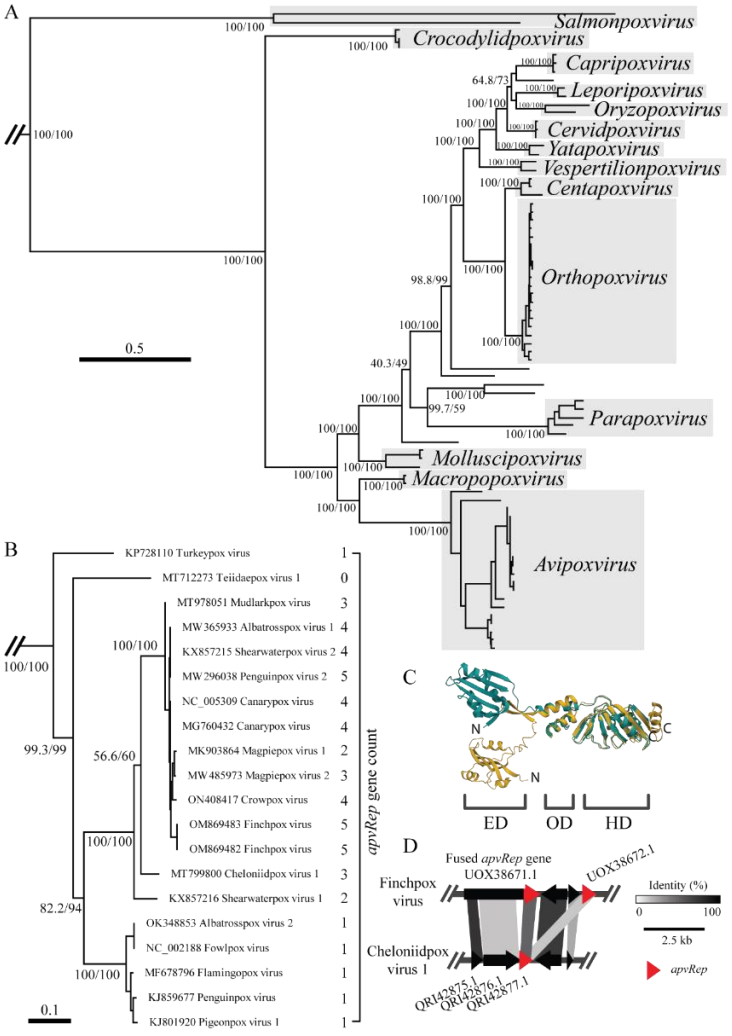
79. Frickey, T. & Lupas, A. CLANS: A Java application for visualizing protein families based on pairwise similarity. Bioinformatics 20, 3702–3704 (2004).

80. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. Genome Res. 21, 487–493 (2011).

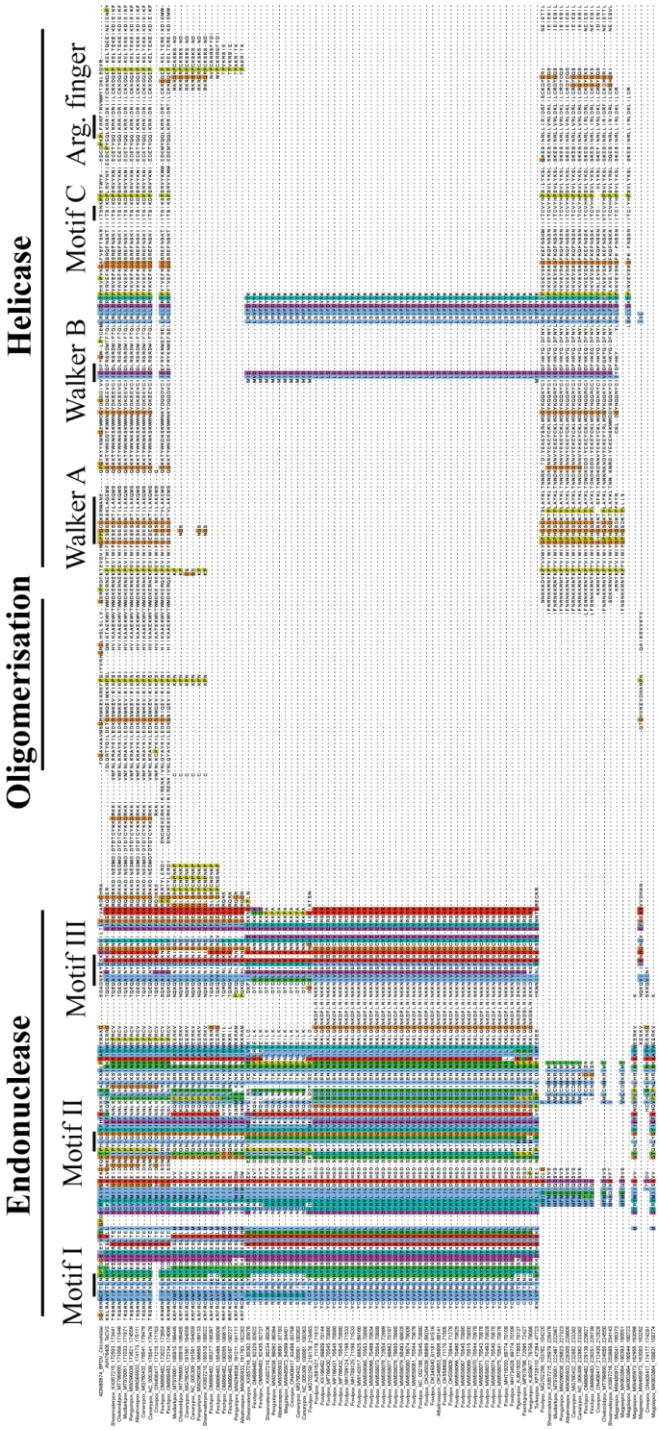
81. Cabanettes, F. & Klopp, C. D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. PeerJ 2018, e4958 (2018).
82. C. M. Kinsella, Predicted Rep structures. Figshare. https://figshare.com/projects/RepStructures/158462. Deposited 31 January 2023.

Supplementary figures

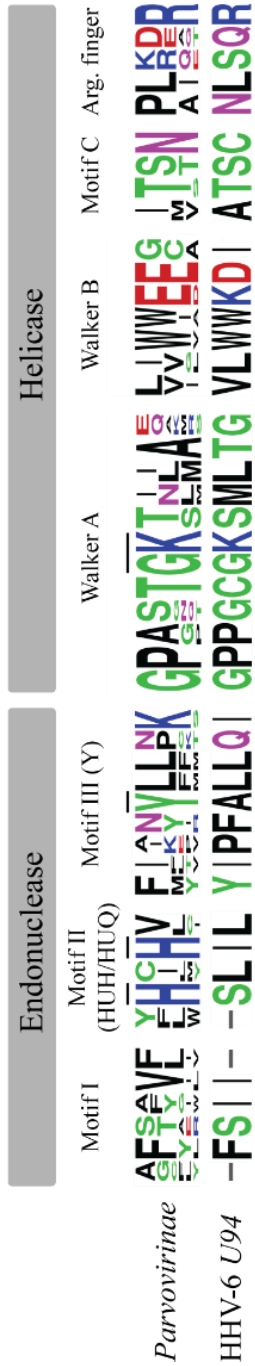
For supplementary tables, see the online version: https://doi.org/10.1073/pnas.2303844120.



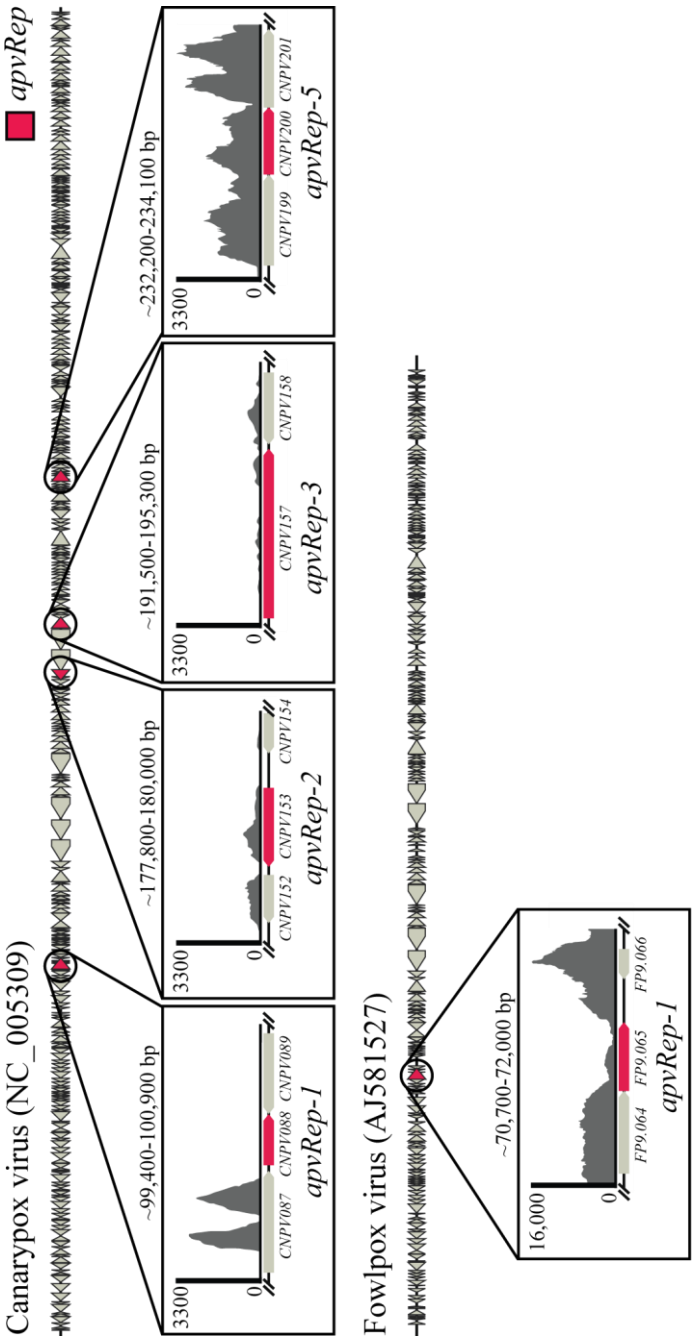
Supplementary Figure 1. A) Rooted maximum-likelihood phylogeny of the subfamily *Chordopoxvirinae*. Proposed genera with only one member are not labelled. Scale bar is in amino acid substitutions per site. Branch supports report SH-aLRT scores on the left and ultrafast bootstrap scores on the right. *Entomopoxvirinae* sequences were used for an outgroup (not shown). B) Focus on the genus *Avipoxvirus* within the same phylogeny. The *apvRep* gene count per-genome is shown at right. C) Superposed alignment between AlphaFold predicted structures of TM-6c Rep (cyan) and CNPV153 (gold) performed with the jFATCAT rigid algorithm; root-mean-square deviation = 1.61, TM-score = 0.53, query (TM-6c) coverage 65%, target coverage 55%. ED = endonuclease domain, OD = oligomerisation domain, HD = helicase domain. N and C denote the respective termini. D) Focused section of a synteny map between finchpox virus (OM869483.1) and cheloniidpox virus 1 (MT799800.1), revealing gene fusion has occurred in the former.



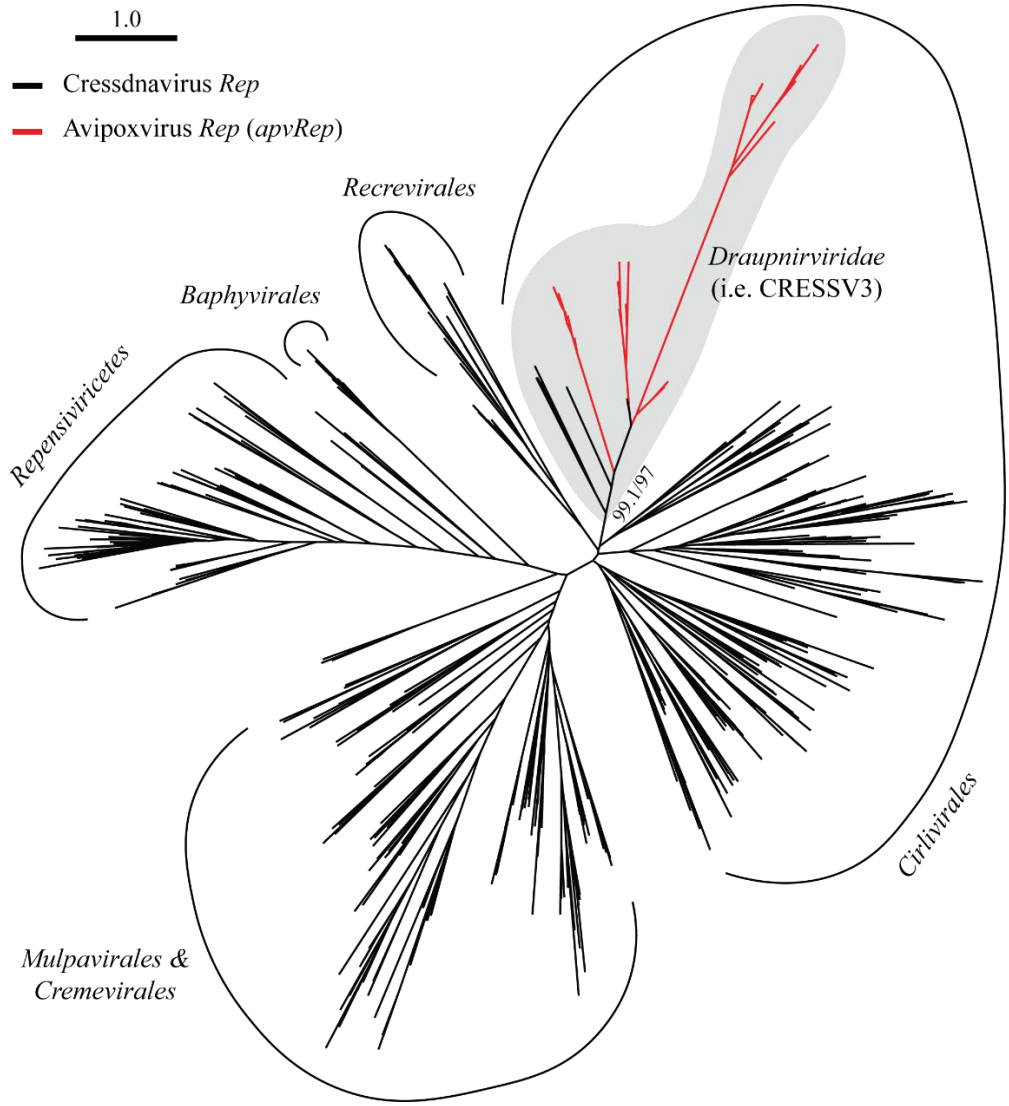
Supplementary Figure 2. Alignment between Reps of *Circoviridae*, *Krikovirus*, and predicted protein sequence of *apvRep* genes. Protein domains are annotated above.



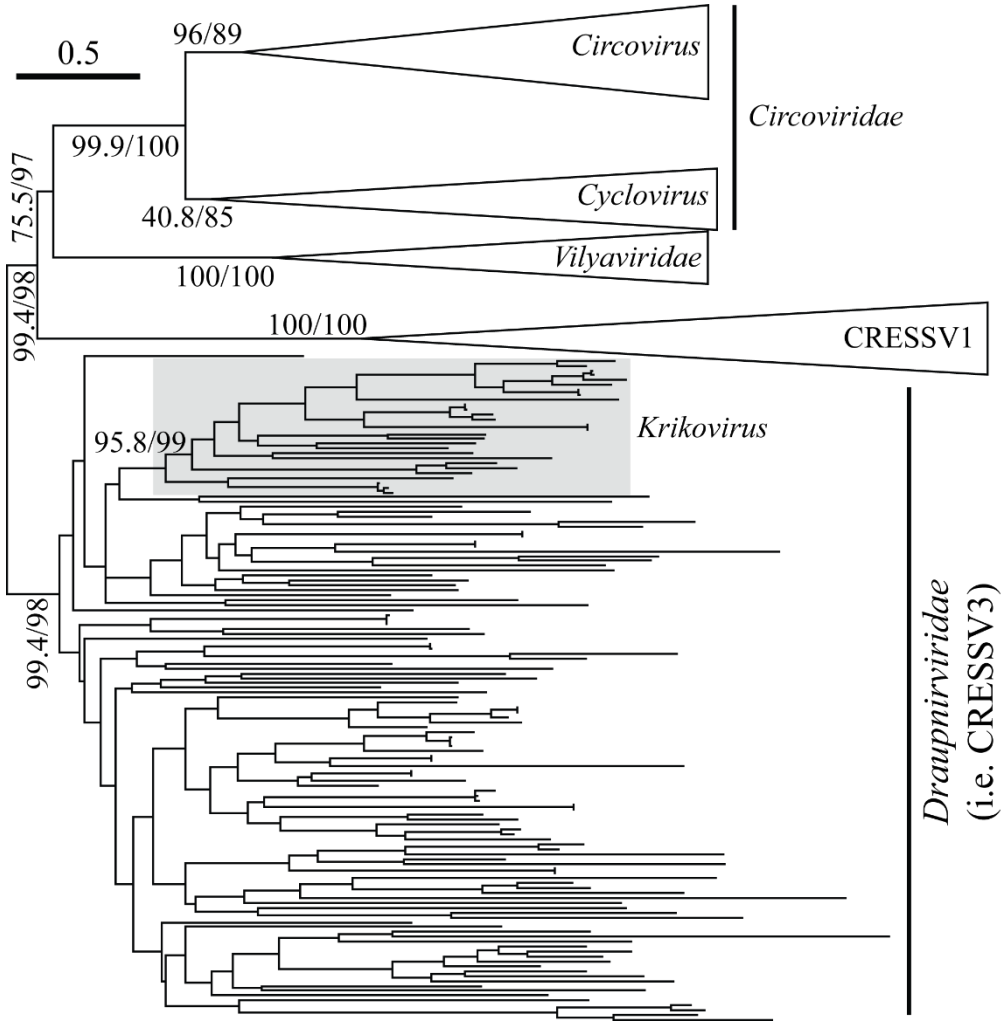
Supplementary Figure 3. Rep protein sequence motifs in the *Parvoviridae* (subfamily *Parvovirinae*), and the *U94* gene of human herpesvirus 6 (AVK93697.1). Arg. = arginine. Residue colours: hydrophobic = black, polar = green, basic = blue, acidic = red, neutral = purple. Key residues discussed in the main text are marked.



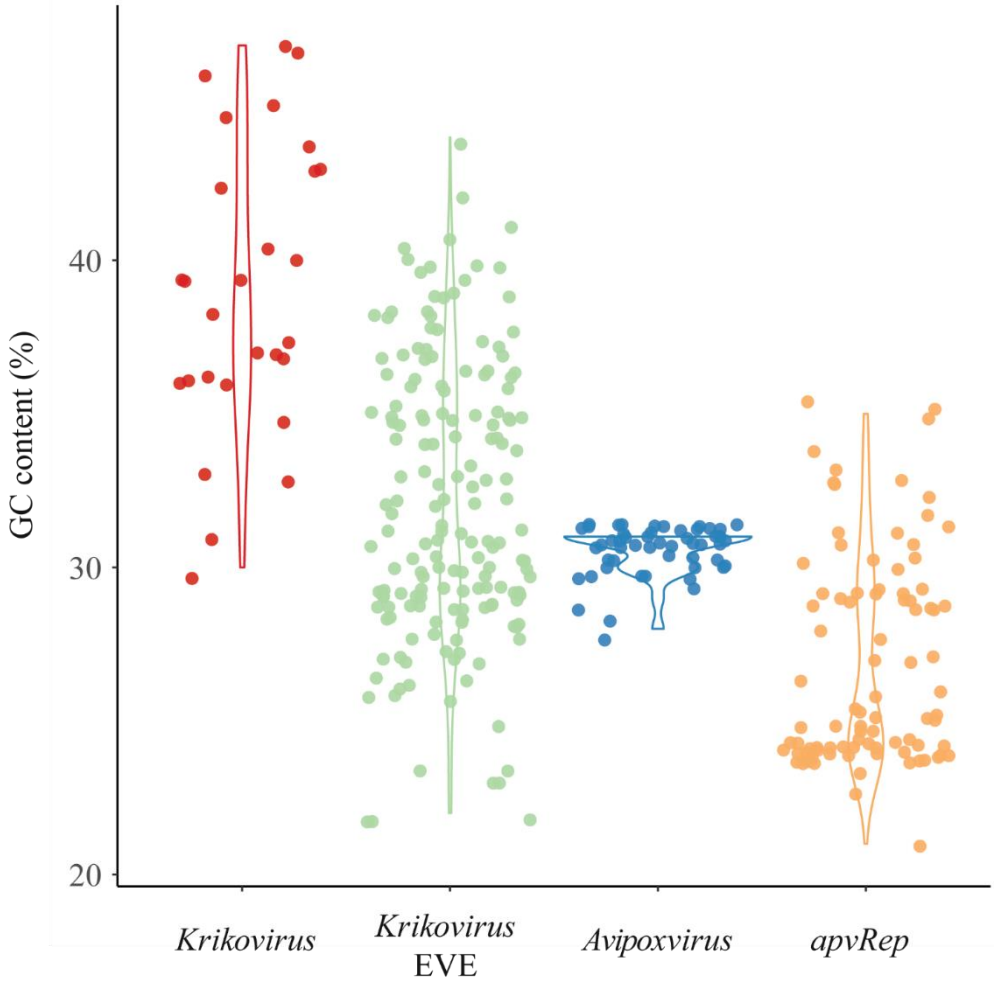
Supplementary Figure 4. Expression of *apvRep* genes in canarypox virus and fowlpox virus infected chicken embryonic stem cells. The y-axes show read coverage per site. At the 16-hour timepoint, there is no evidence of *apvRep-1* expression in canarypox virus, but other *apvRep* genes are expressed at different degrees. At 16-hours, *apvRep-1* is expressed by fowlpox virus.



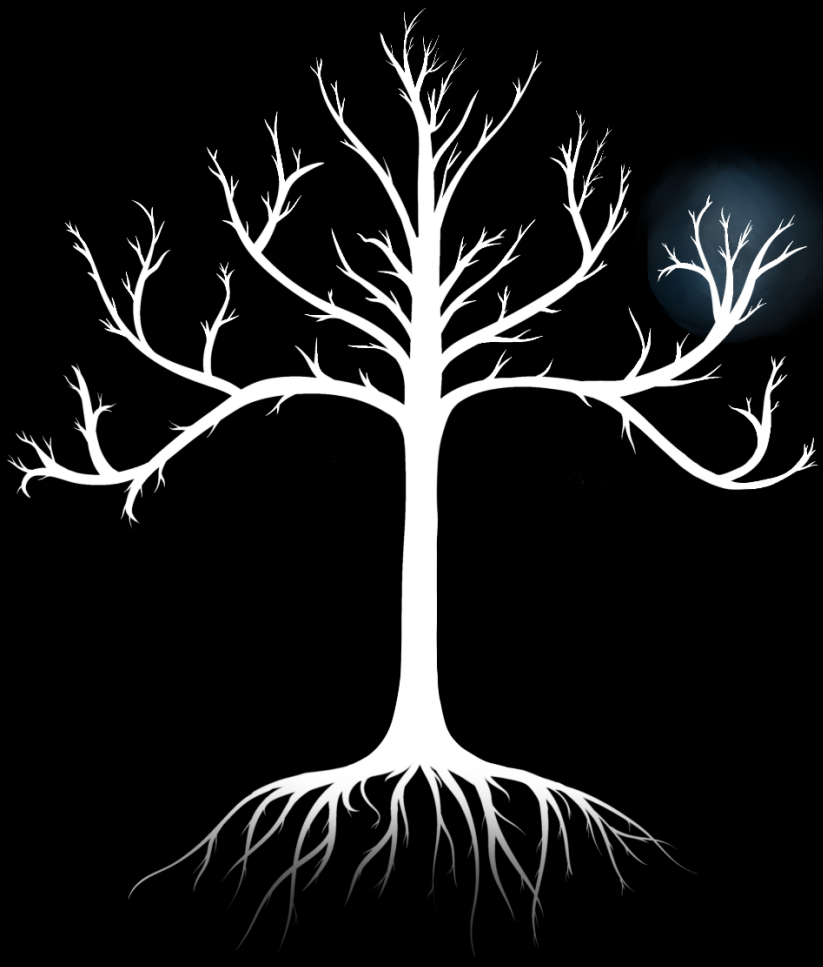
Supplementary Figure 5. Maximum-likelihood phylogeny of representative Rep sequences across the *Cressdnaviricota*, with the addition of *apvRep* predicted sequences. Scale bar is in amino acid substitutions per site. Branch support reports SH-aLRT scores on the left and ultrafast bootstrap scores on the right.



Supplementary Figure 6. Maximum-likelihood phylogeny of selected cressdnavirus Rep lineages. Scale bar is in amino acid substitutions per site. Branch supports report SH-aLRT scores on the left and ultrafast bootstrap scores on the right.



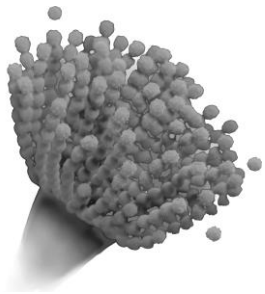
Supplementary Figure 7. GC contents of various sequences, including whole genomes of krikoviruses and avipoxviruses, krikovirus EVEs in animal genomes, and *apvRep* alleles in avipoxvirus genomes.



Chapter 6

Human clinical isolates of pathogenic fungi are host to diverse mycoviruses

Cormac M. Kinsella, Martin Deijs, H. M. Gittelbauer, Lia van der Hoek, Karin van Dijk



Microbiology Spectrum, 2022

https://doi.org/10.1128/spectrum.01610-22

Abstract

Fungi host viruses from many families, and next-generation sequencing can be used to discover previously unknown genomes. Some fungus-infecting viruses (mycoviruses) confer hypovirulence on their pathogenic hosts, raising the possibility of therapeutic application in the treatment of fungal diseases. Though all fungi probably host mycoviruses, many human pathogens have none documented, implying the mycoviral catalogue remains at an early stage. Here, we carried out virus discovery on 61 cultures of pathogenic fungi covering 27 genera and at least 56 species. Using next-generation sequencing of total nucleic acids, we found no DNA viruses but did find a surprising RNA virus diversity of 11 genomes from six classified families and two unclassified lineages, including eight genomes likely representing new species. Among these was the first jivivirus detected in a fungal host (*Aspergillus lentulus*). We separately utilized rolling circle amplification and next-generation sequencing to identify ssDNA viruses specifically. We identified 13 new cressdnaviruses across all libraries, but unlike the RNA viruses, they could not be confirmed by PCR in either the original unamplified samples or freshly amplified nucleic acids. Their distributions among sequencing libraries and inconsistent detection suggest low-level contamination of reagents. This highlights both the importance of validation assays and the risks of viral host prediction on the basis of highly amplified sequencing libraries. Meanwhile, the detected RNA viruses provide a basis for experimentation to characterize possible hypovirulent effects, and hint at a wealth of uncharted viral diversity currently frozen in biobanks.

Introduction

The risk of life-threatening invasive fungal infections (IFIs) has been growing for decades (1, 2), partly due to use of immunosuppressive drugs and chemotherapy, though the cause is thought to be multifactorial (3). A shifting epidemiology has also been observed, with once-rare pathogens becoming significant concerns (4). While historically, *Candida* yeasts and *Aspergillus* moulds have caused the majority of IFIs in cancer patients and hematopoietic stem cell transplant recipients, recently *Rhizopus*, *Mucor*, *Fusarium*, and others have emerged as threats (3–5). Next to that, disseminated infections with dimorphic environmental fungi such as *Histoplasma capsulatum* and *Blastomyces dermatitidis* have been described in neonates and immunocompromised patients within regions of endemicity (6). Meningoencephalitis due to *Cryptococcus neoformans* and *Cryptococcus gattii* is often seen in HIV-positive patients (7), while the recent COVID-19 pandemic has had a tangible impact, with COVID-19-associated pulmonary aspergillosis described in around 30% of ICU patients (8, 9). In Italy and Brazil, up to a 10-fold increase in candidemia has been reported in patients with COVID-19 (10, 11), and in India, a high incidence of mucormycosis is found in COVID-19 patients, with a mortality rate of 35% (12).

Fungal infections are difficult to treat, with only a few options available and resistance to these emerging (e.g., azole resistance in invasive *Aspergillus fumigatus*), while emerging species may be unaffected by standard empirical treatments (3). Available antifungal drugs also have marked side effects, like nephrotoxicity and hepatotoxicity. New treatments that can combat fungal infections would be valuable. Like any living creature, fungi are susceptible to viral infection. Viruses of fungi are called mycoviruses, though this term encompasses a massive genetic diversity spanning viruses of many lineages (13). Known mycoviruses almost all have RNA genomes, though recently three ssDNA mycoviruses of the family *Genomoviridae* were identified (14–16), and endogenous viral elements found in fungal genomes hint that other species of this family also infect fungi (17). If infection by a virus slows or halts growth of a fungal pathogen, for example via cell lysis, this can cause reduced virulence during infection (hypovirulence). Hypovirulence-associated viruses may provide future options for antifungal therapies, and mycoviruses are in fact already applied in the biological control of fungal phytopathogens. An example is chestnut blight, a disease of chestnut trees caused by the fungus *Cryphonectria parasitica*. Infected trees can be treated with the RNA virus Cryphonectria hypovirus 1 (CHV1), resulting in a significantly reduced virulence of the fungus (18, 19). Another example is treatment of *Sclerotinia sclerotiorum* infection of plants using the virus Sclerotinia sclerotiorum hypovirulence-associated DNA virus 1 (SsHADV-1) (14). Infecting the fungus on rapeseed plants reduces disease severity and enhances the rapeseed yield (20). Another mycovirus treatment under development for plants is Rosellinia necatrix megabirnavirus 1 (RnMBV1), which infects the fungal species causing white root rot of fruit trees (21).

Not all mycoviruses are viable as biological control agents, especially for fungal pathogens of humans. It may be important that the virus is not recognized by the innate immune system (e.g., Toll-like receptors [22]) or the adaptive immune system, and thus a low viral antigenicity is preferable. Viruses should be deliverable to a target fungus in the patient using application techniques such as injection or topical administration. For this, an extracellular phase would be ideal, but the majority of known mycoviruses lack one (23, 24). Without an extracellular phase, possible options for virus infection are more complex; for example, hyphal anastomosis between the patient strain and a virus-infected conspecific strain would require addition of a hypovirulent fungus (e.g., infected conidia), which is unlikely a viable approach in human patients. Notably, SsHADV-1 was both the first known DNA mycovirus and the first mycovirus confirmed to have an extracellular phase (24). This raises the possibility that ssDNA viruses represent the likeliest candidates for therapeutic applications in humans (25). In order to identify unknown mycoviruses that may have utility in future therapeutics, we investigated cultured clinical isolates of human-pathogenic fungi, using virus discovery cDNA-AFLP (VIDISCA) next-generation sequencing and Illumina sequencing with or without initial rolling circle amplification.

Results

RNA virus discovery in clinical isolates.

The 61 initial fungal samples belonged to 27 genera and at least 56 species (Table S1 in the supplemental material). VIDISCA sequencing produced an average of 7,600 reads per sample. Bioinformatic analysis identified six samples containing between 1 and 1,306 RNA virus reads belonging to distinct viral lineages. To recover and characterize full genomes of these viruses, Illumina reads were generated (an average of 3.5 million reads per sample after quality control) and assembled from the six positive samples plus one without detected viruses to serve as a control of screening sensitivity. Sequenced isolates belonged to *Rhizopus microsporus*, *R. oryzae*, *Syncephalastrum racemosum*, *Aspergillus niger*, *A. lentulus*, *Cladosporium sphaerospermum*, and *Penicillium vanoranjei*. Surprisingly, upon analysis we found that all resulting assemblies, including the control, contained at least one RNA virus, and two had mixed infections with three viral species each (Table 1). A total of 11 RNA viruses were therefore found instead of the six expected. BLASTx searches showed these belonged to the six families *Partitiviridae*, *Narnaviridae*, *Totiviridae*, *Mitoviridae*, *Endornaviridae*, and *Botourmiaviridae*, plus two unclassified lineages: a jivivirus related to the family *Virgaviridae* and a ribovirus with uncertain relationships (Fig. 1). PCR screening confirmed the presence of all RNA viruses in their respective index samples, and even detected additional positive samples for four of the viruses (Table S3). Three of these additional detections were made in fungi belonging to the same genus as the index (*Rhizopus*, *Aspergillus*, and *Rhizomucor*), but a different species. The last was made in a different genus (*Trichophyton*) to the index (*Aspergillus*), both of which are in the class *Eurotiomycetes*.

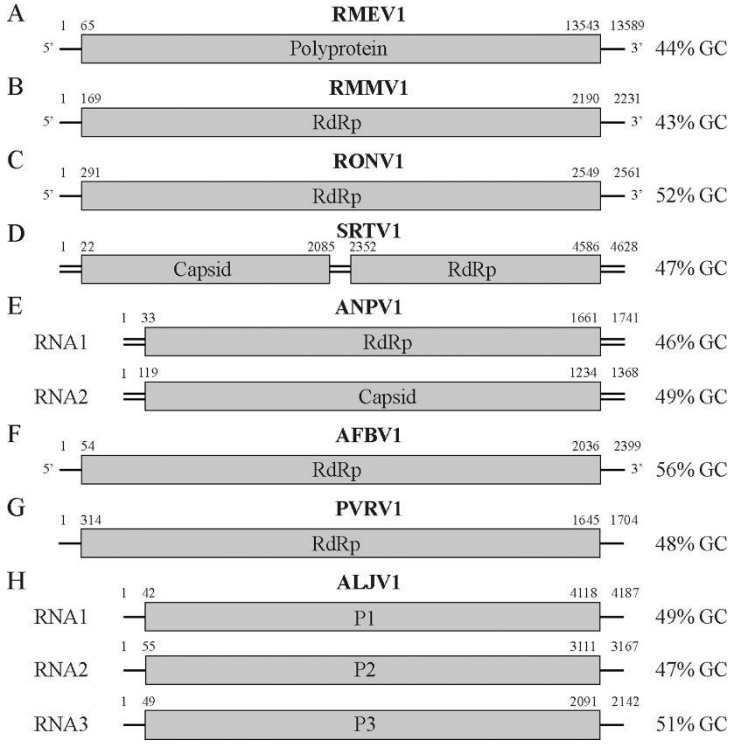


FIG 1 Genome organization of representative RNA viruses for each identified taxonomic group. (A) *Endornaviridae*, Rhizopus microsporus endornavirus 1 (RMEV1, LC671616). (B) *Mitoviridae*, Rhizopus microsporus mitovirus 1 (RMMV1, LC671615). (C) *Narnaviridae*, Rhizopus oryzae narnavirus 1 (RONV1, LC671613). (D) *Totiviridae*, Syncephalastrum racemosum totivirus 1 (SRTV1, LC671614). (E) *Partitiviridae*, Aspergillus niger partitivirus 1 (ANPV1, LC671611 and LC671612). (F) *Botourmiaviridae*, Aspergillus fumigatus botourmiavirus 1 (AFBV1, LC671624). (G) Unclassified, Penicillium vanoranjei associated RNA virus 1 (PVRV1, LC671619). (H) Unclassified, Aspergillus lentulus jivivirus 1 (ALJV1, LC671620, LC671621, and LC671622). Viral sense is not shown for dsRNA viruses and those with unknown sense. Genome sizes are not drawn to scale.

TABLE 1 RNA viruses metagenomically sequenced from clinical isolates of fungi

Sample	Host	Virus family	Virus genus	Putative viral genome	Molecule	Accession
4	*A. niger*	*Partitiviridae*	*Gammapartitivirus*	Aspergillus niger partitivirus 1	dsRNA	LC671611 LC671612
11	*R. oryzae*	*Narnaviridae*	Unclassified	Rhizopus oryzae narnavirus 1	ssRNA(+)	LC671613
13	*S. racemosum*	*Totiviridae*	*Totivirus*	Syncephalastrum racemosum totivirus 1	dsRNA	LC671614
15	*R. microsporus*	*Endornaviridae*	*Alphaendornavirus*	Rhizopus microsporus endornavirus 1	ssRNA(+)	LC671616
15	*R. microsporus*	*Endornaviridae*	*Alphaendornavirus*	Rhizopus microsporus endornavirus 2	ssRNA(+)	LC671617
15	*R. microsporus*	*Mitoviridae*	Unclassified	Rhizopus microsporus mitovirus 1	ssRNA(+)	LC671615
56	*Cladosporium sphaerospermum*	*Botourmiaviridae*	*Penoulivirus*	Erysiphe necator associated ourmia-like virus 69	ssRNA(+)	LC671618
60	*P. vanoranjei*	Unclassified	Unclassified	Penicillium vanoranjei associated RNA virus 1	RNA	LC671619
61	*A. lentulus*	*Botourmiaviridae*	*Magoulivirus*	Aspergillus fumigatus botourmiavirus 1	ssRNA(+)	LC671624
61	*A. lentulus*	*Narnaviridae*	Unclassified	Aspergillus fumigatus narnavirus 2	ssRNA(+)	LC671623
61	*A. lentulus*	Unclassified	Unclassified	Aspergillus lentulus jivivirus 1	RNA	LC671620 LC671621 LC671622

DNA virus discovery in RCA libraries.

Analysis of VIDISCA sequencing reads identified no DNA viruses in the 61 analyzed samples, and we therefore focused instead on the Illumina libraries enriched by rolling circle amplification (RCA) for circular ssDNA. These had an average of 1.7 million reads per library after quality control. Removal of poor quality contigs left 14 that appeared to be of cressdnaviral origin, since they possessed at least partial Rep and Cap proteins with BLASTp identity to known viruses. We found that 12 of the 14 sequences were complete

circular genomes and the last two were truncated. Each was derived from a different sample. Since RCA indiscriminately amplifies any primed circular ssDNA, we explored the possibility that they represented viral contaminants amplified from reagents rather than mycoviruses (26, 27). We first looked at the distribution of reads mapping to each sequence across all RCA libraries. We found that one of the incomplete genomes had a clear signature of contamination, being positive in 11 of 61 samples at a cutoff of 50 reads per million (RPM, 18% prevalence), with 39 samples containing at least one read (Table S4). On closer examination, this sequence was also found to contain a region with BLASTn identity to the fungal isolate species, suggesting a hybrid assembly. Despite this, after trimming off the hybrid region, the contaminant mapping signature remained. We opted to retain the Rep sequence for phylogenetic analysis, but did not upload the nucleotide sequence to the International Nucleotide Sequence Database Collaboration (INSDC) databases due to its uncertain quality (instead providing it at https://figshare.com/projects/Viruses_infecting_clinical_mycology_cultures/128186). The other 13 sequences (Table 2) did not contain hybrid regions, and also showed more specific distributions, being positive in between one and three samples (1.6% to 4.9% prevalence). This aligned more with our expectations of mycoviruses rather than contaminants. Despite this, PCR screening failed to detect the 13 sequences in unamplified index samples, while off-target amplification showed the polymerase was active. We hypothesized that this could be explained by low viral load combined with poor PCR efficiency, and so repeated the RCA step with freshly extracted nucleic acids. PCR on the amplified nucleic acids again failed to detect the viruses, suggesting they represent low-level contamination of a reagent or reagents used upstream of RCA. This result shows the importance of validation assays, and underscores that contaminants will not necessarily be widely distributed across samples, presumably due to a low initial load and related sampling effects. None of the 13 sequences dominated their respective libraries, with 3,869 RPM the maximum normalized read count, in line with low load prior to RCA. Previous work in our laboratory on the circular anelloviruses has shown RCA can amplify genomes to a level allowing complete assembly, even from loads below PCR and qPCR limits of detection (28). Together, the result strongly implies caution is needed in interpreting the biological source of amplified viruses, as incorrect host predictions could easily occur.

TABLE 2 Contaminant circular single-stranded DNA viruses in RCA libraries constructed from clinical isolates of fungi

Sample	Virus order	Virus name	Molecule	Accession
1	*Arfiviricetes*	*Cressdnaviricota* sp. isolate 2020-AMS-AF	ssDNA	LC671629
3	*Arfiviricetes*	*Cressdnaviricota* sp. isolate 2020-AMS-AN	ssDNA	LC671630
12	*Arfiviricetes*	*Cressdnaviricota* sp. isolate 2020-AMS-CB	ssDNA	LC671631
17	*Arfiviricetes*	*Cressdnaviricota* sp. isolate 2020-AMS-MCA	ssDNA	LC671632
19	*Arfiviricetes*	*Cressdnaviricota* sp. isolate 2020-AMS-MP	ssDNA	LC671633
29	*Arfiviricetes*	*Cressdnaviricota* sp. isolate 2020-AMS-MCO	ssDNA	LC671634
31	*Arfiviricetes*	*Cressdnaviricota* sp. isolate 2020-AMS-TR	ssDNA	LC671625
33	*Arfiviricetes*	*Cressdnaviricota* sp. isolate 2020-AMS-TS	ssDNA	LC671626
40	*Arfiviricetes*	*Cressdnaviricota* sp. isolate 2020-AMS-SB	ssDNA	LC671627
42	*Arfiviricetes*	*Cressdnaviricota* sp. isolate 2020-AMS-ED	ssDNA	LC671637
46	*Arfiviricetes*	*Cressdnaviricota* sp. isolate 2020-AMS-AK	ssDNA	LC671636
49	*Arfiviricetes*	*Cressdnaviricota* sp. isolate 2020-AMS-RA	ssDNA	LC671628
54	*Arfiviricetes*	*Cressdnaviricota* sp. isolate 2020-AMS-SP	ssDNA	LC671635

Virus relationships and taxonomy.

Two members of the family *Endornaviridae* were found in a single *Rhizopus microsporus* culture. They were distantly related to each other, sharing 13% amino acid (aa) identity across the RdRp, and interestingly both were distant to all other public endornavirus sequences. The Committee on Taxonomy of Viruses (ICTV) species demarcation for endornaviruses is <75% nucleotide (nt) identity across the genome (29), and both sequences would qualify as new species by this criterion. As the first endornaviruses identified in *R. microsporus*, we tentatively named them Rhizopus microsporus endornavirus 1 and 2 (RMEV1 and RMEV2). The closest relative of RMEV1 was Rhizoctonia solani endornavirus 7 (QDW65434.1) at 15.6% RdRp aa identity, while for RMEV2 it was Phytophthora endornavirus 2 (BCL84886.1) at 17.14% RdRp aa identity. Phylogenetic analysis showed both RMEV1 and RMEV2 cluster within the genus *Alphaendornavirus* (Fig. 2A), consistent with their relatively large genomes of 13,589 bp

and 11,599 bp, respectively. Alphaendornaviruses are currently known to infect fungi, plants, and oomycetes, and hypovirulent effects on hosts have been observed in some cases, for example the alphaendornavirus Helicobasidium mompa endornavirus 1 (30).

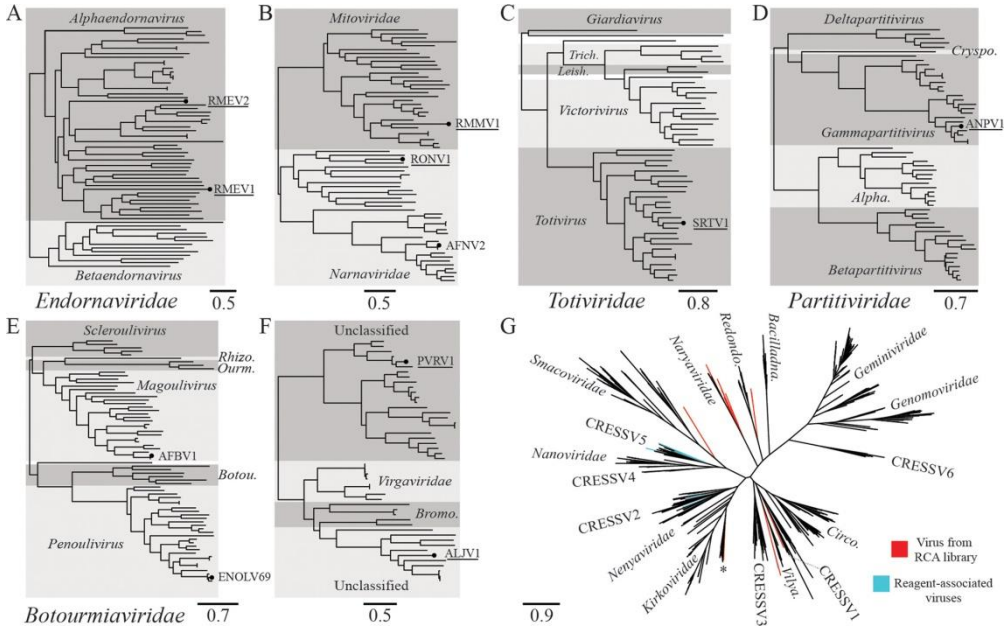


FIG 2 Phylogenetic relationships of viruses. Scale bars refer to amino acid substitutions per site. An underlined virus name denotes a viral genome likely meeting criteria for a new species. (A) *Endornaviridae*, Rhizopus microsporus endornavirus 1 (RMEV1, LC671616), Rhizopus microsporus endornavirus 2 (RMEV2, LC671617). (B) *Mitoviridae* and *Narnaviridae*, Rhizopus microsporus mitovirus 1 (RMMV1, LC671615), Rhizopus oryzae narnavirus 1 (RONV1, LC671613), Aspergillus fumigatus narnavirus 2 (AFNV2, LC671623, isolated here from *Aspergillus lentulus*). (C) *Totiviridae*, Syncephalastrum racemosum totivirus 1 (SRTV1, LC671614); *Trich.*, *Trichomonasvirus*; *Leish.*, *Leishmaniavirus*. (D) *Partitiviridae*, Aspergillus niger partitivirus 1 (ANPV1, LC671611); *Cryspo.*, *Cryspovirus*; *Alpha.*, *Alphapartitivirus*. (E) *Botourmiaviridae*, Aspergillus fumigatus botourmiavirus 1 (AFBV1, LC671624, isolated here from *Aspergillus lentulus*), Erysiphe necator-associated ourmia-like virus 69 (ENOLV69, LC671618, isolated here from *Cladosporium sphaerospermum*); *Rhizo.*, *Rhizoulivirus*; *Ourm.*, *Ourmiavirus*; *Botou.*, *Botoulivirus*. (F) Unclassified RNA viruses: Penicillium vanoranjei-associated RNA virus 1 (PVRV1, LC671619); Aspergillus lentulus jivivirus 1 (ALJV1, LC671620); *Bromo.*, *Bromoviridae*. (G) *Cressdnaviricota*: 14 Rep sequences from viruses found in RCA libraries are highlighted in red, while known reagent-associated viruses are highlighted in blue. Asterisked sequence was a hybrid assembly and therefore not uploaded to INSDC databases. *Bacilladna.*, *Bacilladnaviridae*; *Redondo.*, *Redondoviridae*; *Vilya.*, *Vilyaviridae*; *Circo.*, *Circoviridae*. Alignments, tree files, and the hybrid sequence are available at https://figshare.com/projects/Viruses_infecting_clinical_mycology_cultures/128186.

One member of the family *Mitoviridae* and two members of the family *Narnaviridae* were identified. The mitovirus coinfected the same *R. microsporus* culture as RMEV1 and RMEV2, though while endornavirus replication is cytoplasmic, mitoviruses replicate in fungal mitochondria (31). Mitovirus genus and species demarcation criteria have yet to be defined, but <40% RdRp aa identity has historically been found between defined species (32). On this basis, we suggest the genome be named Rhizopus microsporus mitovirus 1 (RMMV1), with Entomophthora muscae mitovirus 2 (QCF24461.1) as the closest relative (37.4% RdRp aa identity). Narnaviruses were found in cultures of *Rhizopus oryzae* and *Aspergillus lentulus*. They shared only 8% RdRp aa with each other and clustered in different parts of the family tree (Fig. 2B). Genus demarcation for narnaviruses has not been defined, but the species cutoff is <50% RdRp aa identity (32). The former genome met this criterion, and we suggest the name Rhizopus oryzae narnavirus 1 (RONV1) for it, which has 36.6% RdRp aa identity to its closest relative, Erysiphe necator-associated narnavirus 42 (QJT93774.1). The virus found in *A. lentulus* belongs to the previously described species Aspergillus fumigatus narnavirus 2 (AFNV2), sharing 98% RdRp aa identity with accession AXE72934.1. Although no data on biological impact are currently available for these viruses, some mitoviruses and narnaviruses do have the potential to impact fungal biology either by conferring hypovirulence or affecting reproductive capabilities (33, 34).

A member of the family *Totiviridae* was found in a *Syncephalastrum racemosum* culture. Phylogenetic analysis placed it within the genus *Totivirus* (Fig. 2C). Species demarcation criteria for the *Totivirus* genus are not absolute, and largely relate to biological characteristics such as host range (though <50% RdRp aa identity is also considered a probable species cutoff). The closest relative of the virus identified here was Trichoderma koningiopsis totivirus 1 (QGA70771.1) at 60.9% RdRp aa identity. Despite this, we suggest the genome be given the provisional name Syncephalastrum racemosum totivirus 1 (SRTV1). Our rationale is the large phylogenetic distance between the host genera *Syncephalastrum* and *Trichoderma* and the current lack of biological data to support assignment to the same species. Members of the genus *Totivirus* have been previously associated with hypovirulence (35).

A virus belonging to the family *Partitiviridae* was identified in a culture of *Aspergillus niger*. The sequence was phylogenetically placed within the genus *Gammapartitivirus* (Fig. 2D). Criteria for species demarcation within this genus are <90% RdRp aa identity and also <80% capsid aa identity (36), and the sequence identified here meets this, most closely related to Botryosphaeria dothidea virus 1 (KJ722537.1) with 77.1% RdRp aa identity and 54.9% capsid aa identity. We suggest the name Aspergillus niger partitivirus 1 (ANPV1). The finding is in line with the known ascomycete host range of the genus *Gammapartitivirus.*

Two members of the family *Botourmiaviridae* were found, one in the *A. lentulus* culture also containing AFNV2 and another in a *Cladosporium sphaerospermum* culture. The

former belonged to the genus *Magoulivirus*, while the latter belonged to the genus *Penoulivirus* (Fig. 2E). The species demarcation criterion for both these genera is <90% RdRp aa identity, and neither met this; the magoulivirus belongs to Aspergillus fumigatus botourmiavirus 1 (AFBV1, BCH36640.1) with 97.7% RdRp aa identity, while the penoulivirus belongs to Erysiphe necator-associated ourmia-like virus 69 (QKI79899.1) with 90.7% RdRp aa identity. At least one member of the family has previously been shown to be associated with the hypovirulence of its host (37).

The final two RNA viruses identified were both unclassified. One coinfected the *A. lentulus* culture alongside AFNV2 and AFBV1, while the other was found in a *Penicillium vanoranjei* culture. BLASTp searches showed the closest relatives of the *A. lentulus* virus included Citrus virga-like virus (CVLV, ARO38274.1) and Grapevine-associated jivivirus 1 (QIJ25698.1), each with approximately 40% RdRp aa identity across >96% query coverage. Though no ICTV guidelines on species demarcation currently exist for this lineage, we propose this genome be named Aspergillus lentulus jivivirus 1 (ALJV1) on the basis of low sequence identity to relatives (39% RdRp aa identity across the whole protein alignment to CVLV) in combination with a novel host record; notably, this is the first jivivirus identification in an axenic fungal culture. As the first record, no data are currently available regarding the biological impact of jivivirus infection on their fungal hosts. The closest relative of the virus infecting *P. vanoranjei* was an unclassified virus recorded as *Riboviria* sp. (QDH88072.1), at 49% RdRp aa identity. We gave it the temporary name Penicillium vanoranjei-associated RNA virus 1 (PVRV1) until proper taxonomic classification. Notably, BLASTp results suggested ALJV1 was related to the *Virgaviridae* and *Bromoviridae*, while PVRV1 hit one sequence labeled as virga-like (BBB86779.1). We therefore analyzed their relationships together, alongside representatives of both families. This confirmed a close relationship between ALJV1 and members of the *Virgaviridae* and *Bromoviridae* (Fig. 2F). PVRV1 was resolved in a distinct lineage alongside other unclassified viruses, many of which were themselves found associated with fungi.

As described above, analysis of RCA libraries returned 14 cressdnavirus sequences suspected of being contaminants due to PCR validation failure in the original samples. Phylogenetic analysis of the Rep proteins showed they clustered in distinct locations across the *Arfiviricetes* class (Fig. 2G). The only family of cressdnaviruses currently recognized to infect fungi are the family *Genomoviridae*, belonging to the class *Repensiviricetes*. This is concordant with a nonfungal host of these viruses. Largely, the 14 Rep sequences could not be assigned to recognized clusters or families (except one apparent member of *Naryaviridae*), but all remaining sequences were resolved as distant relatives of lineages, including *Kirkoviridae*, *Smacoviridae*, *Naryaviridae*, *Redondoviridae*, CRESSV1, and *Vilyaviridae*. These lineages in particular are conspicuous since all are found associated with the gastrointestinal tracts of humans and other animals, and also in the human respiratory environment in the case of *Redondoviridae* (38). Gastrointestinal viruses are often detected in stool-contaminated wastewater. While unconfirmed, if the viruses detected here occupy similar niches, it may suggest the true contamination source is

recycled water. The 14 viral Reps were mostly unrelated to sequences previously identified as contaminants (26), though one (LC671626) did cluster alongside MZ824233.1 and MZ824234.1.

Discussion

Decades of research have uncovered numerous mycoviruses, with the bulk of sampling effort directed toward industrially relevant hosts, such as plant-pathogenic fungi or edible mushrooms (39, 40). Large-scale efforts to genetically catalogue mycoviruses of human-pathogenic fungi specifically have been limited (40), though there has long been evidence they also host viruses (39), and genomes are now increasingly becoming available on public databases. Here, we investigated the viruses of 75 clinical isolates of medically relevant fungi, covering 27 genera and at least 56 species. We uncovered a remarkable diversity of 11 RNA mycovirus genomes in seven hosts. This probably represents an underestimate of the true RNA virus richness in our sample set, since even within the seven deep-sequenced samples, we detected five viruses not observed with initial VIDISCA screening. While we currently lack data on the biological impact of these viruses, many are related to viruses capable of conferring hypovirulence on their hosts. Despite this, the majority of mycoviruses do not negatively impact their hosts (41), and each must be individually characterized, for example by comparing growth characteristics of infected cultures with virus-free ones. A notable possibility is that clinically isolated fungi may be particularly poor sources for discovery of hypovirulence-associated viruses, since they are competent pathogens upon isolation. Screening of fungi in their alternative niches might therefore be more productive in this regard. Aside from hypovirulence, the therapeutic potential of RNA mycoviruses is generally unfavourable, since all studied to date lack an extracellular stage, possibly due to a physical inability to transit pores in fungal cell walls (25). Discovery of smaller ssDNA viruses may circumvent this barrier.

Detected mycoviruses belonged to six families and two additional unclassified groups. Only in two cases were their closest known relatives also identified in a human-pathogenic fungus (AFBV1 and AFNV2, from *Aspergillus fumigatus*). In five cases, the closest mycovirus relatives were identified in phytopathogenic fungi, one was found in an endophytic species, one in an entomopathogenic fungus, and in two cases the relatives were from uncertain hosts. This is likely partly due to low sampling effort toward human pathogens as mentioned above; however, it probably also reflects the fact that human-pathogenic fungi are phylogenetically nested within nonpathogenic lineages across the fungal radiation (42) and consequently share their viral lineages. Indeed, human-pathogenic fungi are mostly opportunistic rather than obligate pathogens (42), normally filling other ecological roles where mycovirus host switches could occur. For example, *Rhizopus microsporus* is both human- and plant-pathogenic, and its virome might therefore be expected to resemble other phytopathogens.

This study focused on fungi recultured from axenic stocks, with all RNA viruses confirmed by PCR in their original sample extractions. The viral phylogenetic relationships were also concordant with previous mycovirus literature or public sequences (except ALJV1, see below), and we were thus confident they represented true mycoviruses and not contamination. The notable exception was ALJV1, which represents the first unambiguous detection of a jivivirus in a fungus. Previous identifications of the recently named jiviviruses (43) have been plant or plant-pest associated (thrips), though they have also been observed associated with *Plasmopara*-infected grapevines (43). Besides fungi, it is therefore probable they infect plants, and potentially oomycetes. The unclassified lineage containing jiviviruses is related to the families *Virgaviridae* and *Bromoviridae*, both of which infect plants (44, 45). Interestingly, virga-like viruses are also known from a fungus (46), and cucumber mosaic virus (CMV, family *Bromoviridae*) has been observed to naturally infect the phytopathogenic fungus *Rhizoctonia solani*, which can in turn transmit CMV to uninfected plants under laboratory conditions (47). Such cross-kingdom transmission may similarly occur with jiviviruses. Interestingly, most lineages of mycoviruses have plant virus relatives (48), hinting at a deep history of cross-kingdom host shifts during extensive ecological interaction. This is true for most families identified here; indeed, some members of the *Endornaviridae*, *Botourmiaviridae*, *Mitoviridae*, *Partitiviridae*, and *Totiviridae* can all infect plants (48). Cross-kingdom transmission has been suggested to have played a major role in the evolution of mycoviruses (49).

We also detected ssDNA viruses in RCA libraries, 12 with complete genomes. We universally failed to validate these by PCR, both in the original unamplified samples and after repeating RCA. Recently, more attention has been given to the detection of viral contaminants in sequencing libraries, since they can easily result in incorrect assessments of virus–host relationships (26, 50). Here, we found that contaminating sequences can occur in RCA libraries without a wide distribution as may be expected, but rather occurring in between one and three samples each. This serves as a further caution that validation assays are essential to confirm the presence of viruses in samples. Our failure to detect DNA viruses is perhaps unsurprising, given their relative rarity among mycoviruses (13). Despite finding no ssDNA mycoviruses here, we reiterate the rationale that discovering hypovirulence-associated ssDNA viruses with an extracellular stage may represent the best opportunities for application in human therapeutics, and we therefore suggest RCA should still be applied in similar surveys of human pathogens.

Materials and methods

Fungal clinical isolates

Bronchial aspirates, bronchoalveolar lavage fluid, bone marrow, and biopsy specimens sent in for culture of (dimorphic) fungi were inoculated on two containers of brain heart infusion

agar with penicillin and gentamicin. One container was incubated for 3 weeks at 20 to 25°C, and the other at 35 to 37°C. Nails, hair, and skin scrapings were inoculated on dermatophyte test medium agar and Sabouraud agar with gentamicin and chloramphenicol (SabGC). Incubation was for 3 weeks at 25 to 28°C, with one container of SabGC incubated for 3 weeks at 35 to 37°C. All other materials sent in for fungal culture were inoculated on two SabGC containers for 1 week, one at 25 to 28°C and one at 35 to 37°C. Fungal isolates included in this study were recultured from glycerol stocks, and samples were transferred to tubes containing Universal Transport Medium (UTM, Copan). A total of 75 isolates were included, split into two batches (Table S1). Batch one samples (61 diverse isolates) were utilized in virus discovery and PCR screening experiments, while batch two samples (14 isolates of *Mucorales* species) were included later and used only in PCR screening.

Next generation sequencing

Fungal swabs were suspended 1:3 in UTM. Sample suspension (110 µL) was transferred to a reaction tube and centrifuged (10 min at 5,000 *g*) to pellet solid matter and cellular debris. Supernatant was treated with 20 µL TURBO DNase (Thermo Fisher Scientific, Waltham, MA, USA) for 30 min at 37°C to remove naked DNA. Nucleic acids were extracted using the Boom method (51) and were then split according to their use either in VIDISCA or RCA library preparation. For VIDISCA, reverse transcription (RT) was done on 20 µL using nonribosomal hexamer primers (52). This was followed by second-strand synthesis and a cleanup via phenol-chloroform extraction and ethanol precipitation. Double-stranded DNA was digested with MseI restriction enzyme, and sequencing adapters were ligated to the sticky ends. Libraries were amplified before size selection of fragments between 200 and 600 bp, quantification, and pooling. Sequencing was then done on an IonTorrent S5 instrument. For RCA, 4 µL of extracted nucleic acids was incubated with Φ29 DNA polymerase and exonuclease-resistant random primers for 4 h at 30°C. Product was incubated with NEBNext dsDNA fragmentase (New England Biolabs) for 25 min at 37°C and then cleaned up, which was also done after each subsequent step. Fragmented DNA was end repaired for 30 min at 37°C using the Klenow fragment of DNA polymerase I (New England Biolabs) before A-tailing was carried out for 30 min at 37°C with Klenow fragment (3′→5′ exo-, New England Biolabs). NEBNext adapters (1:1,000 dilution, New England Biolabs) were ligated overnight at 16°C using T4 DNA ligase (5 U/µL, Invitrogen). After size selection of fragments >200 bp, adaptor-ligated DNA was treated with USER enzyme (New England Biolabs) and was then enriched and indexed during a 12-cycle PCR. Further size selection to target fragments between 200 and 600 bp was done, before quantification, pooling, and paired-end sequencing (2 × 150 bp) on an Illumina MiSeq instrument. Total nucleic acid metagenomic sequencing was carried out on samples positive for RNA viruses after VIDISCA sequencing. Library preparation up to second-strand synthesis was identical to the VIDISCA protocol described above, except that RT hexamers carried a 5′-phosphate and AMPure XP beads were used for all cleanups. After second-strand synthesis, the protocol matched RCA methodology from fragmentation

onward. The library preparation and sequencing was carried out twice independently. Raw Illumina reads are available under European Nucleotide Archive project accession PRJEB49942.

Virus discovery and genome assembly

VIDISCA sequences were analyzed with a previously published workflow (53). Briefly, reads were aligned to viral proteins using the UBLAST algorithm (54), before reduction of false positives by alignment of hits to the GenBank nt database using BLASTn (55). Visual outputs generated from hit tables were inspected to identify viral content, and samples positive for RNA viruses were selected for RNA deep sequencing. Illumina sequence reads from virus-enriched metagenomic libraries and RCA products were cleaned of adapters and quality trimmed to a Phred score of 30 using BBDuk, from BBMap v38.71. *De novo* assembly was done with SPAdes v3.15.2 (56). Contigs from the metagenomic libraries were aligned to a database of viral proteins using UBLAST to identify putative RNA viruses. Contigs generated from RCA products were aligned to a database of Rep genes covering *Cressdnaviricota* diversity. Matching contigs above 1,500 bp were aligned to the GenBank nt database using BLASTn to remove nonviral sequences. Remaining contigs were self-aligned using the MAFFT online server (https://mafft.cbrc.jp/alignment/server/), and those containing visible misassemblies were discarded. Genome completeness was assessed by the presence of both *Rep* and *Cap* genes plus genome circularity (identical sequence at both contig ends). Circular overlap was trimmed from complete genomes, and they were rotated to begin with the *Rep* gene. For all RNA and DNA genomes, inspection was performed by mapping quality controlled reads to respective sequences using BWA MEM v0.7.17 (57) and manually examining resulting pileups. Contigs were curated to correct minor errors, though assemblies with uncorrectable misassemblies were discarded.

PCR validation and analysis of viral distribution

To confirm viral RNA in index samples, and assess distribution across the others, PCRs were designed and run on freshly extracted and reverse transcribed nucleic acids from all samples (for primers, see Table S2). A 40-cycle first round was performed, with nested PCR carried out if this was negative, and PCR products were Sanger sequenced. For ssDNA viruses, only index samples were screened due to nondetection. Both the original unamplified samples and an aliquot with RCA repeated were screened. To explore the possibility that they represented contaminants of reagents, we also examined ssDNA virus read distribution across all RCA libraries. Since reagents were shared between samples, we expected that contaminants would be detected in all or most sequencing libraries. This was tested by mapping reads from all RCA libraries to the set of cressdnaviral sequences using BWA and examining their read distributions per sample, using 50 reads per million (RPM) as the cutoff for positive detection.

Phylogenetic and genetic distance analyses

For each RNA virus group, representative RNA-dependent RNA polymerase (RdRp) or polyprotein sequences were gathered using the International Committee on Taxonomy of Viruses (ICTV) website and GenBank, and alignment was performed using MAFFT v7.490 (58) with the E-INS-i setting. Open reading frame prediction for genomic segments was done using default ORFfinder settings (https://www.ncbi.nlm.nih.gov/orffinder), though the yeast mitochondrial genetic code was applied for Rhizopus microsporus mitovirus 1 (LC671615), since its UGA codon encodes tryptophan rather than a translation termination signal (31). Maximum likelihood phylogenetic analyses were done using IQ-TREE v1.6.11 (59) with automatic model detection, 1,000 ultrafast bootstrap tests, and 1,000 SH-aLRT tests. Pairwise distances between proteins were calculated for each alignment, using the SIAS web tool with mean length of sequences set as the denominator (http://imed.med.ucm.es/Tools/sias.html). For cressdnaviruses, Rep protein sequences were aligned alongside a database covering recognized and proposed lineages collated for a previously published phylogeny (60), now with the addition of the proposed family *Kirkoviridae* (61) and seven Rep sequences previously shown to be reagent associated (26). Phylogenetic analysis was as above. All alignments and tree files are available at https://figshare.com/projects/Viruses_infecting_clinical_mycology_cultures/128186.

Data availability

Assembled genomes are available from INSDC databases under accessions LC671611–LC671637. Raw Illumina reads are available under European Nucleotide Archive project accession PRJEB49942. Alignments, tree files, and one hybrid sequence are available from https://figshare.com/projects/Viruses_infecting_clinical_mycology_cultures/128186.

For supplementary tables, see the online version: https://doi.org/10.1128/spectrum.01610-22.

Acknowledgements

We thank Joanna Kaczorowska for discussions regarding rolling circle amplification of ssDNA viruses.

This work was supported by a grant from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie agreement No. 721367 (HONOURs).

References

1. Lehrnbecher, T.; Frank, C.; Engels, K.; Kriener, S.; Groll, A.H.; Schwabe, D. Trends in the postmortem epidemiology of invasive fungal infections at a university hospital. J. Infect. 2010, 61, 259–265.

2. McNeil, M.M.; Nash, S.L.; Hajjeh, R.A.; Phelan, M.A.; Conn, L.A.; Plikaytis, B.D.; Warnock, D.W. Trends in mortality due to invasive mycotic diseases in the United States, 1980-1997. Clin. Infect. Dis. 2001, 33, 641–647.

3. Enoch, D.A.; Ludlam, H.A.; Brown, N.M. Invasive fungal infections: A review of epidemiology and management options. J. Med. Microbiol. 2006, 55, 809–818.

4. Krishnan-Natesan, S. Emerging fungal infections in cancer patients - a brief overview. Med. Mycol. Open Access 2016, 2, 12.

5. Marr, K.A. Fungal infections in hematopoietic stem cell transplant recipients. Med. Mycol. 2008, 46, 293–302.

6. Babady, N.E.; Buckwalter, S.P.; Hall, L.; Le Febre, K.M.; Binnicker, M.J.; Wengenack, N.L. Detection of Blastomyces dermatitidis and Histoplasma capsulatum from culture isolates and clinical specimens by use of real-time PCR. J. Clin. Microbiol. 2011, 49, 3204–3208.

7. Park, B.J.; Wannemuehler, K.A.; Marston, B.J.; Govender, N.; Pappas, P.G.; Chiller, T.M. Estimation of the current global burden of cryptococcal meningitis among persons living with HIV/AIDS. AIDS 2009, 23, 525–530.

8. van Arkel, A.L.E.; Rijpstra, T.A.; Belderbos, H.N.A.; van Wijngaarden, P.; Verweij, P.E.; Bentvelsen, R.G. COVID-19-associated pulmonary aspergillosis. Am. J. Respir. Crit. Care Med. 2020, 202, 132–135.

9. Koehler, P.; Cornely, O.A.; Böttiger, B.W.; Dusse, F.; Eichenauer, D.A.; Fuchs, F.; Hallek, M.; Jung, N.; Klein, F.; Persigehl, T.; et al. COVID-19 associated pulmonary aspergillosis. Mycoses 2020, 63, 528.

10. Mastrangelo, A.; Germinario, B.N.; Ferrante, M.; Frangi, C.; Li Voti, R.; Muccini, C.; Ripa, M.; Group, C.-B.S.; Canetti, D.; Castiglioni, B.; et al. Candidemia in coronavirus disease 2019 (COVID-19) patients: Incidence and characteristics in a prospective cohort compared with historical non–COVID-19 controls. Clin. Infect. Dis. 2021, 73, e2838–e2839.

11. Nucci, M.; Barreiros, G.; Guimarães, L.F.; Deriquehem, V.A.S.; Castiñeiras, A.C.; Nouér, S.A. Increased incidence of candidemia in a tertiary care hospital with the COVID-19 pandemic. Mycoses 2021, 64, 152–156.

12. Pal, R.; Singh, B.; Bhadada, S.K.; Banerjee, M.; Bhogal, R.S.; Hage, N.; Kumar, A. COVID-19-associated mucormycosis: An updated systematic review of literature. Mycoses 2021, 64, 1452–1459.

13. Ghabrial, S.A.; Castón, J.R.; Jiang, D.; Nibert, M.L.; Suzuki, N. 50-plus years of fungal viruses. Virology 2015, 479–480, 356–368.

14. Yu, X.; Li, B.; Fu, Y.; Jiang, D.; Ghabrial, S.A.; Li, G.; Peng, Y.; Xie, J.; Cheng, J.; Huang, J.; et al. A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. Proc. Natl. Acad. Sci. 2010, 107, 8387–8392.

15. Hao, F.; Wu, M.; Li, G. Characterization of a novel genomovirus in the phytopathogenic fungus Botrytis cinerea. Virology 2021, 553, 111–116.

16. Li, P.; Wang, S.; Zhang, L.; Qiu, D.; Zhou, X.; Guo, L. A tripartite ssDNA mycovirus from a plant pathogenic fungus is infectious as cloned DNA and purified virions. Sci. Adv. 2020, 6.

17. Zhao, L.; Lavington, E.; Duffy, S. Truly ubiquitous CRESS DNA viruses scattered across the eukaryotic tree of life. J. Evol. Biol. 2021, 34, 1901–1916.

18. Anagnostakis, S.L. Biological control of chestnut blight. Science 1982, 215, 466–471.

19. Macdonald, W.L.; Fulbright, D.W. Biological control of chestnut blight: Use and limitations of transmissible hypovirulence. Plant Dis. 1991, 75.

20. Qu, Z.; Zhao, H.; Zhang, H.; Wang, Q.; Yao, Y.; Cheng, J.; Lin, Y.; Xie, J.; Fu, Y.; Jiang, D. Bio-priming with a hypovirulent phytopathogenic fungus enhances the connection and strength of microbial interaction network in rapeseed. npj Biofilms Microbiomes 2020, 6, 1–13.

21. Chiba, S.; Salaipeth, L.; Lin, Y.-H.; Sasaki, A.; Kanematsu, S.; Suzuki, N. A novel bipartite double-stranded RNA mycovirus from the white root rot fungus Rosellinia necatrix: Molecular and biological characterization, taxonomic considerations, and potential for biological control. J. Virol. 2009, 83, 12801.

22. Eren, R.O.; Reverte, M.; Rossi, M.; Hartley, M.A.; Castiglioni, P.; Prevel, F.; Martin, R.; Desponds, C.; Lye, L.F.; Drexler, S.K.; et al. Mammalian innate immune response to a Leishmania-resident RNA virus increases macrophage survival to promote parasite persistence. Cell Host Microbe 2016, 20, 318–328.

23. Nuss, D.L. Hypovirulence: Mycoviruses at the fungal–plant interface. Nat. Rev. Microbiol. 2005, 3, 632–642.

24. Yu, X.; Li, B.; Fu, Y.; Xie, J.; Cheng, J.; Ghabrial, S.A.; Li, G.; Yi, X.; Jiang, D. Extracellular transmission of a DNA mycovirus and its use as a natural fungicide. Proc. Natl. Acad. Sci. 2013, 110, 1452–1457.

25. Van De Sande, W.W.J.; Vonk, A.G. Mycovirus therapy for invasive pulmonary aspergillosis? Med. Mycol. 2019, 57, S179–S188.

26. Boom, R.; Sol, C.J.; Salimans, M.M.; Jansen, C.L.; Wertheim-Van Dillen, P.M.; van der Noordaa, J. Rapid and simple method for purification of nucleic acids. J. Clin. Microbiol. 1990, 28, 495–503.

27. Endoh, D.; Mizutani, T.; Kirisawa, R.; Maki, Y.; Saito, H.; Kon, Y.; Morikawa, S.; Hayashi, M. Species-independent detection of RNA virus by representational difference analysis using non-ribosomal hexanucleotides for reverse transcription. Nucleic Acids Res. 2005, 33, e65.

28. Kinsella, C.M.; Deijs, M.; van der Hoek, L. Enhanced bioinformatic profiling of VIDISCA libraries for virus detection and discovery. Virus Res. 2019, 263, 21–26.

29. Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010, 26, 2460–2461.

30. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. BMC Bioinformatics 2009, 10, 421.

31. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D.; et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 2012, 19, 455–477.

32. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN] 2013.

33. Katoh, K.; Rozewicki, J.; Yamada, K.D. MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. 2017, 20, 1160–1166.

34. Nibert, M.L. Mitovirus UGA(Trp) codon usage parallels that of host mitochondria. Virology 2017, 507, 96–100.

35. Nguyen, L.-T.; Schmidt, H.A.; von Haeseler, A.; Minh, B.Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 2015, 32, 268–274.

36. Kinsella, C.M.; Bart, A.; Deijs, M.; Broekhuizen, P.; Kaczorowska, J.; Jebbink, M.F.; van Gool, T.; Cotten, M.; van der Hoek, L. Entamoeba and Giardia parasites implicated as hosts of CRESS viruses. Nat. Commun. 2020, 11, 1–10.

37. Li, L.; Giannitti, F.; Low, J.; Keyes, C.; Ullmann, L.S.; Deng, X.; Aleman, M.; Pesavento, P.A.; Pusterla, N.; Delwart, E. Exploring the virome of diseased horses. J. Gen. Virol. 2015, 96, 2721–2733.

38. Porter, A.F.; Cobbin, J.; Li, C.-X.; Eden, J.-S.; Holmes, E.C. Metagenomic identification of viral sequences in laboratory reagents. Viruses 2021, 13, 2122.

39. Naccache, S.N.; Greninger, A.L.; Lee, D.; Coffey, L.L.; Phan, T.; Rein-Weston, A.; Aronsohn, A.; Hackett, J.; Delwart, E.L.; Chiu, C.Y. The perils of pathogen discovery: Origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. J. Virol. 2013, 87, 11966–11977.

40. Kaczorowska, J.; Deijs, M.; Klein, M.; Bakker, M.; Jebbink, M.F.; Sparreboom, M.; Kinsella, C.M.; Timmerman, A.L.; van der Hoek, L. Diversity and long-term dynamics of human blood anelloviruses. J. Virol. 2022.

41. Khalifa, M.E.; Pearson, M.N. Molecular characterisation of an endornavirus infecting the phytopathogen Sclerotinia sclerotiorum. Virus Res. 2014, 189, 303–309.

42. Osaki, H.; Nakamura, H.; Sasaki, A.; Matsumoto, N.; Yoshida, K. An endornavirus from a hypovirulent strain of the violet root rot fungus, Helicobasidium mompa. Virus Res. 2006, 118, 143–149.

43. Hillman, B.; Esteban, R. Family Narnaviridae. In Virus taxonomy: ninth report of the international committee on taxonomy of viruses; King, A., Adams, M., Carstens, E., Lefkowitz, E., Eds.; Elsevier: Amsterdam, 2011; p. 1058.

44. Xu, Z.; Wu, S.; Liu, L.; Cheng, J.; Fu, Y.; Jiang, D.; Xie, J. A mitovirus related to plant mitochondrial gene confers hypovirulence on the phytopathogenic fungus Sclerotinia sclerotiorum. Virus Res. 2015, 197, 127–136.

45. Espino-Vázquez, A.N.; Bermúdez-Barrientos, J.R.; Cabrera-Rangel, J.F.; Córdova-López, G.; Cardoso-Martínez, F.; Martínez-Vázquez, A.; Camarena-Pozos, D.A.; Mondo, S.J.; Pawlowska, T.E.; Abreu-Goodger, C.; et al. Narnaviruses: Novel players in fungal-bacterial symbioses. ISME J. 2020, 14, 1743–1754.

46. Suzaki, K.; Ikeda, K.-I.; Sasaki, A.; Kanematsu, S.; Matsumoto, N.; Yoshida, K. Horizontal transmission and host-virulence attenuation of totivirus in violet root rot fungus Helicobasidium mompa. J. Gen. Plant Pathol. 2005, 71, 161–168.

47. Vainio, E.J.; Chiba, S.; Ghabrial, S.A.; Maiss, E.; Roossinck, M.; Sabanadzovic, S.; Suzuki, N.; Xie, J.; Nibert, M. ICTV virus taxonomy profile: Partitiviridae. J. Gen. Virol. 2018, 99, 17–18.

48. Zhao, Y.; Zhang, Y.; Wan, X.; She, Y.; Li, M.; Xi, H.; Xie, J.; Wen, C. A novel ourmia-like mycovirus confers hypovirulence-associated traits on Fusarium oxysporum. Front. Microbiol. 2020, 11, 3084.

49. Abbas, A.A.; Taylor, L.J.; Dothard, M.I.; Leiby, J.S.; Fitzgerald, A.S.; Khatib, L.A.; Collman, R.G.; Bushman, F.D. Redondoviridae, a family of small, circular DNA viruses of the human oro-respiratory tract that are associated with periodontitis and critical illness. Cell Host Microbe 2019, 25, 719–729.

50. Van De Sande, W.W.J.; Lo-Ten-Foe, J.R.; Van Belkum, A.; Netea, M.G.; Kullberg, B.J.; Vonk, A.G. Mycoviruses: Future therapeutic agents of invasive fungal infections in humans? Eur. J. Clin. Microbiol. Infect. Dis. 2010, 29, 755–763.

51. Kondo, H.; Botella, L.; Suzuki, N. Mycovirus diversity and evolution revealed/inferred from recent studies. Annu. Rev. Phytopathol. 2022, 60.

52. Pearson, M.N.; Beever, R.E.; Boine, B.; Arthur, K. Mycoviruses of filamentous fungi and their relevance to plant pathology. Mol. Plant Pathol. 2009, 10, 115–128.

53. Rokas, A. Evolution of the human pathogenic lifestyle in fungi. Nat. Microbiol. 2022, 7, 607–619.

54. Chiapello, M.; Rodríguez-Romero, J.; Nerva, L.; Forgia, M.; Chitarra, W.; Ayllón, M.A.; Turina, M. Putative new plant viruses associated with Plasmopara viticola-infected grapevine samples. Ann. Appl. Biol. 2020, 176, 180–191.

55. Bujarski, J.; Gallitelli, D.; García-Arenal, F.; Pallás, V.; Palukaitis, P.; Krishna Reddy, M.; Wang, A. ICTV virus taxonomy profile: Bromoviridae. J. Gen. Virol. 2019, 100, 1206–1207.

56. Adams, M.J.; Adkins, S.; Bragard, C.; Gilmer, D.; Li, D.; MacFarlane, S.A.; Wong, S.M.; Melcher, U.; Ratti, C.; Ryu, K.H. ICTV virus taxonomy profile: Virgaviridae. J. Gen. Virol. 2017, 98, 1999–2000.

57. Marzano, S.-Y.L.; Nelson, B.D.; Ajayi-Oyetunde, O.; Bradley, C.A.; Hughes, T.J.; Hartman, G.L.; Eastburn, D.M.; Domier, L.L. Identification of diverse mycoviruses through metatranscriptomics characterization of the viromes of five major fungal plant pathogens. J. Virol. 2016, 90, 6846–6863.

58. Andika, I.B.; Wei, S.; Cao, C.; Salaipeth, L.; Kondo, H.; Sun, L. Phytopathogenic fungus hosts a plant virus: A naturally occurring cross-kingdom viral infection. Proc. Natl. Acad. Sci. 2017, 114, 12267–12272.

59. Roossinck, M.J. Evolutionary and ecological links between plant and fungal viruses. New Phytol. 2019, 221, 86–92.

60. Dolja, V. V.; Koonin, E. V. Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. Virus Res. 2018, 244, 36–52.

61. Asplund, M.; Kjartansdóttir, K.R.; Mollerup, S.; Vinner, L.; Fridholm, H.; Herrera, J.A.R.; Friis-Nielsen, J.; Hansen, T.A.; Jensen, R.H.; Nielsen, I.B.; et al. Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. Clin. Microbiol. Infect. 2019, 25, 1277–1285.

Chapter 7

General discussion

The growing need for computational solutions in virology

The first metagenomic high-throughput sequencing (HTS) study of viruses was published in 2006[1]. Experiments in that work yielded nearly two million sequence reads – more than three orders of magnitude higher than a 2002 shotgun cloning approach[2]. The rapid rise in HTS throughput has continued unabated however, with modern high-end instruments such as the Illumina NovaSeq 6000 capable of delivering 20 billion reads per run[3], more than 10,000 times 2006 capabilities. The corresponding paradigm shift in capacity for virus genome discovery has been invaluable for researchers; however, it has fundamentally changed how topics from group-level taxonomy to individual virus characterisation must be approached. With data generation far outstripping the ability to characterise viruses in the laboratory, computational biology must increasingly bridge the gaps. Since most newly identified viruses will likely *never* reach the laboratory, it is imperative that bridging methods reach accuracy comparable to gold standard laboratory analogues. Advances like the AlphaFold algorithm for atomic-scale protein structure prediction show this could be attainable even for difficult goals[4]. Elsewhere, compiled computational evidence can now approach experimental confidence levels for virus gene prediction[5], while automated taxonomic assignment can in some cases faithfully reproduce results of expert manual curation[6,7]. This thesis addressed additional challenges in virus computational biology, namely methods to discover viruses from metagenomic HTS data, and methods to identify hosts of viruses known only by their genome sequences (i.e., stray viruses).

A computational virus discovery workflow for VIDISCA-NGS

Protocols with diverse design rationales fall under the umbrella of metagenomic HTS; however, each is a laboratory procedure beginning with sampling and ending with nucleic acid sequencing. Subsequent data analysis could be considered as a module within the overall protocol, which itself must vary according to the specific research question and the type of data generated upstream. For example, metagenomic detection of a panel of pathogens will usually involve querying reads or contigs against reference nucleotide sequences using algorithms such as BWA or BLASTn. However, these would usually not be sensitive enough for discovery of unknown viruses. Because protein sequences are more conserved than nucleotides, protein-similarity-based algorithms such as DIAMOND or BLASTp may be used instead, along with known viral proteins for a reference. At remote distance, even protein sequence similarity is lost, and protein secondary or tertiary structure might be the only evidence of homology remaining, again requiring different algorithms such as DALI. If we visualise virus ‘sequence space’ as a dark landscape of unknown proteins, and known proteins as points of light, computational algorithms use these points to illuminate additional sequences nearby in space, which can then be used as future starting points. This allows iterative expansion in the number of known virus sequences and the lit area. Complicating this, researchers do not know the true size of the virus protein landscape, so some dark areas may currently be inaccessible or computationally expensive to access with any similarity-based tool[8,9]. In practise, some of these proteins can be

labelled as viral because they are found alongside known virus marker genes, but this 'guilt by association' cannot be relied on when no marker is found. Similarity-independent solutions are not widely applied in the literature, though methods to flag 'virus-like' sequences have been discussed, for example identification of contigs with unusually high open reading frame density – potentially indicative of a virus[10].

In **chapter 2** we presented an analysis workflow for VIDISCA-NGS data. This incorporated several methods discussed above, such as nucleotide-similarity-based detection of known viruses and protein-similarity-based discovery of unknown ones, plus separation and visualisation of these outputs for rapid sample overviewing. We also developed an approach capable of identifying some viruses escaping similarity-based detection. This relied on the assumption that virus infections produce genomes at high-copy number, and that restriction enzymes used in VIDISCA-NGS library preparation will cut these at identical locations. The resulting biological replicates, further replicated during the PCR stage, should then be at relatively high copy number in output reads. Our approach used clustering of similar reads to identify these 'virus-like' high-copy number sequences, and attempted to remove other entities behaving similarly, such as mitochondrial or ribosomal sequence. We demonstrated the utility of the method in principle via the detection of a novel gokushovirus in highly clustered reads. The concept could equally be applied to data generated by standard random shearing library preparation, except that rather than read clustering, *de novo* assembly followed by detection of high coverage contigs of unknown identity should be used instead. However, our method is not a complete solution. Some viruses aren't found at high-copy number, such as latent viruses and those not currently replicating for other reasons. Also, it is only a method to flag suspicious content, with follow-up required. If the viral identity is still not clear after full genome sequencing, this would need to include laboratory tests for a viral nature, such as filtration of the sample to remove large particles, digestion of background nucleic acids with nucleases, and PCR-based detection of virion protected RNA or DNA.

Computational solutions for stray virus host identification

Modern metagenomic HTS methods are delivering an accelerating torrent of virus genomic data (Figure 1). As a result, the newly recognised viral diversity has precipitated an overhaul of virus taxonomy norms[11], deemphasising data impractical to collect using metagenomics (e.g., virus phenotypic data gathered in a laboratory). Instead, sequence-based evolutionary analyses are now central[12]. The resulting standards expected of researchers have been released as consensus statements endorsed by the International Committee on Taxonomy of Viruses (ICTV)[11,12]. Meanwhile, field standards for other problems in virology have yet to fully adapt to the *status quo*. Principle among these is host identification for stray viruses found using metagenomics. The current gold standard of host confirmation remains isolation in axenic cell culture, but it is not uncommon for a host to be assumed based on the sampled organism, carrying a high risk of misassignment. As with taxonomy however, laboratory characterisation is an unrealistic standard for most viruses.

A miniscule fraction of Earth's cellular species richness is cultured, and a substantial proportion is likely intractable to culture. Even if this could be overcome, screening millions of cultures for virus infectivity is inconceivable for both economic and practical reasons. Solving this issue is of fundamental importance for virology. Coevolution with hosts is a key driver of virus diversification[13,14], and host identity is therefore vital to help inform sequence-based taxonomy. It additionally underpins our understanding of when historical host switches occurred, and efforts to understand why[15], relevant to zoonotic emergence and global health[16,17]. Host identity is a prerequisite to specifically evaluate virus ecological roles[18,19], and their potential medical or veterinary importance[20,21]. Stray virus host identification must therefore follow virus taxonomy by entering the high-throughput computational age, with appropriate methods developed for both prokaryote-infecting and eukaryote-infecting viruses; in this thesis we focused on the latter.

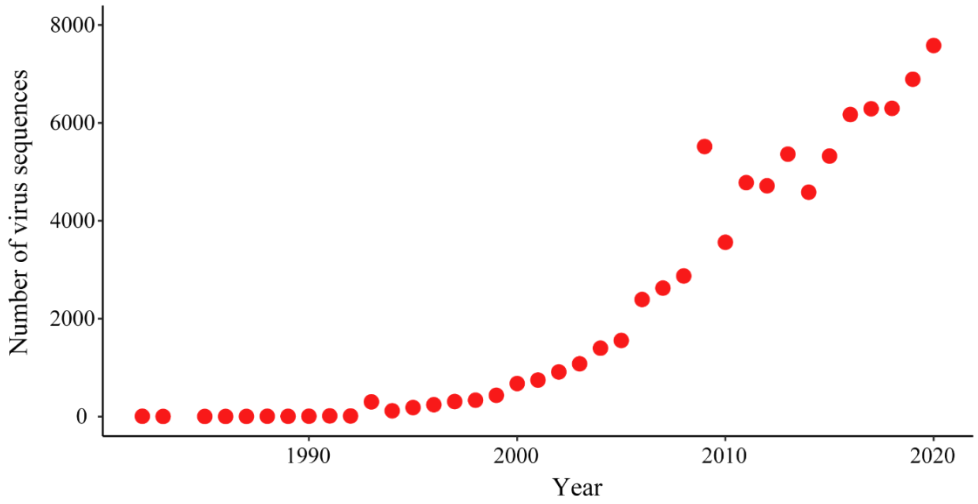


Figure 1. The accelerating virus sequence count on GenBank. A random subsample of virus sequences was taken with deposition dates between 1982 and 2020 (~83,000 sequences of ~11,000,000 in March 2023), and the frequency per year was plotted.

In **chapter 3** we presented the discovery of three families of eukaryote-infecting cressdnaviruses in human stool, the *Naryaviridae*, *Nenyaviridae*, and *Vilyaviridae*. We were able to identity their hosts as the human gut protists, *Entamoeba* for the first two families, and *Giardia* for the last. This work built directly upon the observation of possible cressdnavirus-derived endogenous viral elements (EVEs) in parasite genomes by Gibbs et al.[22]. Since EVEs are occasionally left behind in germline genomes during infection of a host, they are evidence of historical virus host ranges[23]. The emerging field of palaeovirology has already established that viruses of all Baltimore groups can leave a fossil record inside host genomes[24]. For the first time, we used EVEs to identify hosts of stray viruses, and supported this inference with additional evidence. If this method is to be

regularly used instead of cell culture isolation, quality standards must be adopted to take account of possible complications. First, are reported EVEs real features of a genome, or could they instead represent contamination of a genome assembly? Amplification of a putative EVE using PCR targeting the flanking regions is an appropriate solution, yet in many cases researchers don't have access to specimens or cultures to perform this. A hierarchy of computational evidence could be used instead, from examining read coverage across EVE junctions, to comparative genomics across independent assemblies or closely related species showing that integrations are homologous. Shifts in the GC content of EVEs versus exogenous viruses may be evidence of their old age, as could stop codons. If possible, EVE transcription or small RNA responses against them could serve as evidence of their veracity, as we argued in our study. Such quality control is essential and could be partly automated for high-throughput analysis of many assemblies. Second, while EVEs reveal historical infection, it is possible that horizontal shifts in host range occurred subsequently, and that virus lineages have since gone extinct within the original host. Furthermore, some very ancient EVE integrations occurred in host species that no longer exist, and are only conserved in their modern descendants. Independent evidence from modern sampling is therefore useful to link EVE-implicated hosts to modern exogenous viruses via physical cooccurrence, or ideally case-control analyses, as was done for many of the protist hosts described in this thesis. This is admittedly complex for environments such as seawater that lack natural compartmentalisation, when compared to human stool for example, which is compartmentalised by individual. In such cases, it is important for researchers to consider the host as part of the study design. Viral tagging, in which unknown viruses of specific hosts are labelled, and infected cells are sorted[25], or spatially-resolved library preparation methods such as single-cell sequencing[26] are some solutions to this.

A further complicating scenario is when no EVEs of eukaryotic viruses exist or can be detected. This will occur when the host genome has not yet been sequenced, or when EVE sequences are too degraded to be detected by alignment algorithms. Furthermore, it is probable that most virus species never left a genomic fossil record, as EVE integrations are generally rare events[23]. Here, alternative methods to identify hosts should be designed that do not rely on host genome assemblies. In **chapter 4** we reported such an approach, a computational workflow that works from a 'host agnostic' standpoint. The workflow carries out metagenomic analysis of target viruses and also ribosomal RNA (rRNA) sequences present in a sample, and then comparatively identifies which taxa are commonly found in virus positive samples. This allows users to reduce the problem size from hundreds of possible host taxa in a sample, to a much smaller set of potential host candidates. Statistical analysis then reduces the shortlist further, allowing host predictions to be made. The main drawback of this approach is that it results in a host prediction rather than a confirmed host. However, a prediction enables researchers to design targeted experiments (including computational ones) to confirm a candidate, which might otherwise be logistically overwhelming. We showed this in our study, confirming two of four host predictions using EVEs (CRESSV1 and *Kirkoviridae*), and providing strong independent

case-control evidence for a third (*Redondoviridae*). Notably, four months after we published our prediction of *Entamoeba gingivalis* as the host of redondoviruses, the finding was independently confirmed by evidence of cell culture infection[21], suggesting our method approaches experimental accuracy. While the workflow does not require a host genome assembly, it does assume rRNA sequences of the host are available. In practise this allows for a much greater host breadth than assembly-based approaches, because millions of rRNA sequences are available versus thousands of genome assemblies. A drawback is that it requires a cohort of samples, in which some are virus negative and some positive, which again may not be realistic for some sample types lacking spatial compartmentalisation.

Recombination and horizontal gene transfer as tools for host delimitation

One difficulty of host identification for stray viruses is how to set the granularity of analysis. For example, should we aim to determine a host for each individual virus sequence, strain, or species (a very large problem size), or can we assume that closely related viruses share a host (making a smaller problem size, but with a risk of incorrectly assuming shared hosts). An advantage of collective analysis is increased statistical power when examining viruses found across a cohort, as it will increase the number of samples positive for a lineage. However, it is difficult to judge how much divergence between two viruses should be permitted for collective analysis, before we must assume they infect a different host. In **chapters 3 and 4** we developed approaches based on viral recombination to allow collective analysis of viruses without risking incorrect host assumption. In **chapter 3**, having identified *Entamoeba* as the host genus of both *Naryaviridae* and *Nenyaviridae*, we observed that recombination has occurred between the two families. This was intriguing, not least because of the large genetic distance between analogous genes carried by each, showing that the two families occasionally exchange complete replicative or structural modules without losing replication competence. Recombination is often considered as a confounding factor for phylogenetic analysis, but here we argued that it is also useful for the purposes of host determination. Recombination between viruses demonstrates a shared host, as it can only occur during viral replication of coinfecting strains or species. This rationale should allow grouping of viruses that overlap in host range, even prior to any host suspicion. Interestingly, this rationale is broadly analogous to an approach used in automated virus taxonomy assignment[6], consistent with host codivergence as a key driver of virus speciation[14]. In **chapter 4** we therefore applied this rationale prior to any host prediction analysis to determine which gastrointestinal cressdnavirus genomes likely overlapped in host range. We additionally included an analysis of virus sequence-level occurrence between different sample types and cohorts, as distribution biases proved helpful to reveal when a different host should likely be assumed.

As discussed, recombination involves genetic transfer between the same or distinct species. In the example of naryaviruses and nenyaviruses recombination does not result in an increase in the number of genes. Instead, complete genetic modules are replaced in the progeny of a ‘recipient’ virus (i.e., allelic replacement). Other kinds of inter-species genetic

transfer do increase the genetic material of the recipient[27], collectively referred to as horizontal gene transfer (HGT). An example of HGT already discussed is the integration of viruses inside eukaryotic genomes to form EVEs. Both allelic replacement and HGT are useful for revealing aspects of virus host identity. In **chapter 5**, we combined lessons learned from viral recombination and virus-to-host HGT, and extended them to a rare case of virus-to-virus HGT. We reported that members of the genus *Avipoxvirus*, dsDNA pathogens in the family *Poxviridae*, realm *Varidnaviria*, were recipients of *Rep* HGT from ssDNA cressdnaviruses, realm *Monodnaviria*. Donors belonged to the genus *Krikovirus* within the unofficial lineage CRESSV3, which we renamed as the *Draupnirviridae*. Because the hosts of avipoxviruses were already known to be birds and other saurians, we predicted the same hosts for krikoviruses, arguing that HGT between viruses must have occurred during a coinfection. A wide screen for EVEs was done to test this, confirming krikoviruses have infected saurians since at least the late Mesozoic Era, ~100 million years ago. Detection of HGT between distant viral realms is relatively rare, though it has influenced the biology and evolution of several viruses[28,29], and HGT events between RNA viruses and *Rep*-encoding plasmids are even thought to have given rise to the cressdnaviruses[30,31]. The apparent rarity of HGT between viral realms may be explained by the effects of these events on viral fitness. It is likely that the vast majority of virus-to-virus HGT events are lethal or confer substantial fitness costs, and are therefore rapidly purged from gene-dense virus genomes subject to stringent selection. This differs from EVEs, which often find themselves in gene-sparse eukaryotic genomes, large parts of which evolve neutrally. Many EVEs likely have little or no fitness impact on the host, and may be fixed by population genetic processes, with positive fitness effects not essential for long-term survival[32]. Extending this rationale, large scale surveys of virus-to-virus HGT should be conducted, as any verified events are likely to have enhanced the fitness of their recipients, explaining their conservation. These events may therefore represent a rich source of virus adaptive evolution case studies, in addition to potentially providing host determination data. The *Rep* genes we found captured by avipoxviruses (*apvReps*) are essentially conserved at the genus level, and show strong evidence of purifying selection at the sequence level. This is convincing evidence of functionality, and suggests they conferred increased fitness to the poxviruses. The relationship found between *apvRep* gene count and avipoxvirus genome size further suggests a role in pathogen adaptive evolution, though the details remain obscure. Further work on *apvRep* functions would be informative for understanding non-canonical roles of Rep proteins, since even though *apvReps* are highly conserved, they display a large variation in domain structure and often have inactivated functional motifs. This suggests that canonical functions are either absent or reduced in most cases. Genes such as *apvRep-1* encode just the endonuclease domain while *apvRep-5* encodes just the helicase domain, and understanding why both are conserved in different avipoxviruses is of interest. Furthermore, evaluating the potential pathogenicity of krikoviruses and confirming their transmission by mosquito vectors represent promising future research directions.

The evolutionary history of cressdnaviruses

Chapters 3, 4, and 5 all dealt with aspects of cressdnavirus evolution via exploration of their hosts. We proposed four novel families (*Naryaviridae*, *Nenyaviridae*, *Vilyaviridae*, and *Draupnirviridae*) of which the first three have been accepted by the ICTV[33]. We proposed eukaryotic hosts of eight lineages (the four above, plus *Redondoviridae*, *Kirkoviridae*, CRESSV1, and CRESSV19). For each lineage, some members infected various protist hosts, while in the case of *Draupnirviridae*, a protist-infecting lineage apparently also emerged in animals. This thesis has significantly expanded our awareness of the host ranges of cressdnaviruses, as previously only five lineages were known to infect eukaryotic hosts. The number of newly recognised protist-infecting viruses implies micro-eukaryotes have been important sources of cressdnavirus spillover into plants, fungi, and animals throughout history. The case of *Draupnirviridae* is particularly striking, as emergence of animal viruses from protist-infecting ones within a single cressdnavirus family could suggest an extreme age for the phylum as a whole. Indeed, we can date the genus *Krikovirus* to at least 100 million years ago, and it is possibly older. Relevant to this topic is work suggesting that the family *Smacoviridae* infects archaea[34], a surprising result that though unconfirmed in culture does have some emerging support[35,36]. While this theoretically could be explained by a primordial origin of some cressdnaviruses in prokaryotes, phylogenetic analysis instead tends to support a eukaryote to prokaryote host switch[30], running counter to the general assumption that viral exchange does not occur between eukaryotic and prokaryotic domains of life. Future work to resolve the origins of various cressdnaviruses should begin with more research into their hosts, since at least 42 stray lineages remain to be solved. Wider CRISPR screening may uncover additional prokaryote-infecting lineages, and it will be important to understand their phylogenetic distribution within the phylum, as this could elucidate where, when, and how many times they evolved from eukaryote-infecting viruses.

Cressdnaviruses vary in genome complexity, but include some of the smallest and simplest known viruses. For example, porcine circovirus 2 (PCV2) has a genome of only ~1,760 nt and a virion diameter of ~17 nm[37]. All cressdnaviruses share a homologous *Rep* gene[38], and detectable protein sequence similarity between these makes them the clear choice for phylum-level phylogenies. As a result, cressdnavirus taxonomy is heavily oriented around Rep sequence relationships. The other core gene possessed by all cressdnaviruses is *Cap*, encoding the capsid protein (Cap). All Caps share a characteristic single jelly roll fold, but often have no detectable sequence similarity as a result of relatively rapid sequence evolution and a polyphyletic evolutionary history[29,39,40]. As a result it is not uncommon to see them described as evolutionarily non-homologous, yet structural analyses tend to suggest they are extremely distant relatives nested within RNA virus capsid lineages[39]. As mentioned, it is thought that cressdnaviruses originated from multiple independent Cap to plasmid HGT events, with further Cap replacements by other RNA viruses following later[29,40]. Unpublished analyses not presented in this thesis suggest that cressdnavirus Caps can be grouped into a handful of major classes, e.g., circo-like, gemini-like, and noda-like.

If homologous, it would be interesting to uncover their evolutionary relationships using structure-based methods such as homologous structure finder[41], however it may be that useful evolutionary information has already been obscured over time. For understanding the broad-scale determinants of cressdnavirus host tropism and major historical host switches, Cap evolutionary history is likely more relevant than that of Rep. As the structural protein, Cap is involved in host interaction and presumably is the main determinant of host range. Even if accurate phylum-level analyses are not possible, further family-level analyses could be informative to unravel host tropism for individual species, for example among the many stray members of *Circoviridae*. Data on Cap relationships will also provide additional resolution to evolutionary hypotheses on the origins of cressdnaviruses[30]. Only a few cressdnavirus Cap proteins have experimentally solved structures, yet the arrival of atomic-scale structure prediction software such as AlphaFold[4] or ESM-2[42] may allow sufficient throughput to explore this question.

From host identification to virus characterisation

Upon identification of stray virus hosts using computational methods, in some cases virus characterisation may become feasible. For example, the parasitic hosts of *Naryaviridae*, *Nenyaviridae*, and *Vilyaviridae* (*Entamoeba* and *Giardia*) can all be cultured in the laboratory, and animal infection models exist. *In vitro* infection experiments with pure virus are potentially feasible, allowing the study of any effects on the host. If viruses cause significant cytopathic effect, lysis, or culture debilitation, clinical impact studies could be warranted. The determinants of *E. histolytica* invasive disease remain partly unsolved[43], and any hypovirulent or hypervirulent effects of virus infection would be of high interest to understand in a clinical context. The 2019 discovery of redondoviruses and their strong association to human gum disease (periodontitis)[20] represents another example. Our finding that redondoviruses infect *E. gingivalis* explains this association, as the host is itself tightly correlated with periodontitis[44]. However, it is not known if redondovirus infection has an impact on *E. gingivalis* pathogenicity, and thus an indirect role in human disease. In **chapter 6**, we looked for viruses infecting human-pathogenic fungi (mycoviruses) that were frozen in a clinical biobank at the Amsterdam UMC. While we did not find mycoviral cressdnaviruses, we did identify a number of RNA viruses. Exploration of their pathogenicity to hosts could similarly be interesting in the context of virus-therapy, though we suspect RNA mycoviruses are not ideal candidates for this as all characterised thus far lack an extracellular stage. Cressdnaviruses of the family *Genomoviridae* could be the ideal candidates for biological control of fungi, as they have an extracellular stage and at least one confers strong hypovirulence on its fungal host[45]. This topic echoes widely discussed concerns of a post-antibiotic world, in which bacterial resistance eventually negates conventional treatment options. This possibility has revived the idea of phage-therapy for treatment of bacterial infections, as viruses can evolve around their hosts defence mechanisms, and have been doing so for billions of years. It is striking that 100-year-old ideas of virologists[46,47] may return to prominence again, underscoring not only their insight, but also the outsized impact that tiny viruses have on the wider world.

References

1. Angly, F. E. et al. The marine viromes of four oceanic regions. PLOS Biol. 4, e368 (2006).

2. Breitbart, M. et al. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. 99, 14250–14255 (2002).

3. Illumina. Output per flow cell for various read lengths, NovaSeq 6000 system. (2023). Available at: https://emea.illumina.com/systems/sequencing-platforms/novaseq/specifications.html. (Accessed: 17th March 2023)

4. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).

5. Olo Ndela, E. et al. Reekeekee- and roodoodooviruses, two different Microviridae clades constituted by the smallest DNA phages. Virus Evol. 9, veac123 (2023).

6. Bin Jang, H. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat. Biotechnol. 37, 632–639 (2019).

7. Aiewsakun, P. & Simmonds, P. The genomic underpinnings of eukaryotic virus taxonomy: Creating a sequence-based framework for family-level virus classification. Microbiome 6, 1–24 (2018).

8. Boratto, P. V. M. et al. Yaravirus: A novel 80-nm virus infecting Acanthamoeba castellanii. Proc. Natl. Acad. Sci. 117, 16579–16586 (2020).

9. Perdigão, N. et al. Unexpected features of the dark proteome. Proc. Natl. Acad. Sci. 112, 15898–15903 (2015).

10. Lauber, C. & Seitz, S. Opportunities and challenges of data-driven virus discovery. Biomolecules 12, 1073 (2022).

11. Simmonds, P. et al. Virus taxonomy in the age of metagenomics. Nat. Rev. Microbiol. 15, 161–168 (2017).

12. Simmonds, P. et al. Four principles to establish a universal virus taxonomy. PLOS Biol. 21, e3001922 (2023).

13. Ghafari, M., Simmonds, P., Pybus, O. G. & Katzourakis, A. A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses. Curr. Biol. 31, 4689-4696.e5 (2021).

14. Simmonds, P., Aiewsakun, P. & Katzourakis, A. Prisoners of war — host adaptation and its constraints on virus evolution. Nat. Rev. Microbiol. 17, 321–328 (2018).

15. Aiewsakun, P. & Katzourakis, A. Marine origin of retroviruses in the early Palaeozoic Era. Nat. Commun. 8, 1–12 (2017).

16. Parry, R. & Asgari, S. Discovery of novel crustacean and cephalopod flaviviruses: Insights into the evolution and circulation of flaviviruses between marine invertebrate and vertebrate hosts. J. Virol. 93, 432–451 (2019).

17. Van Heuverswyn, F. & Peeters, M. The origins of HIV and implications for the global epidemic. Curr. Infect. Dis. Rep. 9, 338–346 (2007).

18. Shiah, F. K. et al. Viral shunt in tropical oligotrophic ocean. Sci. Adv. 8, 2829 (2022).

19. Wilhelm, S. W. & Suttle, C. A. Viruses and nutrient cycles in the sea: Viruses play critical roles in the structure and function of aquatic food webs. Bioscience 49, 781–788 (1999).

20. Abbas, A. A. et al. Redondoviridae, a family of small, circular DNA viruses of the human oro-respiratory tract that are associated with periodontitis and critical illness. Cell Host Microbe 25, 719–729 (2019).

21. Keeler, E. L. et al. Widespread, human-associated redondoviruses infect the commensal protozoan Entamoeba gingivalis. Cell Host Microbe 31, 58-68.e5 (2023).

22. Gibbs, M. J., Smeianov, V. V., Steele, J. L., Upcroft, P. & Efimov, B. A. Two families of Rep-like genes that probably originated by interspecies recombination are represented in viral, plasmid, bacterial, and parasitic protozoan genomes. Mol. Biol. Evol. 23, 1097–1100 (2006).

23. Katzourakis, A. & Gifford, R. J. Endogenous viral elements in animal genomes. PLoS Genet. 6, e1001191 (2010).

24. Barreat, J. G. N. & Katzourakis, A. Paleovirology of the DNA viruses of eukaryotes. Trends Microbiol. 30, 281–292 (2022).

25. Deng, L. et al. Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. Nature 513, 242–245 (2014).

26. Yoon, H. S. et al. Single-cell genomics reveals organismal interactions in uncultivated marine protists. Science 332, 714–717 (2011).

27. Lawrence, J. G. & Retchless, A. C. The interplay of homologous recombination and horizontal gene transfer in bacterial speciation. Methods Mol. Biol. 532, 29–53 (2009).

28. Caselli, E. et al. Human herpesvirus 6 (HHV-6) U94/REP protein inhibits betaherpesvirus replication. Virology 346, 402–414 (2006).

29. Diemer, G. S. & Stedman, K. M. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. Biol. Direct 7, 1–14 (2012).

30. Kazlauskas, D., Varsani, A., Koonin, E. V. & Krupovic, M. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. Nat. Commun. 10, 1–12 (2019).

31. Krupovic, M. Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. Current Opinion in Virology 3, 578–586 (2013).

32. Holmes, E. C. The evolution of endogenous viral elements. Cell Host Microbe 10, 368–377 (2011).

33. Krupovic, M. & Varsani, A. Naryaviridae, Nenyaviridae, and Vilyaviridae: Three new families of single-stranded DNA viruses in the phylum Cressdnaviricota. Arch. Virol. 167, 2907–2921 (2022).

34. Díez-Villaseñor, C. & Rodriguez-Valera, F. CRISPR analysis suggests that small circular single-stranded DNA smacoviruses infect Archaea instead of humans. Nat. Commun. 10, 294 (2019).

35. Medvedeva, S., Borrel, G., Krupovic, M. & Gribaldo, S. A global virome of methanogenic archaea highlights novel diversity and adaptations to the gut environment. Res. Sq. (2023).

36. Li, R., Wang, Y., Hu, H., Tan, Y. & Ma, Y. Metagenomic analysis reveals unexplored diversity of archaeal virome in the human gut. Nat. Commun. 13, 1–12 (2022).

37. Ellis, J. et al. Isolation of circovirus from lesions of pigs with postweaning multisystemic wasting syndrome. Can. Vet. J. 39, 44–51 (1998).

38. Ilyina, T. V. & Koonin, E. V. Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. Nucleic Acids Res. 20, 3279–3285 (1992).

39. Krupovic, M. & Koonin, E. V. Multiple origins of viral capsid proteins from cellular ancestors. Proc. Natl. Acad. Sci. 114, E2401–E2410 (2017).

40. Kazlauskas, D. et al. Evolutionary history of ssDNA bacilladnaviruses features horizontal acquisition of the capsid gene from ssRNA nodaviruses. Virology 504, 114–121 (2017).

41. Ravantti, J., Bamford, D. & Stuart, D. I. Automatic comparison and classification of protein structures. J. Struct. Biol. 183, 47–56 (2013).

42. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130 (2023).

43. Faust, D. M. & Guillen, N. Virulence and virulence factors in Entamoeba histolytica, the agent of human amoebiasis. Microbes Infect. 14, 1428–1441 (2012).

44. Bao, X., Weiner, J., Meckes, O., Dommisch, H. & Schaefer, A. S. Entamoeba gingivalis exerts severe pathogenic effects on the oral mucosa. J. Dent. Res. 100, 771–776 (2021).

45. Yu, X. et al. A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. Proc. Natl. Acad. Sci. 107, 8387–8392 (2010).

46. D'Hérelle, F. Bacteriophage as a treatment in acute medical and surgical infections. Bull. N. Y. Acad. Med. 7, 329–348 (1931).

47. Twort, F. W. An investigation on the nature of ultra-microscopic viruses. Lancet 186, 1241–1243 (1915).

Summary

Computational discovery of viruses and their hosts

The field of virology is now 125 years old, and during that time it has experienced radical evolution in the techniques available for discovering unknown viruses, summarised in **chapter 1**. During the last two decades alone, metagenomic high-throughput sequencing (HTS) methods have transformed our understanding of virus diversity and evolution. They achieved this by enabling efficient discovery of virus genomes from diverse environments – without their isolation in host cell culture. A key disadvantage of this is that the host species identity is usually unknown, meaning the full evolutionary, ecological, or medical significance of findings cannot be realised.

This thesis first focused on developing computational tools to analyse metagenomic HTS data and find previously unknown viruses. In **chapter 2**, a new computational workflow was developed incorporating both sequence-dependent and sequence-independent virus identification techniques, and this was subsequently applied in various research projects.

In metagenomic HTS studies, it is currently typical for hosts of newly discovered viruses to go unreported, since these are considered difficult to identify. With the gold standard of host identification (cell culture isolation) being highly impractical for the majority of metagenomically identified viruses, the second and main focus of this thesis was to develop methods to computationally pinpoint virus hosts. In **chapter 3**, we discovered three families of circular ssDNA viruses (i.e., cressdnaviruses) in human stool samples, naming them the *Naryaviridae*, *Nenyaviridae*, and *Vilyaviridae*. We found that each family had historically integrated genetic material inside the genomes of respective hosts (*Entamoeba* and *Giardia* parasites), leaving so-called endogenous viral elements (EVEs), some of which may now function in antiviral defence. Using case-control analyses we could show that the association between the viruses and respective hosts still exists today, and also found that different virus families infecting the same host had recombined with each other – a phenomenon that we subsequently applied as a tool for demonstrating shared host ranges.

In **chapter 4**, we considered how hosts might be found when EVEs were not detected upstream. We reasoned that if a virus was replicating in a sampled environment, then its host must be present also, and that across a large cohort of samples (some virus negative, some positive) a detectable co-occurrence pattern should exist between virus and host. We developed a ‘host agnostic’ computational workflow that began with viral and eukaryotic metagenomic analyses, before reporting of over-represented eukaryotes in virus positive samples (i.e., potential hosts). Subsequent statistical analysis allowed host

prediction and targeted confirmation. Applying this methodology, we predicted protistan hosts for four further lineages of gastrointestinal cressdnaviruses, including *Entamoeba gingivalis* as the host of the *Redondoviridae* – discovered recently in humans and strongly-linked to gum disease. Of the four host predictions, independent evidence has now confirmed three. This study highlighted the utility of both recombination and horizontal gene transfer for host delimitation purposes.

In **chapter 5** we extended this utility, finding that a group of cressdnaviruses donated their *Rep* gene to poxviruses infecting birds and other saurian relatives. We reasoned that for this donation to occur, the two virus lineages must have shared the same vertebrate host range. Among cressdnaviruses only the family *Circoviridae* had representatives known to infect birds, and we therefore assumed these were the likely donors, but were surprised to find they instead belonged to an unclassified lineage, CRESSV3. We renamed this lineage as the family *Draupnirviridae*, and showed that the genus *Krikovirus* was the sublineage interacting with poxviruses. By looking for *Krikovirus* EVEs in putative host lineages, we confirmed the prediction that they have a vertebrate-tropism, showing that they left traces in some saurian genomes over 100 million years ago. Further, we found that the avian-infecting poxviruses likely gained fitness advantages upon receiving *Rep* copies from krikoviruses.

In **chapter 6** we explored the viral content of medically important fungal isolates, with a particular interest in identifying new cressdnaviruses, since these represent a potential source of anti-fungal biocontrol agents. While we did not find these, we did identify a wealth of novel RNA virus diversity. Within this was one representative of a group never previously observed infecting fungi, the jiviviruses.

Finally, in **chapter 7**, the results were discussed in light of the current academic literature, and ideas for future research were presented.

Samenvatting

Computationele ontdekking van virussen en hun gastheren

Het veld van de virologie is nu 125 jaar oud en heeft in die tijd een radicale evolutie doorgemaakt in de technieken die beschikbaar zijn voor het ontdekken van onbekende virussen, samengevat in **hoofdstuk 1**. Alleen al de afgelopen twee decennia hebben metagenomische high-throughput sequencing (HTS)-methoden ons begrip van virusdiversiteit en evolutie getransformeerd. Deze methoden hebben het mogelijk gemaakt om virusgenomen uit verschillende omgevingen op een efficiënte manier te detecteren - zonder dat deze moeten worden gekweekt in cellen. Een cruciaal nadeel hiervan is dat de identiteit van de gastheersoort meestal onbekend blijft, wat betekent dat de volledige evolutionaire, ecologische of medische betekenis van bevindingen niet kan worden bepaald.

Ten eerste richtte dit proefschrift zich op het ontwikkelen van computationele tools om metagenomische HTS-gegevens te analyseren en voorheen onbekende virussen te vinden. In **hoofdstuk 2** werd een nieuwe computationele workflow ontwikkeld waarin zowel sequentie-afhankelijke als sequentie-onafhankelijke virusidentificatietechnieken zijn opgenomen, en deze werkwijze werd vervolgens toegepast in verschillende onderzoeksprojecten.

In metagenomische HTS-onderzoeken wordt momenteel de gastheer van nieuw ontdekte virussen meestal niet gerapporteerd, aangezien ze moeilijk te identificeren zijn. Doordat de gouden standaard van gastheeridentificatie (isolatie met behulp van celkweek) niet praktisch is voor de meeste metagenomisch geïdentificeerde virussen, lag de tweede en belangrijkste focus van dit proefschrift op het ontwikkelen van methoden om virus gastheren computationeel te identificeren. In **hoofdstuk 3** ontdekten we drie families van circulaire ssDNA-virussen (cressdnavirussen) in menselijke ontlasting, die we de *Naryaviridae*, *Nenyaviridae* en *Vilyaviridae* noemden. We hebben ontdekt dat elk van deze drie virus families genetisch materiaal integreerde in het genoom van de gastheren (*Entamoeba*- en *Giardia*-parasieten), waardoor endogene virale elementen (EVE's) zijn ontstaan. Deze EVE’s zijn mogelijk onderdeel van een antivirale afweer. Met behulp van case-control analyses hebben we aangetoond dat de associatie tussen de virussen en respectievelijke gastheren nog steeds bestaat, en ook dat virusfamilies die dezelfde gastheer delen kunnen recombineren– een fenomeen dat we vervolgens als hulpmiddel hebben toegepast om gedeelde gastheren aan te tonen.

In **hoofdstuk 4** hebben we bekeken hoe de gastheer kan worden aangetoond als er geen EVE’s aanwezig zijn. We redeneerden dat als een virus zich repliceert in een specifieke omgeving, de gastheer ook aanwezig moet zijn in die omgeving, en dat er in een groot cohort aan monsters (met virus positieve en negatieve klinische materialen) een

detecteerbaar patroon zou moeten bestaan tussen virus en gastheer. We ontwikkelden een 'gastheer-agnostische' computationele workflow waarin eerst virale en eukaryotische metagenomische analyses werden uitgevoerd, waardoor oververtegenwoordigde eukaryoten (d.w.z. potentiële gastheren) in virus-positieve monsters te herkennen zijn. Vervolgens werd door middel van statistische analyse de gastheer voorspeld. Door deze methodologie toe te passen voorspelden we de gastheren voor vier lijnen van gastro-intestinale cressdnavirussen. Een van deze gastheren is *Entamoeba gingivalis* als gastheer van de *Redondoviridae. Entamoeba gingivalis* is een parasiet die sterk verbonden is met parodontitis. Van de vier gastheervoorspellingen zijn er tot nu toe drie bevestigd met onafhankelijk bewijs. Deze studie benadrukt het nut van zowel recombinatie-onderzoek als horizontale genoverdracht-onderzoek om de gastheer van een virus te identificeren.

In **hoofdstuk 5** breidden we deze toepassing uit, en ontdekten we dat een groep cressdnavirussen hun *Rep*-gen doneert aan sommige pokkenvirussen. Het gaat om pokkenvirussen die vogels en andere saurische verwanten infecteren. We redeneerden dat, om deze overdracht te laten plaatsvinden, de twee viruslijnen eenzelfde scala aan gewervelde gastheren moeten hebben gedeeld. Van de cressdnavirussen zijn alleen vertegenwoordigers van de *Circoviridae* familie bekend die vogels kunnen infecteren, en dus de meest waarschijnlijke donoren. We kwamen echter tot de verassende ontdekking dat niet de *Circoviridae* de donorvirussen waren, maar een niet-geclassificeerde cressdnavirussen lijn (CRESSV3). We hernoemden deze lijn de *Draupnirviridae* familie, en toonden aan dat het *Krikovirus* geslacht de sublijn was die de interactie had met pokkenvirussen. Door te zoeken naar EVE's van het *Krikovirus* in vermeende gastheren, bevestigden we dat ze inderdaad een vertebraat-tropisme hebben. Meer dan 100 miljoen jaar geleden hebben deze virussen al sporen achtergelaten in enkele saurische genomen. Daarnaast hebben we ontdekt dat de pokkenvirussen die vogels kunnen infecteren waarschijnlijk een fitness voordeel hadden na ontvangst van krikovirus *Rep*-gen kopieën.

In **hoofdstuk 6** onderzochten we de virussen die medisch belangrijke schimmel isolaten infecteren, met een bijzondere interesse in het identificeren van nieuwe cressdnavirussen, aangezien deze een potentiële bron van antischimmel bestrijdingsmiddelen vormen. Hoewel we cressdnavirussen niet hebben gevonden, hebben we wel een grote hoeveelheid nieuwe RNA-virusdiversiteit geïdentificeerd. Hierin bevond zich een vertegenwoordiger van een groep virussen waarvan nooit eerder was aangetoond dat ze schimmels kunnen infecteren, namelijk de jivivirussen.

In **hoofdstuk 7** worden de resultaten van dit proefschrift besproken in het licht van de huidige academische literatuur en ideeën voor toekomstig onderzoek gepresenteerd.

Author affiliations

Aldert Bart*, Amsterdam UMC, Laboratory of Clinical Parasitology, Department of Medical Microbiology and Infection Prevention, University of Amsterdam, Meibergdreef 9, Amsterdam 1105 AZ, The Netherlands; Amsterdam Institute for Infection and Immunity, Postbus 22660, Amsterdam 1100 DD, The Netherlands.

*Present address: Department of Medical Microbiology, Tergooi MC, Van Riebeeckweg 212, Hilversum 1213 XZ, The Netherlands.

Christin Becker, Department of Periodontology, Oral Surgery and Oral Medicine, Institute for Dental and Craniofacial Sciences, Berlin Institute of Health, Charité—Universitätsmedizin Berlin, Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin, Germany.

Patricia Broekhuizen, Amsterdam UMC, Laboratory of Clinical Parasitology, Department of Medical Microbiology and Infection Prevention, University of Amsterdam, Meibergdreef 9, Amsterdam 1105 AZ, The Netherlands.

Matthew Cotton, MRC/UVRI & LSHTM Uganda Research Unit, 3FC6+Q3, Entebbe, Uganda; MRC-University of Glasgow Centre for Virus Research, G61 1QH, Glasgow, UK.

Martin Deijs, Amsterdam UMC, Laboratory of Experimental Virology, Department of Medical Microbiology and Infection Prevention, University of Amsterdam, Meibergdreef 9, Amsterdam 1105 AZ, The Netherlands; Amsterdam Institute for Infection and Immunity, Postbus 22660, Amsterdam 1100 DD, The Netherlands.

Marieke Gittelbauer, Amsterdam UMC, Laboratory of Mycology, Department of Medical Microbiology and Infection Prevention, University of Amsterdam, Meibergdreef 9, Amsterdam 1105 AZ, The Netherlands.

Maarten F. Jebbink, Amsterdam UMC, Laboratory of Experimental Virology, Department of Medical Microbiology and Infection Prevention, University of Amsterdam, Meibergdreef 9, Amsterdam 1105 AZ, The Netherlands; Amsterdam Institute for Infection and Immunity, Postbus 22660, Amsterdam 1100 DD, The Netherlands.

Joanna Kaczorowska, Amsterdam UMC, Laboratory of Experimental Virology, Department of Medical Microbiology and Infection Prevention, University of Amsterdam, Meibergdreef 9, Amsterdam 1105 AZ, The Netherlands; Amsterdam Institute for Infection and Immunity, Postbus 22660, Amsterdam 1100 DD, The Netherlands.

Arne S. Schaefer, Department of Periodontology, Oral Surgery and Oral Medicine, Institute for Dental and Craniofacial Sciences, Berlin Institute of Health, Charité—

Universitätsmedizin Berlin, Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin, Germany.

Lia van der Hoek, Amsterdam UMC, Laboratory of Experimental Virology, Department of Medical Microbiology and Infection Prevention, University of Amsterdam, Meibergdreef 9, Amsterdam 1105 AZ, The Netherlands; Amsterdam Institute for Infection and Immunity, Postbus 22660, Amsterdam 1100 DD, The Netherlands.

Karin van Dijk, Amsterdam UMC, Laboratory of Mycology, Department of Medical Microbiology and Infection Prevention, University of Amsterdam, Meibergdreef 9, Amsterdam 1105 AZ, The Netherlands.

Tom van Gool, Amsterdam UMC, Laboratory of Clinical Parasitology, Department of Medical Microbiology and Infection Prevention, University of Amsterdam, Meibergdreef 9, Amsterdam 1105 AZ, The Netherlands; Amsterdam Institute for Infection and Immunity, Postbus 22660, Amsterdam 1100 DD, The Netherlands.

Author contributions

Chapter 2:

Conceptualisation, CMK, LvdH; methodology, CMK, LvdH, MD; software, CMK; investigation, CMK, MD; visualisation, CMK; supervision, LvdH; funding acquisition, LvdH; writing - original draft, CMK; writing - review and editing, CMK, LvdH.

Chapter 3:

Conceptualisation, CMK, LvdH; methodology, CMK, LvdH; software, CMK; investigation, CMK, LvdH, MC, AB, MD, PB, JK, MJ; writing - original draft, CMK; visualisation, CMK; resources, LvdH, TG; supervision, LvdH; funding acquisition, LvdH; writing - review and editing: CMK, LvdH, AB, MC, JK, TG.

Chapter 4:

Conceptualisation, CMK, LvdH, AS; methodology, CMK; software, CMK; validation, CMK, MD, LvdH, AS, CB; formal analysis, CMK; investigation, CMK, MD, LvdH, PB, CB, AB; resources, LvdH, AS, TvG, PB; data curation, CMK; writing - original draft, CMK; writing - review and editing, CMK, LvdH, AB, AS, TG, PB; visualisation, CMK; supervision, LvdH, AS; project administration, CMK; funding acquisition, LvdH.

Chapter 5:

Conceptualisation, CMK; data curation, CMK; formal analysis, CMK; funding acquisition, LvdH; investigation, CMK; methodology, CMK; software, CMK; supervision; LvdH, visualisation, CMK; writing – original draft, CMK; writing – review and editing, CMK, LvdH.

Chapter 6:

Conceptualisation, KvD, LvdH, CMK; formal analysis, CMK; investigation, MD, HMG, CMK; resources, KvD, LvdH; data curation, CMK; writing - original draft preparation, CMK, LvdH; writing - review and editing, CMK, LvdH, KvD, MD, HMG; visualisation, CMK; supervision, LvdH; funding acquisition, LvdH.

About the author

Cormac Michael Kinsella was born in the United Kingdom on the 5th of October 1991. His academic studies began with a BSc in zoology at Cardiff University. It was there he first got a taste for academic research, with Cardiff offering students the opportunity to conduct a research training year at a separate institution. Cormac spent that year at the University of Bristol, working on dinosaur evolution with Prof. Michael Benton. He then returned to Cardiff to work on lepidopteran life history with Dr. Mark Jervis, and eventually fish parasitology and evolutionary ecology with Dr. Jessica Stephenson and Prof. Joanne Cable. This latter experience, along with an excellent parasitology course given by Prof. Cable, sparked an interest in parasite and pathogen evolution, which led him to enrol on a research-oriented MSc in evolutionary biology. During this he worked on an avian 'genomic parasite' at Uppsala University with Dr. Alexander Suh, at LMU München with Prof. Michael Boshart on *Trypanosoma* parasites, and at Harvard University and the Broad Institute on Jamestown Canyon virus with Dr. Anne Piantadosi and Prof. Pardis Sabeti. During this project he identified previously unknown viruses in mosquito RNA sequencing data, beginning an appreciation of the vast diversity of viruses. As a result, he became interested in their discovery, pursuing the topic for his doctorate with Dr. Lia van der Hoek. With PhD studies complete, Cormac will begin postdoctoral research with Prof. Aris Katzourakis at the University of Oxford, studying the evolution of various virus groups.

PhD portfolio

PhD student: Cormac M. Kinsella
PhD period: 2017-2023
Promoter: Dr. Lia van der Hoek
Copromoters: Prof. Dr. Ben Berkhout
Dr. Aldert Bart

1. PhD training

General courses	Year	ECTS
AMC crash course	2018	0.2

Specific courses	Year	ECTS
Media training	2017	0.6
Infectious diseases	2017	1.2
Flavivirus serology	2018	0.9
Protein production	2018	1.2
Direct next-generation sequencing after shearing	2018	1.5
High throughput RT-PCR and degenerative oligonucleotide design	2018	1.5
Virus discovery using VIDISCA next-generation sequencing	2018	1.5
Advanced virus discovery using VIDISCA next-generation sequencing	2018	1.8
Primate animal models	2018	1.2

High-performance computing	2018	0.5
Novel diagnostics	2019	0.9
Bioinformatics	2019	1.1
Microarrays for virus detection	2019	0.9
Commercialisation of vaccines	2019	1.5

Seminars & workshops	**Year**	**ECTS**
Weekly department seminars	2017-2021	1
Weekly PhD student seminars	2017-2021	1
HONOURs annual meetings	2017-2019	1
Outbreak antenna meeting	2017	0.3
24th International Bioinformatics Workshop on Virus Evolution and Molecular Epidemiology (VEME)	2019	1
CADDE Genomic Epidemiology Workshop	2019	1

Oral presentations	**Year**	**ECTS**
Invited talk, University of Exeter	2019	0.5
3rd Uppsala Transposon Symposium	2019	1
International Virus Bioinformatics Meeting, online	2020	0.5
International Symposium on ssDNA Viruses, Sète	2022	1
6th Uppsala Transposon Symposium	2022	1

Invited talk, European Virus Bioinformatics Center, ECR viromics webinar series	2023	0.5
European Congress of Virology, Gdańsk	2023	1

Poster presentations	**Year**	**ECTS**
European Congress of Virology, Rotterdam	2019	0.5
24th International Bioinformatics Workshop on Virus Evolution and Molecular Epidemiology (VEME), Hong Kong	2019	NA
CADDE Genomic Epidemiology Workshop, São Paulo	2019	NA

2. Teaching

Supervising	**Year**	**ECTS**
Nadine van Kleef, BSc stage	February - June 2019	2
Lorenzo Hogendorp, HBO stage	October 2019 - June 2020	2
Tianne Spreij, MSc stage	February 2020 - April 2021	2
Hamed Alahmad, MSc stage	January - June 2022	2

Lecturing	**Year**	**ECTS**
Virus discovery using VIDISCA next-generation sequencing	2018	0.5

3. Parameters of Esteem

Awards & grants	Year
Best poster award, 24th International Bioinformatics Workshop on Virus Evolution and Molecular Epidemiology (VEME) evolutionary hypothesis testing module	2019
Best poster award, CADDE Project Genomic Epidemiology Workshop	2019
Amsterdam institute for Infection and Immunity, €500 travel grant for IS3DV	2022
Best talk award (runner up), 6th Uppsala Transposon Symposium	2022

List of publications

Peer-reviewed publications (this thesis)

PNAS, 2023	**Kinsella, C.M.**, van der Hoek, L. Vertebrate-tropism of a cressdnavirus lineage implicated by poxvirus gene capture.
Virus Evolution, 2022	**Kinsella, C.M.**, Deijs, M., Becker, C., Broekhuizen, P., van Gool, T., Bart, A., Schaefer, A., van der Hoek, L. Host prediction for disease-associated gastrointestinal cressdnaviruses.
Microbiology Spectrum, 2022	**Kinsella, C.M.**, Deijs, M., Gittelbauer, H.M., Jebbink, M.F., van Dijk, K., van der Hoek, L. Human clinical isolates of pathogenic fungi are host to diverse mycoviruses.
Nature Communications, 2020	**Kinsella, C.M.**, Bart, A., Deijs, M., Broekhuizen, P., Kaczorowska, J., Jebbink, M.F., van Gool, T., Cotten, M., van der Hoek, L. *Entamoeba* and *Giardia* parasites implicated as hosts of CRESS viruses.
Virus Research, 2019	**Kinsella, C.M.**, Deijs, M., van der Hoek, L. Enhanced bioinformatic profiling of VIDISCA libraries for virus detection and discovery.

Peer-reviewed publications (other)

*Shared first authorship

Virus Evolution, 2023	Kaczorowska, J., Timmerman, A.L., Deijs, M., **Kinsella, C.M.**, Bakker, M., van der Hoek, L. *Anellovirus evolution during long-term chronic infection*.
Fluids and Barriers of the CNS, 2022	**Kinsella, C.M.***, Edridge, A.W.D.*, van Zeggeren, I.E.*, Deijs, M., van de Beek, D., Brouwer, M.C., van

	der Hoek, L. *Bacterial ribosomal RNA detection in cerebrospinal fluid using a viromics approach*.
Journal of Virology, 2022	Kaczorowska, J., Deijs, M., Klein, M., Bakker, M., Jebbink, M.F., Sparreboom, M., **Kinsella, C.M.**, Timmerman, A.L., van der Hoek, L. *Diversity and long-term dynamics of human blood anelloviruses*.
EClinicalMedicine, 2021	Kullberg, R., Hugenholtz, F., Brands, X., **Kinsella, C.M.,** Peters-Sengers, H., Butler, J.M., Deijs, M., Klein, M., Faber, D.R., Scicluna, B.P., van der Poll, T., van der Hoek, L., Wiersinga, J., Haak, B. *Rectal bacteriome and virome signatures and clinical outcomes in community-acquired pneumonia: An exploratory study*.
mBio, 2021	Piantadosi, P., Mukerji, S.S., Ye, S., Leone, M.J., Freimark, L., Park, D.J., Adams, G., Lemieux, J., Kanjilal, S., Solomon, I.H., Ahmed, A.A., Goldstein, R., Ganesh, V., Ostrem, B., Cummins, K.C., Thon, J., **Kinsella, C.M.**, Rosenberg, E.S., Frosh, M.P., Goldberg, M.B., Cho, T., Sabeti, P. *Enhanced virus detection and metagenomic sequencing in patients with meningitis and encephalitis*.
mSystems, 2021	Haak, B., Argelaguet, R., **Kinsella, C.M.**, Kullberg, R., Lankelma, J., Deijs, M., Klein, M., Jebbink, M., Hugenholtz, F., Kostidis, S., Giera, M., Hakvoort, T., de Jonge, W., Schultz, M., van Gool, T., van der Poll, T., de Vos, W., van der Hoek, L., Wiersinga, J. *Integrative transkingdom analysis of the gut microbiome in antibiotic perturbation and critical illness*.
PLoS Pathogens, 2020	**Kinsella, C.M.**, Santos, P.D., Postigo-Hidalgo, I., Folgueiras-González, A., Passchier, T.C., Szillat, K.P., Akello, J.O., Álvarez-Rodríguez, B., Martí-Carreras, J. *Preparedness needs research: how fundamental science and international collaboration accelerated the response to COVID-19*.
Nature Medicine, 2020	Edridge, A.W.D., Kaczorowska, J., Hoste, A., Bakker, M., Klein, M., Loens, K., Jebbink, M., Matser, A.,

	Kinsella, C.M., Rueda, P., Ieven, M., Goossens, H., Prins, M., Sastre, P., Deijs, M., van der Hoek, L. *Seasonal coronavirus protective immunity is short-lasting*.
Emerging Microbes & Infections, 2020	**Kinsella, C.M.**, Paras, M.L., Smole, S., Mehta, S., Ganesh, V., Chen, L.H., McQuillen, D. P., Shah, R., Chan, J., Osborne, M., Hennigan, S., Halpern-Smith, F., Brown, C.M., Sabeti, P., Piantadosi, A. *Jamestown Canyon virus in Massachusetts: Clinical case series and vector screening*.
Nature Communications, 2019	**Kinsella, C.M.***, Ruiz-Ruano, F.J.*, Dion-Côté, A.-M., Charles, A.J., Gossmann, T.I., Cabrero, J., Kappei, D., Hemmings, N., Simons, M.J.P., Camacho, J.P.M., Forstmeier, W., Suh, A. *Programmed DNA elimination of germline development genes in songbirds*.
Genes, 2019	Edridge, A.W.D., Deijs, M., van Zeggeren, I.E., **Kinsella, C.M.**, Jebbink, M.F., Bakker, M., van de Beek, D., Brouwer, M.C., van der Hoek, L. *Viral metagenomics on cerebrospinal fluid*.
Journal of the Geological Society, 2018	Benton, M.J., Bernardi, M., **Kinsella, C.M.** *The Carnian Pluvial Episode and the origin of dinosaurs*.
Ecology and Evolution, 2016	Stephenson, J.F., **Kinsella, C.M.**, Cable, J., van Oosterhout, C. *A further cost for the sicker sex? Evidence for male-biased parasite-induced vulnerability to predation*.

Conference reports

Viruses, 2020	Hufsky, F., Beerenwinkel, N., Meyer, I.M., Roux, S., Cook, G.M., **Kinsella, C.M.**, Lamkiewicz, K., Marquet, M., Nieuwenhuijse, D.F., Olendraite, I., Paraskevopoulou, S., Young, F., Dijkman, R., Ibrahim, B., Kelly, J., Mercier, P.L., Marz, M., Ramette, A., Thiel, V. *The International Virus Bioinformatics Meeting 2020*.

Acknowledgements

Thank you to **Dr. Lia van der Hoek**, both for leading the HONOURs programme – which was a privilege to join – and for being my advisor and promoter throughout this journey. I have learned so much in your group that I will take forward in my career. Scientifically, you have always encouraged academic freedom and creativity, and first and foremost, the projects we came up with have been fun to do. I am also deeply grateful for your support and flexibility upon the birth of our daughter Adielle. I know that you care very much about the members of your group both personally and professionally, and I am ever thankful for that.

Thank you to my copromoters, **Prof. Dr. Ben Berkhout** and **Dr. Aldert Bart**. You have always shown great enthusiasm and interest in the work of this thesis, and I have appreciated your valuable insight throughout.

Thank you to my PhD committee, **Prof. Dr. Menno de Jong**, **Prof. Dr. Colin Russel**, **Prof. Dr. Marion Koopmans**, **Dr. Mart Krupovic**, and **Dr. Jelle Matthijnssens** for your valuable time, I am very grateful for it.

Thank you to all the collaborators who have made this thesis possible, as well as those who have provided logistic support. In particular, **Aldert**, **Patricia**, and **Tom** in experimental parasitology, **Marieke** and **Karin** in medical mycology, and **Arne** and **Christin** at Charité – thank you for everything. Thank you to the **participants and organisers of the Amsterdam Cohort Studies**, and to **Margreet** for all of your help sourcing samples. Thanks to those at the **SURF Cooperative** for the compute hours, technical assistance, and infrastructure maintenance. Thanks to **Ad de Groof** for sharing data. Thanks to all scientists who share their sequence data online – your generosity has facilitated much of the work in this thesis.

Thank you to the members of the virus discovery group. **Martin** and **Maarten** – huge thanks for the numerous times you've helped in the lab, taught me methods, and answered various questions, it was great to work with you both. **Arthur**, **Joanna**, **Ferdy**, **Anne**, **and Lisa** – thanks for everything, all the discussions, collaborations, and experiences we've shared – I wish you the best of luck in your careers. **Michelle**, my gym buddy, thanks for being so fun to work with and joining me on the journey to the 100 kg squat. Good luck with your PhD! Thanks to my students **Nadine**, **Lorenzo**, **Tianne**, and **Hamed** – you taught me a great deal and it was a pleasure working with each of you.

Thanks to **all members of the HONOURs programme** – we've shared so much, and I have amazing memories of travel throughout Europe with you – hiking in the Swiss mountains with Kevin, long lab evenings running protein expressions with Tim, swapping music with Kinga, and so many more. It's been amazing to watch the development of so many talented people, and I'm really excited to see what you all do next.

Thanks to **all members of K3** past and present for being so welcoming. Many of you helped me directly during my PhD – for which I'm grateful – or joined on forays to Café Gollem – for which I'm even more grateful.

My paranymphs, **Joe** and **Mathieu**. You've belayed, you've quaffed, you've commiserated. What more can I say? Thanks for being stalwart companions throughout my PhD and beyond.

Joe and **Cosimo** – why are we not signed yet? Perhaps a new drummer is needed.

Mathieu, **Nina**, **Thoma**, and **Maya** – so many great times at your house, meals and conversations that I miss.

To all other **friends in the Netherlands and abroad**, my sincere thanks for being there. Special thanks to **Marit Melssen** for Dutch translation!

Huge thanks to my parents **Michael** and **Deirdre**, my siblings **Thérèse** and **Joe**, their partners **Pete** and **Ellie**, and my niece **Genevieve**. Thanks also to my Estonian family, **Tiina**, **Agmo**, **Karo**, **Einar**, **Saskia**, **Andre**, and **Elina**. The support you've all given me along the way has been inestimable.

I am deeply thankful for the life of **Judy Kinsella** (1922-2022), who encouraged me from the beginning of my academic journey, radiating equal enthusiasm for the artistic and the scientific – a lesson I haven't forgotten.

Kristel, ma olen sulle igavesti võlgu. Sa oled mind kõiges toetanud, ja ma armastan sind. Thank you, **Adielle Iiris Kinsella,** for keeping my sleep schedule in firm check, and for always reminding your daddy what matters in life.