Explorations of the plant virosphere

Jonathon C.O. Mifsud

Department of Biological Sciences

Macquarie University, New South Wales, 2109

11 November 2020

Submitted as part of the requirements for completion of the degree of Master of Research

Declaration

This work is presented as a ‘thesis by publication’. Chapter II is written as a manuscript for submission to *eLife* and follows the journal’s guidelines set out for publication with some exceptions: line numbers within the manuscript have been removed to conform with the rest of the thesis, figures have been inserted within the text at appropriate positions to allow ease of reading and comprehension and the conclusions section is lengthier than a submission to *eLife* may require.

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Jonathon C.O. Mifsud

11/11/2020

Chapter II of this thesis comprises research that is intended as a manuscript for publication and as such was a collaboration with other researchers, which I led. Collaborators for this work included: Rachael V. Gallagher, Macquarie University and Jemma L. Geoghegan, University of Otago. JCOM, JLG and RVG conceived the ideas; JCOM, JLG and RVG develop the methodology; JCOM conducted the formal analysis; JCOM led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Table of Contents

Acknowledgements

A very special thanks to my supervisors, Rachael Gallagher, and Jemma Geoghegan, for their advice, support, and guidance throughout this project. I would also like to thank Richard Miller for computational support and Andrew Allen for statistical advice. Thanks also go to my fellow colleagues in the Virus Evolution, and the Plant Conservation and Ecology lab. This work would not of been possible without the data generously provided by the One Thousand Plant Transcriptomes Initiative.

This project was funded by the Linnean Society of New South Wales through the William Macleay Scientific Research Fund.

Summary

Plant viruses are near-ubiquitous across natural and managed ecosystems and are known to cause significant economic damage, influence host phenotypes and modulate host-insect and microbial interactions. Metagenomic surveys of plants have recently revealed the enormous diversity of viruses they carry. Yet, this new knowledge has predominantly come from cultivated species - a small and biased subset of the plant kingdom. Next-generation sequencing technology has led to an explosion in open-source transcriptomic data which is an untapped resource for virus discovery. Here, I surveyed the transcriptomes of 960 plant species to quantify the diversity and abundance of plant viruses across 422 plant families and multiple functional groupings (i.e. dispersal syndromes, fruit types, growth forms, longevity classes, and woodiness types). In total, 3,673 plant virus transcripts were found in 415 plant species across the plant kingdom. Virome composition was associated with plant growth form and phylogenetic lineage. Notably, high virus abundance is associated with plants with a climbing habit while ancient plant lineages (algae, gymnosperms) had significantly lower virus diversity compared to more recently evolved groups, like the basal eudicots. I identified 29 potentially novel viruses including the discovery of several single-stranded RNA virus families (i.e. *Benyviridae*, *Tymoviridae* and *Secoviridae*) for the first time in lower plants or algae highlighting that non-cultivated plants likely harbour a multitude of viruses, of which the vast majority are undescribed. With the knowledge generated herein, we can begin to resolve long-held questions about the origins and diversification of plant viruses.

Chapter I: Introduction

Despite viruses being the simplest biological entity, they are an extremely diverse group. Since viruses were first isolated from diseased tobacco plants in the late 1880's virologists have traditionally focused on viruses that cause disease in humans, animals, and crops (1, 2). While these viruses are of great importance to economies and public health, the hosts from which they have been isolated represent a small subset of cellular organisms on Earth. In fact, over 80% of virus genomes catalogued in the NCBI taxonomy database belong to five virus genera, all of which are known to be pathogenic to humans (3). This anthropogenic bias combined with a dependence on culture and polymerase chain reaction (PCR) based methods of virus discovery has resulted in a restricted view of the virosphere (4, 5).

With advancements in next-generation sequencing technology, PCR based methods such as 'consensus PCR' which were once considered the gold standard for virus detection are now often relegated to confirming metagenomic results. For the purpose of virus discovery, these methods suffer from several limitations, namely, the inability to detect non-culturable viruses or in the case of consensus PCR, those that share little or no sequence similarity to known viruses (4).

The limitations of traditional virus discovery techniques are being overcome with the application of metagenomics – more specifically metatranscriptomics (bulk RNA sequencing) (Figure 1). Using viral metagenomics (hereafter viromics), it is now possible to rapidly discover novel and highly divergent viruses (6). This process involves high-throughput sequencing of total nucleic acids (DNA or RNA) isolated from a given sample (e.g. animal or plant tissue, or environmental samples such as soil or water). Before sequencing, ribosomal RNA (rRNA) is frequently depleted to increase the signal of viral genetic material in proportion to that of the host and other microbial organisms as these are typically more abundant than that of virus RNA (7). Following high-throughput sequencing (HTS), raw reads are assembled *de novo* and used in similarity searches against nucleotide and non-redundant protein databases (see Nooij, Schmitz (8) for a detailed review on these methods).

Viromics has fundamentally changed our understanding of virus diversity and their evolutionary patterns (6). Since the adoption of viromics, we have seen the focus of virus discovery expand from examining human-centric virus-host systems to include under-sampled host lineages including invertebrates (9) and chordates (10, 11). Studies of diverse hosts have discovered a multitude of new virus species, genera, and families (6, 9, 10). Such discoveries have made it apparent that we have only sampled a minuscule proportion of the virosphere and have highlighted the importance of continuing to survey under-sampled host lineages including archaea and areas of the plant kingdom (e.g. alga and gymnosperms) (12-15).

Characterising the phytovirosphere in a viromics era

The phytovirosphere encompasses the total assemblage of viruses across the plant kingdom including not only those that infect land plants but also algae. Viruses are important plant pathogens responsible for almost 50% of all emerging plant disease (16). Outbreaks of viruses (e.g. begomoviruses) cause extensive economic damage and may result in food shortages, particularly in developing countries where outbreaks are often uncontrolled (17). As such, the focus of plant virology has traditionally been on pathogenic viruses in species of economic importance (18). Investigations of the viruses that inhabited non-cultivated species were largely focused on detecting known pathogenic viruses in weeds inhabiting the agro-ecological interface (19).

With the advent of metagenomic next-generation sequencing (mNGS), we can now characterise viruses in the context of entire plant communities and ecosystems (20). Surveys of non-cultivated species have revealed the enormous diversity of viruses that infect plants (Figure 1) (21, 22). In particular, the viruses that infect cultivated species appear as only a subset of the virus diversity now known to infect non-cultivated plants (21, 23). These surveys have also revealed an abundance of vertically transmitted viruses (e.g. partitivirids and endornavirids) which are extremely widespread in plants but have been largely overlooked in a pre-viromics era as they caused symptomless and persistent infection (24-27). Viromics has also uncovered the common occurrence of mixed infections in both cultivated and non-cultivated plants (28, 29).

Metagenomic surveys have also highlighted that much of the phytovirosphere remains unclassified. At least 30% of sequences or single reads obtained from viromic surveys of plants have no detectable homology in GenBank (21, 22). Interestingly, no significant difference was found in studies examining the relatedness of cultivated and non-cultivated virus-like sequences to known viruses (21). This suggests that much of the phytovirosphere remains undiscovered even for well-researched hosts.

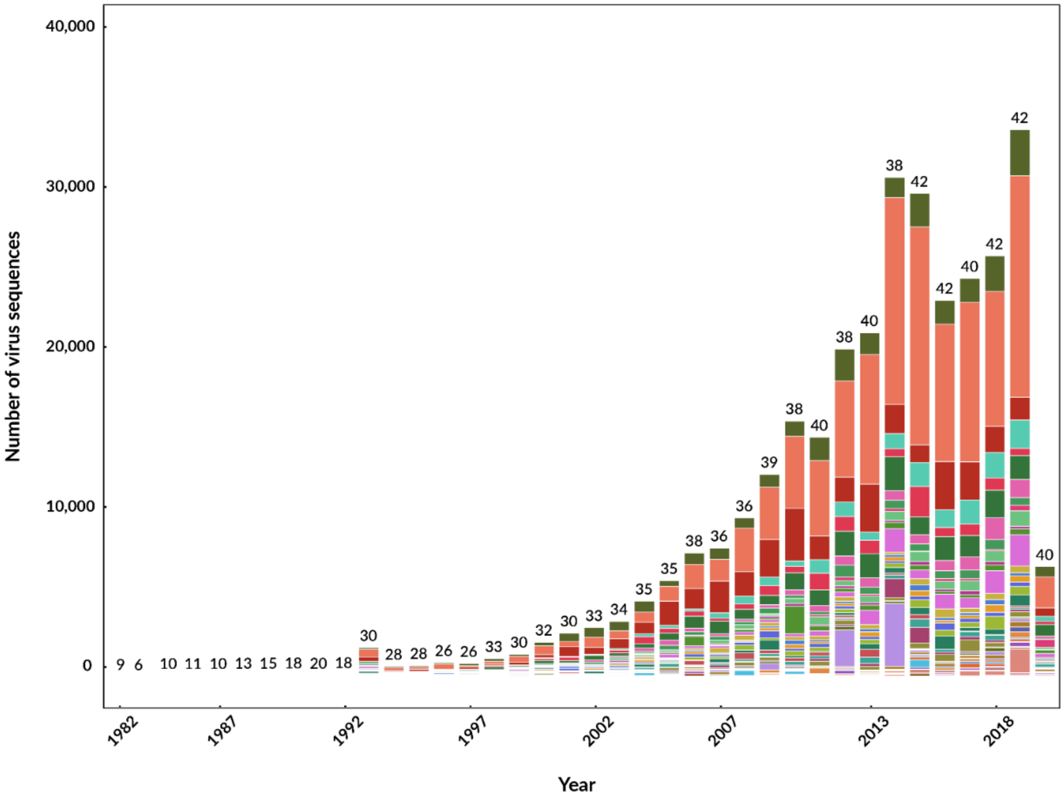


Figure 1. Expansion of the phytovirosphere in the metagenomic era. The number of phytoviruses submitted to GenBank or International Nucleotide Sequence Database Collaboration (INSDC) databases by year of submission (1982-2020). Each color/bar stack represents a virus family (n = 42, the family *Retroviridae* was excluded). Counts above each bar represent the number of distinct virus families that were submitted for a given year. Currently, 168 virus families are recognised by the International Committee on Taxonomy of Viruses (ICTV) (30). A data frame of all nucleotide virus sequences and their release dates were obtained from NCBI Virus (31) which catalogues data from Refseqs, all complete and partial NCBI viral sequences as well as proviral sequences in Genbank. The Virus-Host database (32) was used to find all virus families which contain viruses known to infect plants and used filter the NCBI viral sequences.

Despite these recent discoveries, the phytovirosphere is predominantly based on a small and biased subset of the plant kingdom – cultivated species. Indeed, 69% of all full-length virus genomes in angiosperms have been isolated from cultivated plant species despite these species representing ~0.17% of known plant diversity (see SI Document 1 for a commentary piece submitted alongside this thesis which discusses this topic in more depth). Land plants compose ~80% of the world's biomass with 49% of the world's habitable land composed of forests and shrublands (33, 34). Yet the viruses that infect these plants are rarely studied despite the likelihood that they shape demographic and ecological dynamics. Perennial species - especially long-lived trees - are largely underrepresented despite dominating natural ecosystem (35). Furthermore, little, if anything, is known about the viruses in lower plants and algae (15, 27, 36).

Filling knowledge gaps in the phytovirosphere

In the decades to come and with the continual advancements of mNGS, a representative sampling of the entire phytovirosphere is an ambitious but achievable goal. As we move towards this goal an expanding view of the phytovirosphere can provide a foundation for examining the ecological role of viruses and how ecosystems shape virus evolution. Such advancements have already revolutionised virus taxonomy and shown, with appropriate quality control, viruses can be classified from sequence data alone (36). However, with the continual growth of plant virus ecology there is a definite need for in silico analyses to be complemented by biological characterizations to fully understand the plant virosphere. Below I discuss the areas where this research is likely to be most beneficial.

Determinants of plant virus diversity and emergence How environmental and host ecological factors shape the diversity and abundance of viruses is of great importance to understanding the evolution of their diversification and emergence. We have begun to explore the impacts of ecological factors on phytovirome diversity and abundance outside of single virus-host interactions (37). To date, studies of plant virus ecology have predominately explored how climatic (e.g. temperature, relative humidity, and rainfall) (38) or biotic (e.g. host species richness, density, and identity) factors determine infection risk (39). Where ecological factors influence infection risk, they may also determine phytovirome composition. In uncultivated systems such as natural vegetation, studies have found numerous associations between virus composition and ecological factors including host plant identity (40), latitude (41), land-use type (21) and host density (42) among others. The importance of host traits for determining virome composition has recently been considered in fish (43) and birds (44) although in plants these associations are seldom considered.

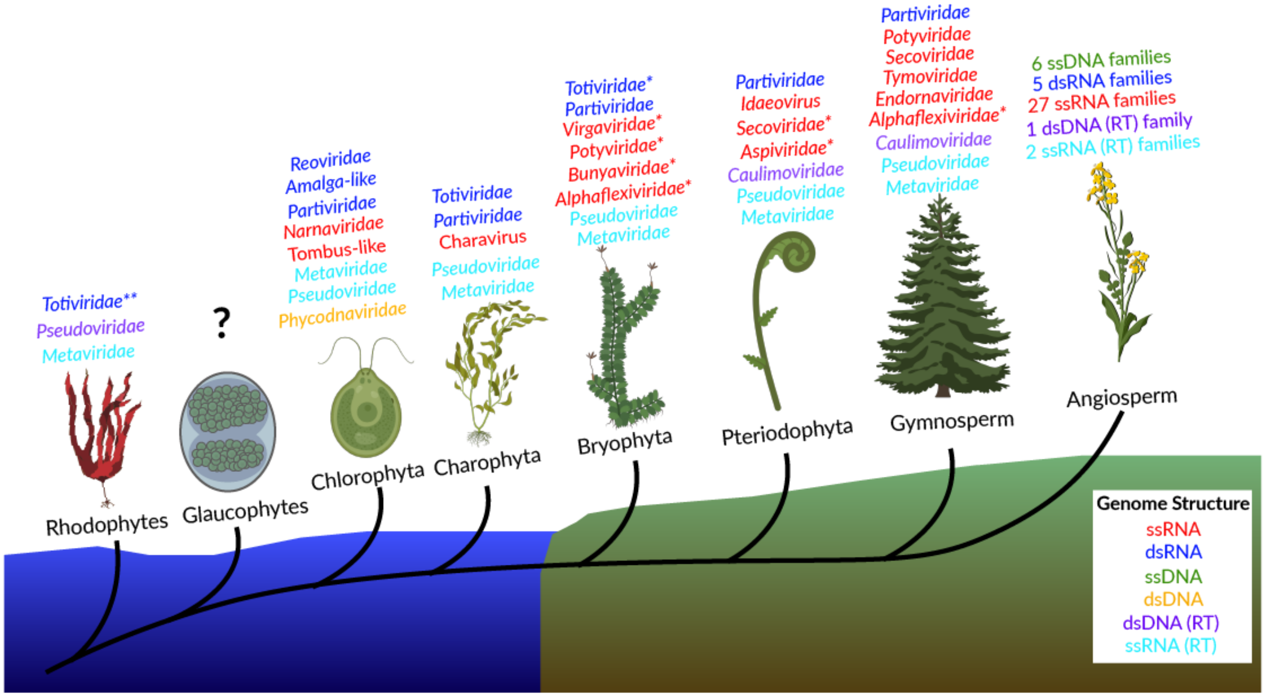
The movement of viruses from non-cultivated plants to crops is well documented (e.g. the emergence of maize streak virus to maize (*Zea mays*) in Africa, likely as a result of intra- or interspecific recombination with indigenous viruses infecting nearby non-cultivated grass (45, 46)). Global agricultural intensification will likely lead to virus spillovers between uncultivated plants and crop species due to the increased proximity between natural and cultivated ecosystems. Surveys of uncultivated plants – especially at the fault lines between agricultural and unmanaged areas – will likely be key to better understanding the processes behind phytovirus emergence.

Global change and the role of viruses Exploration of the phytovirosphere in uncultivated plants also raises important questions about the ecological role of plant viruses in natural systems. The effect of plant viruses on their host is not always one of pathogenicity; viral infection in plants may often occur without adverse signs or symptoms. This is especially true of viruses that infect plants in non-cultivated ecosystems (47), although this should not imply that virus infection in wild plant species is harmless. Viruses rely on the use of host intracellular machinery for genome replication and viral gene expression. In response, plants employ multiple defence mechanisms against viral replication and movement. Together this may suggest that avoiding infection is overall evolutionarily beneficial for the host (48, 49). It has been suggested that viruses are viewed as symbionts where their effect on a host is not static but fluctuates on a scale between pathogenic and mutualistic (50). For example, the integration of endogenous pararetroviral sequences in plant genomes can have mixed effects on host fitness. Under abiotic or genomic stress endogenized banana streak virus, petunia vein clearing virus and tobacco vein clearing virus sequences can be reactivated leading to the assembly of viral proteins and display of viral symptoms (51-54). In different environmental or genomic contexts, pararetroviral sequences may provide immunity against infection from other viruses through the generation of small interfering RNAs (55). The activation of viruses under conditions of abiotic or biotic stress may play an important role in setting or maintaining plant species range boundaries which are commonly thought to exhibit more physiologically stressful conditions for species survival – an idea dating back to Darwin (56).

Viruses are a major component of the biodiversity within ecosystems and thus, are likely indispensable members of natural systems which exert selective pressure on populations. This is most apparent in aquatic ecosystems where viral lysis of microbial hosts is key to the regulation of community composition and nutrient cycling (57-59). In terrestrial ecosystems, viruses - including those that infect plants - may serve a similar function and be important for shaping responses to global changes in climate, soil nutrients and species invasions. The emergence of viral pathogens in unmanaged ecosystems may maintain overall genetic richness by preventing the monopolization of genetically homogenous plants potentially increasing the ability of the ecosystem to adapt to future environmental change (60). Viruses may also influence whether invasive species successfully establish in new ecosystems. The enemy release hypothesis proposes that upon expansion into new ecosystems, plants are potentially liberated from virulent pathogens including viruses allowing introduced plants to reallocate resources towards growth and development (61-63). Invasive plants may increase the virus incidence and amplify vector numbers in native populations and thus increasing the likelihood of a successful invasion (64). On the contrary, the novel viruses that the invasive host acquires in its new environment may cause high host mortality resulting in failed or reversed invasions (65). Together these findings suggest viruses may hold an important role in the resilience and health of terrestrial ecosystems.

Evolutionary history of plant virus lineages The virome of land plants contains viruses belonging to two realms, *Riboviria* (RNA and reverse transcribing viruses) and *Monodnaviria* (single-stranded DNA viruses). In particular, positive-sense single-stranded RNA ((+)ssRNA) dominate virus genomic diversity in angiosperms, however only a very small subset of this virus diversity has been found in gymnosperms and lower plants (15, 66-69) (Figure 2). Unlike land plants, chlorophytes (green algae) are commonly infected by double-stranded DNA (dsDNA viruses) from the family *Phycodnaviridae* (Figure 2) (27, 36).

The current consensus of plant evolutionary history consists of the emergence of the glaucophytes, rhodophytes (red algae) and the split of the green lineage (Chloroplastida) into chlorophytes (green algae) and streptophytes approximately 1 billion years ago (70). Land plants (embryophytes) evolved from within the streptophytes approximately 400 million years ago (70). The bryophytes (hornworts, liverworts, and mosses) are the earliest branch of the land plants followed by the lycophytes, ferns, gymnosperms, and angiosperms (Figure 2) (71).



Created with BioRender.com

Figure 2. The phytovirosphere across divergent lineages of plants and algae. A schematic tree of the evolution of major plant groups. The virus families known to infect each plant group are shown above each branch and colored by their genome structure. Lineage branches are not drawn to scale. The host range of each virus family was obtained using the Virus-Host database (32) and literature searches. To our knowledge, no viruses have been found in the Glaucophytes. *RdRp fragments related to these virus families were recovered from plant transcriptomes but were too short to provide clear phylogenetic assignment (15). **The totiviruses detected in the rhodophytes were identified from a red macroalgal holobiont as such the host association of these viruses is uncertain (67, 72). The data generated within in thesis expands the known virus family host range for the Bryophyta, Pteridophyta, and brown alga (Phaeophyta, not pictured here).

The variation seen in virus composition across the major plant groups raises the question of how has the composition of the phytovirosphere transitioned across the evolutionary history of green plants and the establishment of plants on land? Terrestrialisation has likely necessitated numerous morphological and physiological changes which in turn may shape virus composition (27). For example, the terrestrialisation of plants likely resulted in viruses losing the benefits of marine environments such as the protection from UV rays and routes of virus dissemination and transmission. The anatomy of land plants also serves additional challenges for virus transmission. The ability to generate a hydrophobic surface layer or cuticle is ubiquitous among all extant embryophytes and presents an impenetrable barrier for viruses (73). Successful virus infection and plant-plant transmission require viruses to move between plant cells via the plasmodesmata. It is thought that the plasmodesmata in land plants evolved in parallel with green algal relatives, and that the last common ancestor of land plants and the green algae clade Zygnematales lacked a plasmodesmata (74). The narrow channels of the plasmodesmata may act as a barrier for large virions or dsDNA – a potential explanation for the absence of dsDNA viruses in land plants (27).

Another major force shaping the evolution of the plant virome is cross-kingdom transmission of viruses between invertebrates, fungi and plants enabled by their tight ecological associations. Indeed, viromics studies have revealed that plant virus groups hold ancient relationships with those that infect invertebrates, animals, and fungi (75). In general, plant viruses share replication and morphological structures with eukaryotic viruses - particularly those infecting arthropods and fungi (59). Such findings have led to the theory that much of the land plant virosphere was obtained by cross-species transmission between plants and invertebrates, fungi, and protists rather than through co-divergence from their algal ancestors. This conclusion is supported by the finding that species richness, a key determinant of virome diversity is ~100 times lower in the Zygnematophyceae — a sister clade to land plants — compared to vascular plants (7, 76, 77). Additionally, both vascular plant and algae richness is dwarfed by the speciose terrestrial arthropods (78). Indeed, across several virus phyla, plant viruses sit within a wider branch of arthropod and arthropod/vertebrate viruses (9, 79, 80).

It is important to note that our ability to phylogenomically reconstruct the pathway of plant virus lineage evolution is dependent on an adequate sampling of the breadth of the plant and algae kingdom. We know little if anything about the viruses that infect the green plant lineages that emerged between the chlorophytes and flowering plants (32).

Transcriptome mining as a method of virus discovery

While metagenomics has led to countless advances in the field of plant virology it has also raised many outstanding questions (as discussed above) (23). At the basis of answering many of these questions is a need to expand our view of the phytovirosphere to include a diverse range of non-cultivated hosts and environments. Viromics protocols are often straightforward and do not require a great degree of wet-lab expertise. However, conducting a broad virus survey across many plant species presents additional challenges. The abundance of primary and secondary metabolites which may vary within and between species can interfere with RNA isolation (81). In a recent study of 695 plant species, no single protocol was optimal for the isolation of total RNA (82). Instead, 18 distinct protocols using a wide variety of commercially available kits and non-commercial lab protocols were used. As such, RNA isolation from many plant species is likely time consuming and expensive.

With the development of low-cost, low-error and high-throughput sequencing technology, the amount of available transcriptomic data has increased rapidly (Figure 3) (83). In the new genomic era, the secondary analysis of transcriptomic data - known as transcriptome mining - has become a cheap and efficient method of virus discovery (84). The methodology of mining transcriptomic data mirrors that of the traditional viromics methodology minus that of DNA/RNA extraction. There is variation in the methodology used between studies but generally, these approaches are classified on whether they use reads or assembled contigs in similarity searches. The choice of approach is largely dependent on the aim of your study as discussed further in Figure 4.

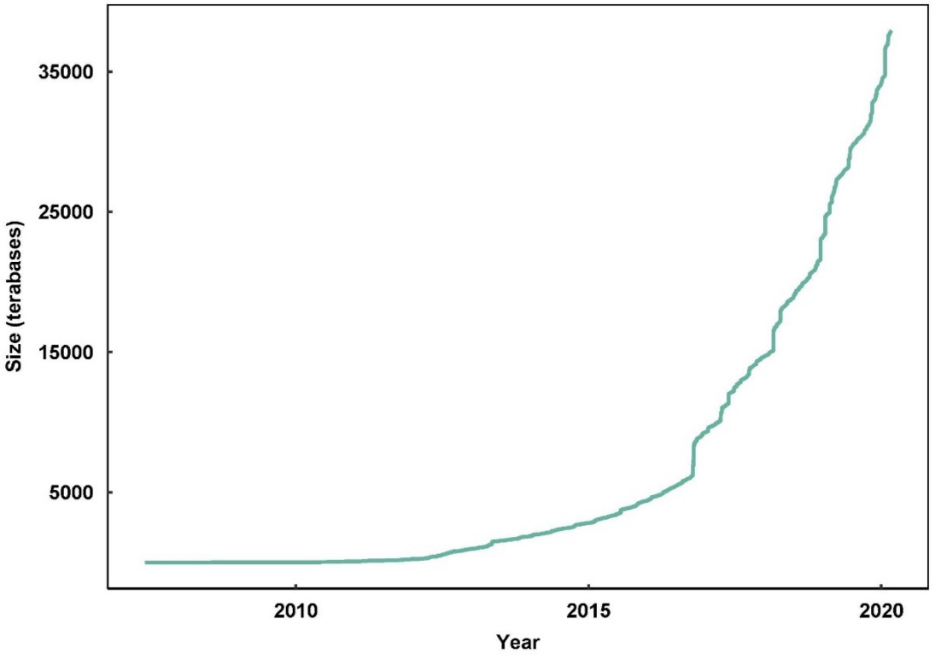


Figure 3. Rapid growth of the NCBI Sequence Read Archive (SRA) data. The number of terabases (10^{12} base pairs) of genetic information stored in the SRA by year of submission from the year 2007 until 2020. Data obtained from SRA overview page (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=announcement)

Transcriptome mining has resulted in several substantial virus discoveries in vertebrates (85), fungi (86), and plants (15, 87), including the discovery of a highly divergent lineage of plant viruses named plastrovirus which are the first plant viruses with astrovirus-like genome architecture. Plastroviruses appear to be a potential intermediate in the evolutionary transition between viruses with astro-like features and poty-like features (87). This discovery was a result of the screening of 6600 plant transcriptome projects against a database of vertebrate astroviruses.

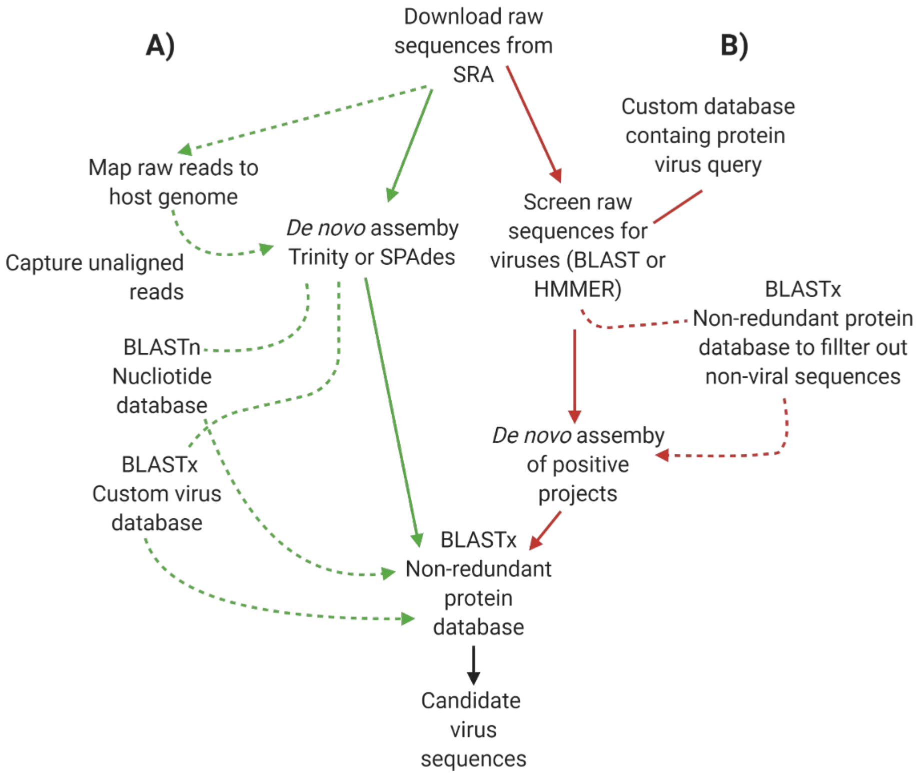

A)
Download raw sequences from SRA
B)
Custom database containg protein virus query
Map raw reads to host genome
Screen raw sequences for viruses (BLAST or HMMER)
De novo assemby Trinity or SPAdes
Capture unaligned reads
BLASTx Non-redundant protein database to fillter out non-viral sequences
BLASTn Nucliotide database
De novo assemby of positive projects
BLASTx Custom virus database
BLASTx Non-redundant protein database
Candidate virus sequences

Figure 4. Approaches to mining for viruses in the Sequence Read Archive (SRA). Green arrows indicate an assembly-based mining protocol while red arrows indicate an assembly-free protocol. Dotted lines indicate optional steps. A) The assembly-based approach is suited for describing the virome of a sample as similarity searches with large databases such as the NCBI non-redundant protein database (nr) contain many distinct virus families as well as non-viral sequences which make excluding non-viral hits more straightforward. The need for assembly combined with the use of large reference databases makes this protocol more computationally expensive. B) The assembly-free protocol involves screening raw reads for the presence of a virus using conserved sequences (e.g. RNA-dependent RNA polymerase (RdRp)). This approach is suited for the targeted discovery of a group of viruses, or to screen libraries where the presence of viruses is thought to be infrequent. Where positive hits are found, raw reads commonly undergo assembly and similarities searching in a similar manner to A). Overall, this approach can decrease the computational power required by reducing the number of libraries that undergo assembly. However, this protocol is potentially more susceptible to false positives (88). Figure created with BioRender.com.

Currently, several ‘1K’ transcriptome projects have been completed or are underway including the One Thousand Plant Transcriptomes Initiative (1KP) (89), 1K Insect Transcriptome Evolution (www.1kite.cngb.org) and the Transcriptomes of 1,000 Fishes (FISHT1K) (www.fisht1k.org). In particular, the 1KP (which contains all the transcriptomes analysed in this thesis) has sequenced the transcriptomes for over 1000 plant species across the breadth of the plant kingdom, so to provide a representative sample at the species-level for the majority of plant families (i.e. at least one species sequenced per plant family, where feasible). Projects of this nature provide an opportunity to examine virus diversity and abundance across the breadth of the host phylogeny.

Conclusion

Metagenomics has truly revolutionised plant virology revealing that non-cultivated plants harbour a diverse virome with an abundance of symptomless and vertically transmitted viruses (27, 59). Such discoveries have provided numerous insights into the evolution and diversification of plant viruses while raising many questions about their emergence and role in terrestrial ecosystems. However, this new knowledge has come from a small and biased subset of the plant kingdom – cultivated species.

Using the transcriptomes available from the 1KP I will conduct a comprehensive survey of plant viruses across the plant kingdom including numerous unsampled gymnosperms and lower plants. The aims of my thesis are, therefore:

1. Determine whether plant virome composition was conserved across the evolutionary history of plant clades;

2. Determine whether patterns of virus abundance and diversity evident in these transcriptomes were associated with plant functional traits;

3. Assess the identity and number of existing plant virus families that can be detected within the broad plant and algal lineages represented in the 1KP transcriptomes.

The viruses discovered in this thesis will fill, in part, knowledge gaps across the plant kingdom and reveal insights into the origins and diversification of plant viruses.

References

1. Beijerinck M. Concerning a contagium viwm fluidum as cause of the spot disease of tobacco leaves. Phytopathology Classics. 1898;7(1):33-52.

2. Hull R. Plant virology: Academic press; 2013.

3. Kitson E. The development and application of new computational tools for working with viral metagenomic data: University of British Columbia; 2019.

4. Kumar A, Murthy S, Kapoor A. Evolution of selective-sequencing approaches for virus discovery and virome analysis. Virus Res. 2017;239:172-9.

5. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harbor Symp Quant Biol; 1986: Cold Spring Harbor Laboratory Press.

6. Zhang YZ, Chen YM, Wang W, Qin XC, Holmes EC. Expanding the RNA virosphere by unbiased metagenomics. Annu Rev Virol. 2019;6:119-39.

7. Firth C, Lipkin WI. The genomics of emerging pathogens. Annu Rev Genomics Hum Genet. 2013;14(1):281-300.

8. Nooij S, Schmitz D, Vennema H, Kroneman A, Koopmans MP. Overview of virus metagenomic classification methods and their biological applications. Front Microbiol. 2018;9:749.

9. Shi M, Lin X-D, Tian J-H, Chen L-J, Chen X, Li C-X, et al. Redefining the invertebrate RNA virosphere. Nature. 2016;540(7634):539.

10. Shi M, Lin X-D, Chen X, Tian J-H, Chen L-J, Li K, et al. The evolutionary history of vertebrate RNA viruses. Nature. 2018;556(7700):197-202.

11. Geoghegan JL, Pirotta V, Harvey E, Smith A, Buchmann JP, Ostrowski M, et al. Virological sampling of inaccessible wildlife with drones. Viruses. 2018;10(6).

12. Geoghegan JL, Holmes EC. Predicting virus emergence amid evolutionary noise. Open Biol. 2017;7(10):170189.

13. Krupovic M, Cvirkaite-Krupovic V, Iranzo J, Prangishvili D, Koonin EV. Viruses of archaea: structural, functional, environmental and evolutionary genomics. Virus Res. 2018;244:181-93.

14. Zhang YZ, Shi M, Holmes EC. Using metagenomics to characterize an expanding virosphere. Cell. 2018;172(6):1168-72.

15. Mushegian A, Shipunov A, Elena SF. Changes in the composition of the RNA virome mark evolutionary transitions in green plants. BMC Biol. 2016;14(1):68.

16. Anderson JT. Plant fitness in a rapidly changing world. New Phytol. 2016;210(1):81-7.

17. Rybicki EP. A top ten list for economically important plant viruses. Arch Virol. 2015;160(1):17-20.

18. Wren JD, Roossinck MJ, Nelson RS, Scheets K, Palmer MW, Melcher U. Plant virus biodiversity and ecology. PLoS Biol. 2006;4(3):80.

19. da Silva SJ, Castillo-Urquiza GP, Júnior BTH, Assunção IP, Lima GS, Pio-Ribeiro G, et al. High genetic variability and recombination in a begomovirus population infecting the ubiquitous weed *Cleome affinis* in northeastern Brazil. Arch Virol. 2011;156(12):2205-13.

20. Roossinck MJ, Martin DP, Roumagnac P. Plant virus metagenomics: advances in virus discovery. Phytopathology. 2015;105(6):716-27.

21. Bernardo P, Charles-Dominique T, Barakat M, Ortet P, Fernandez E, Filloux D, et al. Geometagenomics illuminates the impact of agriculture on the distribution and prevalence of plant viruses at the ecosystem scale. ISME J. 2018;12(1):173-84.

22. Roossinck MJ, Saha P, Wiley GB, Quan J, White JD, Lai H, et al. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol Ecol. 2010;19:81-8.

23. Stobbe AH, Roossinck MJ. Plant virus metagenomics: what we know and why we need to know more. Front Plant Sci. 2014;5:150.

24. Roossinck MJ, Sabanadzovic S, Okada R, Valverde RA. The remarkable evolutionary history of endornaviruses. J Gen Virol. 2011;92(11):2674-8.

25. Fukuhara T. Endornaviruses: persistent dsRNA viruses with symbiotic properties in diverse eukaryotes. Virus Genes. 2019;55(2):165-73.

26. Sabanadzovic S, Valverde RA, Brown JK, Martin RR, Tzanetakis IE. Southern tomato virus: The link between the families *Totiviridae* and *Partitiviridae*. Virus Res. 2009;140(1-2):130-7.

27. Dolja VV, Krupovic M, Koonin EV. Deep roots and splendid boughs of the global plant virome. Annu Rev Phytopathol. 2020;58:23-53.

28. Mascia T, Gallitelli D. Synergies and antagonisms in virus interactions. Plant Sci. 2016;252:176-92.

29. Moreno Goncalves AB, Lopez-Moya JJ. When viruses play team sports: mixed infections in plants. Phytopathology. 2019;110(1):29-48.

30. Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Dempsey DM, Dutilh BE, et al. Changes to virus taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2019). Arch Virol. 2019;164(9):2417-29.

31. Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP, Ostapchuck Y, et al. Virus Variation Resource–improved response to emergent viral outbreaks. Nucleic Acids Res. 2017;45(D1):D482-D90.

32. Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, et al. Linking virus genomes with host taxonomy. Viruses. 2016;8(3):66.

33. Bar-On YM, Phillips R, Milo R. The biomass distribution on Earth. Proc Natl Acad Sci U S A. 2018;115(25):6506-11.

34. Ellis EC, Klein Goldewijk K, Siebert S, Lightman D, Ramankutty N. Anthropogenic transformation of the biomes, 1700 to 2000. Glob Ecol Biogeogr. 2010;19(5):589-606.

35. Mifsud JC, Geoghegan JL, Gallagher RV. Examining the diversity of the phytovirosphere. Funct Ecol. (Manuscript under review).

36. Coy SR, Gann ER, Pound HL, Short SM, Wilhelm SW. Viruses of eukaryotic algae: diversity, methods for detection, and future directions. Viruses. 2018;10(9):487.

37. Simmonds P, Adams MJ, Benko M, Breitbart M, Brister JR, Carstens EB, et al. Virus taxonomy in the age of metagenomics. Nat Rev Microbiol. 2017;15(3):161-8.

38. Maclot F, Candresse T, Filloux D, Malmstrom CM, Roumagnac P, van der Vlugt R, et al. Illuminating an ecological blackbox: using high throughput Sequencing to characterize the plant virome across scales. Front Microbiol. 2020;11:2575.

39. Garrett KA, Dendy SP, Frank EE, Rouse MN, Travers SE. Climate change effects on plant disease: genomes to ecosystems. Annu Rev Phytopathol. 2006;44:489-509.

40. Islam W, Zhang J, Adnan M, Noman A, Zainab M, Jian W. Plant virus ecology: a glimpse of recent accomplishments. Appl Ecol Environ Res. 2017;15(1):691-705.

41. Thapa V, McGlinn DJ, Melcher U, Palmer MW, Roossinck MJ. Determinants of taxonomic composition of plant viruses at the Nature Conservancy's Tallgrass Prairie Preserve, Oklahoma. Virus Evol. 2015;1(1).

42. Seabloom EW, Borer ET, Mitchell CE, Power AG. Viral diversity and prevalence gradients in North American Pacific Coast grasslands. Ecology. 2010;91(3):721-32.

43. Rodríguez-Nevado C, Gavilán R, Pagán I. Host abundance and identity determine the epidemiology and evolution of a generalist plant virus in a wild ecosystem. Phytopathology. 2020;0(1):94-105.

44. Geoghegan JL, Giallonardo FD, Wille M, Ortiz-Baez AS, Costa VA, Ghaly T, et al. Host evolutionary history and ecology shape virome composition in fishes. bioRxiv.

45. Wille M, Eden JS, Shi M, Klaassen M, Hurt AC, Holmes EC. Virus-virus interactions and host ecology are associated with RNA virome structure in wild birds. Mol Ecol. 2018;27(24):5263-78.

46. Martin DP, Willment JA, Billharz R, Velders R, Odhiambo B, Njuguna J, et al. Sequence diversity and virulence in *Zea mays* of Maize streak virus isolates. Virology. 2001;288(2):247-55.

47. Roossinck MJ, García-Arenal F. Ecosystem simplification, biodiversity loss and plant virus emergence. Curr Opin Virol. 2015;10:56-62.

48. Prendeville HR, Ye X, Jack Morris T, Pilson D. Virus infections in wild plant populations are both frequent and often unapparent. Am J Bot. 2012;99(6):1033-42.

49. Incarbone M, Dunoyer P. RNA silencing and its suppression: novel insights from in planta analyses. Trends Plant Sci. 2013;18(7):382-92.

50. Calil IP, Fontes EP. Plant immunity against viruses: antiviral immune receptors in focus. Ann Bot. 2016;119(5):711-23.

51. Roossinck MJ. Plants, viruses and the environment: Ecology and mutualism. Virology. 2015;479:271-7.

52. Lockhart BE, Menke J, Dahal G, Olszewski N. Characterization and genomic analysis of tobacco vein clearing virus, a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. J Gen Virol. 2000;81(6):1579-85.

53. Harper G, Osuji JO, Heslop-Harrison J, Hull R. Integration of banana streak badnavirus into the *Musa* genome: molecular and cytogenetic evidence. Virology. 1999;255:207-13.

54. Ndowora T, Dahal G, LaFleur D, Harper G, Hull R, Olszewski NE, et al. Evidence that badnavirus infection in *Musa* can originate from integrated pararetroviral sequences. Virology. 1999;255:214-20.

55. Richert-Pöggeler KR, Noreen F, Schwarzacher T, Harper G, Hohn T. Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in *Petunia*. EMBO J. 2003;22(18):4836.

56. Staginnus C, Gregor W, Mette MF, Teo CH, Borroto-Fernández EG, da Câmara Machado ML, et al. Endogenous pararetroviral sequences in tomato (*Solanum lycopersicum*) and related species. BMC Plant Biol. 2007;7(1):24.

57. Darwin C. On the origin of species by means of natural selection. London: J Murray. 1859.

58. Proctor LM, Fuhrman JA. Viral mortality of marine bacteria and cyanobacteria. Nature. 1990;343(6253):60-2.

59. Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, Brussaard CPD, et al. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. Nat Commun. 2017;8(1):15955.

60. Coutinho FH, Gregoracci GB, Walter JM, Thompson CC, Thompson FL. Metagenomics sheds light on the ecology of marine microbes and their viruses. Trends Microbiol. 2018;26(11):955-65.

61. Lefeuvre P, Martin DP, Elena SF, Shepherd DN, Roumagnac P, Varsani A. Evolution and ecology of plant viruses. Nat Rev Microbiol. 2019;17(10):632-44.

62. Blossey B, Notzold R. Evolution of increased competitive ability in invasive nonindigenous plants: a hypothesis. J Ecol. 1995;83(5):887-9.

63. Williamson M. Biological invasions: Springer Science & Business Media; 1996.

64. Mitchell CE, Power AG. Release of invasive plants from fungal and viral pathogens. Nature. 2003;421(6923):625-7.

65. Malmstrom CM, McCullough AJ, Johnson HA, Newton LA, Borer ET. Invasive annual grasses indirectly increase virus incidence in California native perennial bunchgrasses. Oecologia. 2005;145(1):153-64.

66. Verhoeven KJ, Biere A, Harvey JA, Van Der Putten WH. Plant invaders and their novel natural enemies: who is naive? Ecol Lett. 2009;12(2):107-17.

67. Vlok M, Gibbs AJ, Suttle CA. Metagenomes of a freshwater charavirus from British Columbia provide a window into ancient lineages of viruses. Viruses. 2019;11(3).

68. Lachnit T, Thomas T, Steinberg P. Expanding our understanding of the seaweed holobiont: RNA viruses of the red alga *Delisea pulchra*. Front Microbiol. 2016;6:1489.

69. Koh SH, Li H, Admiraal R, Jones MGK, Wylie SJ. Catharanthus mosaic virus: a potyvirus from a gymnosperm, *Welwitschia mirabilis*. Virus Res. 2015;203:41-6.

70. Charon J, Marcelino VR, Wetherbee R, Verbruggen H, Holmes EC. Metatranscriptomic identification of diverse and divergent RNA viruses in green and Chlorarachniophyte algae cultures. Viruses. 2020;12(10).

71. de Vries J, Archibald JM. Plant evolution: landmarks on the path to terrestrial life. New Phytol. 2018;217(4):1428-34.

72. Cheng S, Xian W, Fu Y, Marin B, Keller J, Wu T, et al. Genomes of subaerial Zygnematophyceae provide insights into land plant evolution. Cell. 2019;179(5):1057-67. e14.

73. Rousvoal S, Bouyer B, López-Cristoffanini C, Boyen C, Collén J. Mutant swarms of a totivirus-like entities are present in the red macroalga *Chondrus crispus* and have been partially transferred to the nuclear genome. J Phycol. 2016;52(4):493-504.

74. Budke JM, Goffinet B, Jones CS. The cuticle on the gametophyte calyptra matures before the sporophyte cuticle in the moss *Funaria hygrometrica* (*Funariaceae*). Am J Bot. 2012;99(1):14-22.

75. Brunkard JO, Zambryski PC. Plasmodesmata enable multicellularity: new insights into their evolution, biogenesis, and functions in development and immunity. Curr Opin Plant Biol. 2017;35:76-83.

76. Koonin EV, Dolja VV, Krupovic M. Origins and evolution of viruses of eukaryotes: The ultimate modularity. Virology. 2015;479-480:2-25.

77. Willis KJ. State of the world's plants report-2017: Royal Botanic Gardens; 2017.

78. Guiry MD. How many species of algae are there? J Phycol. 2012;48(5):1057-63.

79. Stork NE. How many species of insects and other terrestrial arthropods are there on Earth? Annu Rev Entomol. 2018;63:31-45.

80. Wolf YI, Kazlauskas D, Iranzo J, Lucia-Sanz A, Kuhn JH, Krupovic M, et al. Origins and evolution of the global RNA cirome. mBio. 2018;9(6).

81. Li C-X, Shi M, Tian J-H, Lin X-D, Kang Y-J, Chen L-J, et al. Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. elife. 2015;4:e05378.

82. Bilgin DD, DeLucia EH, Clough SJ. A robust plant RNA isolation method suitable for Affymetrix GeneChip analysis and quantitative real-time RT-PCR. Nat Protoc. 2009;4(3):333.

83. Johnson MT, Carpenter EJ, Tian Z, Bruskiewich R, Burris JN, Carrigan CT, et al. Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. PLoS One. 2012;7(11):e50226.

84. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem. 2009;55(4):641-58.

85. Greninger AL. A decade of RNA virus metagenomics is (not) enough. Virus Res. 2018;244:218-29.

86. Lauber C, Seitz S, Mattei S, Suh A, Beck J, Herstein J, et al. Deciphering the origin and evolution of hepatitis B viruses by means of a family of non-enveloped fish viruses. Cell Host Microbe. 2017;22(3):387-99.e6.

87. Gilbert KB, Holcomb EE, Allscheid RL, Carrington JC. Hiding in plain sight: New virus genomes discovered via a systematic analysis of fungal public transcriptomes. PLoS One. 2019;14(7):e0219207.

88. Lauber C, Seifert M, Bartenschlager R, Seitz S. Discovery of highly divergent lineages of plant-associated astro-like viruses sheds light on the emergence of potyviruses. Virus Res. 2019;260:38-48.

89. Soueidan H, Schmitt L-A, Candresse T, Nikolski M. Finding and identifying the viral needle in the metagenomic haystack: trends and challenges. Front Microbiol. 2015;5:739.

90. Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, et al. One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 2019;574(7780):679-85.

Chapter II: Explorations of the plant virosphere

Abstract

Our knowledge of plant viruses has predominantly come from cultivated species - a small and biased subset of the plant kingdom. Next-generation sequencing technology has led to an explosion in transcriptomic data which is an untapped resource for virus discovery. Here, we surveyed the transcriptomes of 960 plant species to quantify the diversity and abundance of plant viruses across 422 plant families and multiple functional groupings. In total, 3,637 plant virus transcripts were found in 415 plant species. Virome composition was associated with plant growth form and phylogenetic lineage. More specifically, virus abundance was greater in plants with a climbing habit while ancient plant lineages (algae, gymnosperms) had lower virus diversity compared to recently evolved groups (basal eudicots). The discovery of several RNA virus families for the first time in lower plants and algae highlights that non-cultivated plants likely harbour a multitude of viruses, of which the majority are undescribed.

Introduction

Viruses are responsible for almost 50% of all emerging plant disease (1). Historically, virus identification and characterisation have focused on pathogenic viruses that infect species of economic importance with 69% of the phytovirosphere — the total assemblage of viruses across the plant kingdom — found in cultivated species despite these species representing under 0.17% of all known plant diversity (2, 3). With the advent of high throughput sequencing technology virologists can now comprehensively screen for novel and known viruses in plant species or communities (4). Despite this, the vast majority of plant lineages remain unsurveyed (3).

Our ability to infer the origins and diversification of the phytovirosphere from phylogenomic data requires adequate sampling of the viruses across the plant kingdom. Several key plant groups are currently missing or severely underrepresented in our quantification of the phytovirosphere, including gymnosperms, lower plants, algae, and some angiosperm orders (3). Improving sampling across these groups will help uncover the evolutionary history of plant virus lineages. For instance, the inheritance of virus groups from algal ancestors can be examined, as can the acquisition of viruses through cross-species transmission from plant-associated organisms such as invertebrates, fungi, or protists. Importantly, we can begin to investigate how the key evolutionary transitions of plants – such as terrestrialisation – have shaped the contemporary land plant virome (5).

The vast majority of known plant virus genera (68%) are positive-sense single-stranded RNA (+ssRNA) viruses (5). Unlike land plants, algae are commonly infected by double-stranded DNA (dsDNA) viruses particularly from the *Phycodnaviridae* (6, 7). The evolutionary pathways resulting in the compositional differences between land plants and algae are not yet known, although several theories have been suggested. Such theories include the complex anatomy, cell architecture and innate immune system of land plants to explain the absence of dsDNA viruses in land plants, while the enormous diversification of +RNA viruses is likely a result of cross-species transmission from invertebrates and fungi and, to a lesser extent, the inheritance of viruses from algal ancestors (see Dolja, Krupovic (7)).

Globally, there is enormous variation in the ecological strategies of plants and their traits (8). Considering this, key traits which capture important aspects of plant function, such as growth form and whole plant longevity, may also shape the species composition of viral communities at the host species level. To date, studies have found associations between virus composition and host plant identity (9), latitude (10), land-use type (11), and host density (12) among others, though associations between functional traits and virus composition are seldom considered.

In the metagenomic era, the secondary analysis of transcriptomic data — known as transcriptome mining — has become an inexpensive and efficient method of virus discovery that leverages previous investment (13). To this end, we 'mined' the transcriptome data generated by the One Thousand Plant Transcriptomes Initiative (1KP) using sequence homology searches of known plant viruses. The 1KP project provides a major untapped source of transcriptome data for virus discovery drawn from species across the breadth of the plant kingdom including streptophyte and chlorophyte green algae, bryophytes, ferns, gymnosperms and angiosperms. (14, 15). Unlike previous sequencing efforts, the plant species chosen by the 1KP were not biased towards the model organisms and crop species (15)

The aims of this chapter of my thesis were to:

1. Determine whether plant virome composition was conserved across the evolutionary history of plant clades;

2. Determine whether patterns of virus abundance and diversity evident in these transcriptomes were associated with plant functional traits;

3. Assess the identity and number of existing plant virus families that can be detected within the broad plant and algal lineages represented in the 1KP transcriptomes.

This study will uncover more of the phytovirosphere across the plant kingdom, revealing insights into the origins and diversification of plant viruses.

Methods

Transcriptome data generation

The 1KP generated RNA seq libraries from 1147 plant species across the breadth of the plant kingdom (14). Due to the diversity of species examined, samples were obtained from multiple sources including field collections, greenhouses, culture collections and laboratory specimens (16). For the majority of species, young leaves or shoots were collected, although occasionally a mix of vegetative and reproductive tissues was used. To avoid RNA degradation, RNA extraction was performed immediately after tissue collection or tissue was frozen in liquid nitrogen and stored in a -80˚C until extraction (16). Several extraction protocols were used including CTAB and TRIzol (see Johnson, Carpenter (16) for complete details). All sequencing was conducted at BGI-Shenzhen, China, using a combination of in-house protocols or TruSeq chemistry (16). All libraries were prepared from polyA RNA. Paired-end sequencing was initially completed using Illumina GAII machines (11% of libraries) with a ~72bp read length but later the HiSeq platform was used (89% of libraries) with a 90 bp read length (16).

Discovering viruses in the 1KP

Raw plant transcriptomes (n = 1079) from the 1KP were downloaded from the NCBI Short Read Archive (SRA) database (BioProject accession PRJEB21674) and converted to FASTQ format using the SRA Toolkit program fastq-dump in combination with the parallel-fastq-dump wrapper (17). 100 transcriptomes within the BioProject were not publicly available (released 22/08/2019) at the commencement of this study and thus, were not analysed. To reduce the downstream computing resources needed, raw sequences were mapped to their respective host genome scaffold using bowtie2 (18). Genome scaffolds were assembled as part of a previous study (14). Where genome scaffolds were not available (n = 2) all reads were assembled *de novo*. Trinity RNA-seq (v2.1.1) was used to quality trim and assemble *de novo* the unaligned reads captured from mapping (Figure 1) (19). The assembled contigs were then assigned to known virus families and annotated through similarity searches against the NCBI nucleotide database (nt), the non-redundant protein database (nr) and a custom viral RNA dependent RNA polymerases (RdRp) database using BLASTN and Diamond (BLASTX) (20, 21). To filter out weak BLAST sequence matches an e-value cut-off of 1×10^{-10} was used meaning that we would not expect to observe a sequence match by chance alone. To identify potential false positives, putative viral contigs were manually compared across the three BLAST searches (nt, nr and RdRp) to ensure matches to virus-associated sequences were consistent.

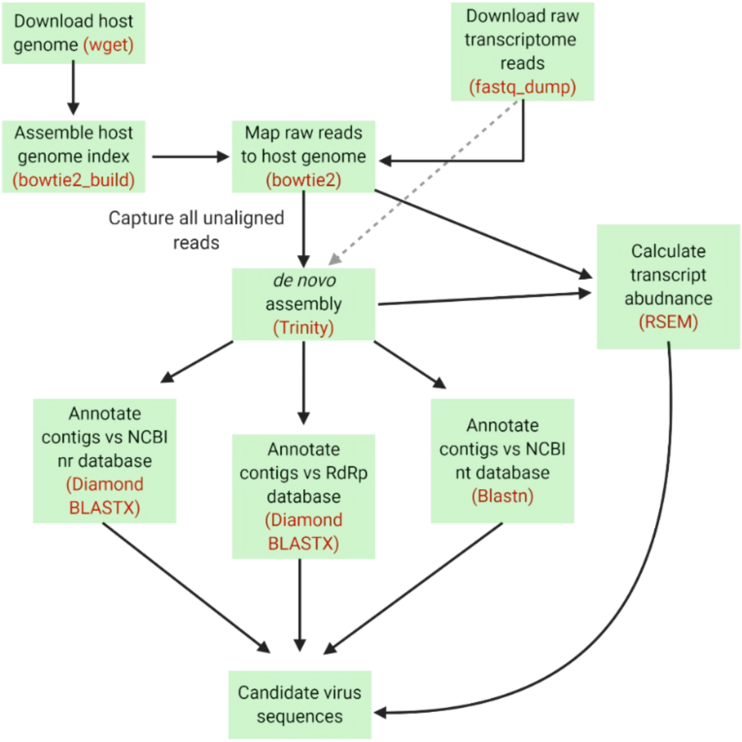

Download host genome (wget)
Download raw transcriptome reads (fastq_dump)
Assemble host genome index (bowtie2_build)
Map raw reads to host genome (bowtie2)
Capture all unaligned reads
de novo assembly (Trinity)
Calculate transcript abudnance (RSEM)
Annotate contigs vs NCBI nr database (Diamond BLASTX)
Annotate contigs vs RdRp database (Diamond BLASTX)
Annotate contigs vs NCBI nt database (Blastn)
Candidate virus sequences

Figure 1. Bioinformatic steps for the virus-like discovery pipeline. Bioinformatics programs used are highlighted in red. The dotted line represents the pipeline structure where a host genome scaffold is not available. In this case, raw reads were sent directly to Trinity for assembly. Our pipeline is based upon the assembly-based transcriptome mining protocol discussed in Chapter 1, Figure 4 (A). Figure created with BioRender.com.

Virus filtering and abundance calculations

For all diversity and abundance analyses, we focused on virus families known to infect plants or algae. As we rely on sequence-based homology searches for virus detection it is important to note that such methodologies are biased towards viruses that share homology to existing virus families. Together, the Virus-Host database (22) and the International Committee on Taxonomy of Viruses (ICTV) Virus taxonomy database were used to develop a list of plant virus families and genera to filter out virus-like contigs associated with vertebrate, invertebrate or fungi hosts based upon their top BLASTx and BLASTn matches. Packages within the Tidyverse collection (v1.3.0) in RStudio were utilised to complete these tasks (23-25). Where the host was ambiguous (e.g. belonged to a family or genera known to infect both plant and fungal species) the contig was inspected manually. Furthermore, for these contigs, we investigated whether reads belonging to other eukaryotes were present in the sequencing libraries as the 1KP samples were not axenic. To achieve this, we obtained taxonomic identification for raw reads in each library – prior to the removal of host reads – by aligning them to the

NCBI nt database using the KMA aligner and the CCMetagen program (26, 27). Library contamination was also assessed by the 1KP and used to inform our host assignment (see Carpenter et al. (15)).

The relative abundance of each transcript within the host transcriptome was calculated using RNA-Seq by Expectation-Maximization (RSEM) (v1.2.28) (28). To account for variation in the number of unaligned reads between libraries after mapping, contig abundance was standardised by the total number of unaligned paired reads – hereafter referred to as "total standardised abundance of viral transcripts". Contigs under 200 nucleotides in length were excluded from further analysis. The 1KP project includes multiple libraries for 96 plant species often with each library associated with different host tissue. To address this, results were pooled within species and contig abundance was recalculated to account for pooling.

Collation of plant trait information

We searched for data on five plant functional traits – dispersal syndrome, fruit type, growth form, longevity, and woodiness – that are hypothesised to affect the transmission, diversity, and abundance of viruses across the plant kingdom. Justifications for our choice of traits in this exploratory analysis are provided in Table 1. Trait information was sought for all plant species in the 1KP project which were found to host a plant virus (n = 415). Available plant functional trait data was collated from four databases: BIEN (29), BROT (30), TRY (31) and USDA Plants (32). Species taxonomy between each database was standardised using the Global Name Resolver implemented within the *taxize* R package (v0.9.94) (33).

All traits were coded as categorical variables (e.g. longevity = annual, perennial) and observations from different databases were standardised into categories based on the consensus trait state across sources (Table 1). For instance, where species had more than one trait state listed across databases (i.e. woody in BIEN and non-woody in BROT) the state with the most observations across all databases were assigned. In cases where there was no clear trait assignment (i.e. tied values), the species was excluded from analyses of trait-virome associations. We also examined associations between plant 'usage types' (e.g. crop, non-crop) and virome composition. Data on usage type was collected from the World Economic Plants resource in GRIN-Global, which is based on World Economic Plants: A Standard Reference (34). We defined crop-associated species broadly as those which matched the query: animal food, bee food, fuels, harmful host organism, human food, food additives, environmental, invertebrate food, medicines, non-vertebrate poisons, social, vertebrate poisons, weed and including all subclasses in GRIN-Global.

Table 1. Functional traits and species attributes used in analyses of virus abundance and diversity.

Trait/attribute	Trait state**	Justification	Source
Dispersal syndrome*	abiotic (wind, water), biotic (vertebrate, invertebrate)	Plant dispersal syndromes may influence the movement of plant-associated viruses within and between ecosystems (35).	BIEN (29), BROT (30), TRY (31)
Fruit type*	wet (e.g. berry, drupe), dry (e.g. capsule, nut)	Fruits with high and low water content provide contrasting environments for virus replication/survival, with wet fruits potentially associated with higher virus abundance and diversity (36).	BIEN (29), BROT (30) TRY(31)
Growth form	climber, herb, shrub, tree	As virus abundance and diversity may differ between tissue types, variation in biomass allocation patterns between growth forms may translate into differences in virus composition (37).	BIEN (29), BROT (30), TRY (31), USDA Plants (32)
Longevity	annual, perennial	Perennial species may accumulate a greater diversity of viruses over their life cycle, relative to annual species. Annual species may invest less in defences against viruses and in turn will be more susceptible to infection (38)	BIEN (29), BROT (30), TRY (31),USDA Plants (32)
Woodiness	woody, non-woody	Differences in water content between herbaceous (non-woody) and woody species shape the internal environment of the host and may lead to changes in the abundance and diversity of viruses (37).	BIEN (29), TRY (31)
Usage type	crop associated or non-crop associated	Crop species are typically planted in monocultures at high density relative to wild species and this may result in higher virus abundance but potentially lower virus diversity (11).	GRIN-GLOBAL (39)

*A lack of available trait data across a large number of One Thousand Plant Transcriptomes Initiative (1KP) species precluded the formal testing of these ideas in this thesis.

** See SI Table 1 for botanical definitions of each trait's states listed in each category.

Host clade

Each plant host was assigned to each clade in a previous study based upon their phylogenetic positioning and lineage information (14). To increase the number of species in each group we reduced the number of clades from 25 down to eight (core eudicots, basal eudicots, monocots, basalmost angiosperms, gymnosperms, fern and fern allies, non-vascular and lastly, algae) through combining those which were closely related or potentially overlapping. For example, green algae, red algae and Chromista were collapsed into a single category, 'algae' (SI Table 2).

Virus abundance and diversity

To examine whether plant virome composition was associated with functional traits and the phylogenetic lineage of their hosts we examined the effect of five host attributes: (1) Virus abundance (i.e. the total standardised abundance of viral transcripts in a library); (2) Alpha diversity indices (i.e. the diversity of virus families per plant species); and, (3) Beta diversity (the diversity of virus families between plant species). Host attributes included host growth form, lifespan, woodiness, usage type and clade (which we used as an approximation of phylogenetic position). Trait coverage was poor for several plant groups including algae, ferns and fern allies and non-vascular plant species, as such these groups were excluded in analyses of functional traits. Algae, ferns and fern allies and non-vascular plant species were included in the analysis of host clade. We also excluded all libraries (n = 2) where a host genome scaffold was not available. All analyses of virus abundance and diversity were conducted in R (v3.6.0). A Box-Cox transformation was used as the total standardised abundance of viral transcripts was not normally distributed. Using modified scripts from the Rhea project (40) library virome richness and alpha diversity (Shannon effective) were calculated at the virus family level using the untransformed standardised virus abundance (41). Using generalised linear models, we compared virus family diversity and abundance to host functional traits, phylogenetic relatedness (host clade), and usage type. Model significance was assessed using a Likelihood Ratio Test. Where an explanatory variable had more than two factor levels, a pairwise comparison (Tukey posthoc) was conducted using the glht function within the *multcomp* package (42) to identify divergent pairs. The differences in virus family diversity and abundance between samples (beta diversity) were investigated using a Bray Curtis dissimilarity matrix. To determine which viral families were contributing the most to differences between host clades and growth forms, an indicator species analysis was performed using the indicspecies package (v1.7.9) in R with 999 permutations (43).

Phylogenetic reconstructions of plant viruses

Virus phylogenies of the plant-associated viruses discovered here were inferred using the maximum likelihood method. We combined our translated virus contigs with known virus protein sequences from each respective virus family taken from NCBI/GenBank (20, 21). Sequences were then aligned using the E-INS-I algorithm implemented in the program MAFFT (v7.450) with default parameters (44). Sites of ambiguity were removed using trimAl (v1.2) (45). A maximum likelihood approach implemented in IQ-TREE with 1000 bootstrap replicates was used to analyse phylogenetic data (46). The LG amino acid substitution model was selected. Phylogenetic trees were annotated with FigTree (v1.4.4) (47) and further edited in Adobe Illustrator (https://www.adobe.com). To determine whether a virus was novel, we followed the criteria as specified by The International Committee on Taxonomy of Viruses (ICTV) (http://www.ictvonline.org/).

Results

We characterised the viruses found in the transcriptomes of 960 plant species within the 1KP project. The transcriptomes represented a broad taxonomic sampling across the Archaeplastida (green plants, glaucophytes and red algae). Sequencing libraries had a median of 25,187,714 reads (range 10,156,464–46,650,336). A median of 82% of reads (range 1%-96%) in these libraries mapped to host genome scaffolds and were subsequently removed. *De novo* assembly of the sequencing reads resulted in a median of 36,015 contigs (range 1,396–146,217) per library, with a total of 41,256,176 contigs generated (SI Table 3).

Diversity and abundance of plant viruses

In total, virus-like transcripts were found in 603 plant species although only 69% of these were found to be plant-associated. That is, numerous identified sequences shared high similarity to non-plant associated viruses including those known to infect fungi, invertebrate and vertebrate hosts. Among these transcripts, 34% of these were unclassified, while the remaining transcripts were largely classified within the *Orthomyxoviridae* (25%), *Rhabdoviridae* (invertebrate associated) (17%), *Partitiviridae* (fungus associated) (10%), *Mimiviridae* (10%) and *Adenoviridae* (7%). Although some of these viruses could represent plant infection it remains challenging to discern their source and thus, they were excluded from further analyses. Hereafter, the viruses discussed are those we deemed as plant-associated and these were drawn from 21 viral families known to infect plants (SI Figure 1).

We detected transcripts closely associated with viruses containing single and double-stranded DNA and RNA genomes. The majority of virus-like sequences belonged to families with ssRNA genomes (60.9%) or reverse-transcribing dsDNA viruses (21.8%). The ssRNA virus transcripts were predominately classified within the *Betaflexiviridae* (30%), *Potyviridae* (19%), *Secoviridae* (16%) and *Alphaflexiviridae* (10%) (SI Table 4). Notably, all dsDNA viruses were exclusively reverse-transcribing viruses from the *Caulimoviridae*. dsDNA virus transcripts with similarities to the *Phycodnaviridae* were detected across the algae samples. These phycodna-like virus transcripts frequently encoded the chitinase and DNA ligase genes which are homologous to those in distant organisms including fungi and bacteria. Due to the difficulties discerning whether these transcripts represent *Phycodnaviridae* sequences or contaminates, we excluded all phycodnavirus related sequences. We failed to detect any sequences which shared homology with several plant virus families including *Reoviridae*, *Nanoviridae*, *Phenuiviridae* and *Fimoviridae* (see Caveats for further discussion).

There was a large range of total viral abundance in each library (5.38% × 10^{-6}–30.57% reads). Viruses with +ssRNA genomes accounted for the vast majority (99.8%) of virus abundance detected (SI Figure 1). As expected, virus discovery was concentrated in the flowering plants (angiosperms) which have the highest number of previously identified viruses. For instance, plant virus-like sequences were frequently discovered in the core eudicots and monocots (i.e. 73% of libraries in which plant virus transcripts were found (Figure 2)). The discovery rate of plant viruses was highest in the basalmost angiosperms (57%) and monocots (50%). No significant difference in virus abundance was observed between Genome Analyzer II and Illumina HiSeq 2000 platforms (p=0.327).

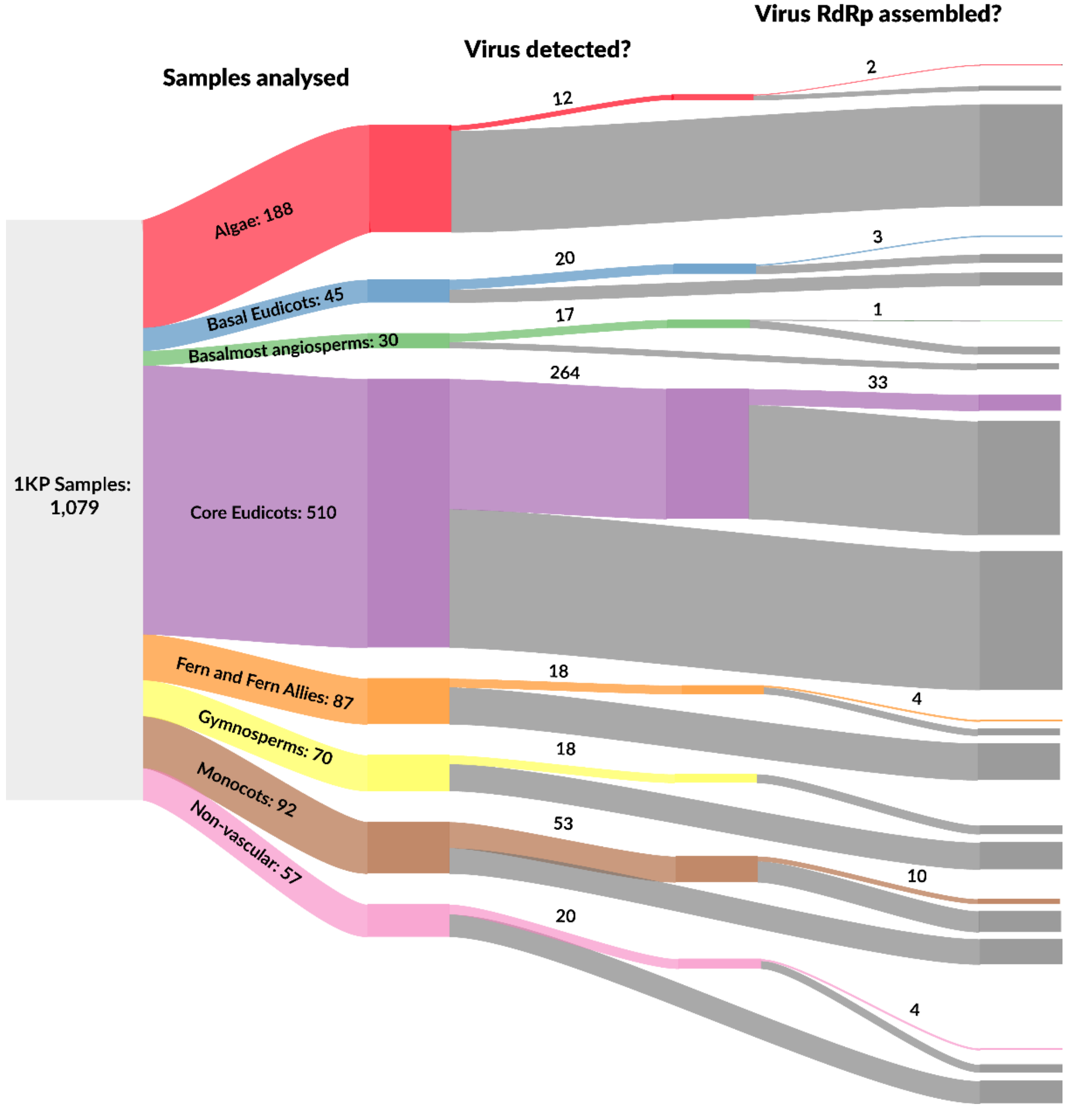


Figure 2. A Sankey diagram summarising the One Thousand Plant Transcriptomes Initiative (1KP) plant hosts queried in this study and the viruses they contain. The height of each column is indicative of the number of 1KP samples analysed. Column 1: splits into eight nodes (colour bands) based upon the sample host species taxonomic grouping. The taxonomic information used to group host species was provided by the 1KP project (14). Column 2: the number of samples in which plant virus-like transcripts were found. Grey is used to indicate samples in which no plant virus-like transcripts were found. Column 3: indicates the number of distinct virus RdRp contigs found in each taxonomic grouping. The online tool SankeyMATIC (48) was used to generate the Sankey diagram.

Virome composition in relation to host functional traits and clades

We examined whether plant virome composition was associated with functional traits and phylogeny of their hosts. Our analyses which included all plant virus transcripts revealed associations between the phylogenetic placement of species (indicated by host clade) and virome diversity. Notably, we observed significant associations between host clade and Shannon diversity (χ^2=388, df=7, p=0.025) in both algae (Tukey: z=3.210, p=0.026) and gymnosperms (Tukey: z=-3.110, p=0.035). These groups had significantly lower virus diversity relative to the basal eudicots. However, no significant difference in virus richness was observed between host clades (χ^2=388, df=7, p=0.058), although both algae and gymnosperms exhibited low virus richness (Figure 4). Host clade was also a significant predictor of beta diversity (the differences in diversity of viruses between host species) although it explains little of the variability in our model (R^2=0.057, df=7, p=0.001).

While host clade was not found to be significantly associated with total viral abundance in our model (p=0.209), an indicator species analysis found several associations between virus families and host clades. As expected, the *Marnaviridae* were characteristic of algae but also non-vascular plants (exclusively species of moss) (p=0.001, A=0.873, B=0.219). Furthermore, *Alphaflexiviridae* were found to be indicators of four clades, basalmost angiosperms, fern and fern allies, monocots and non-vascular plants (p=0.017, A=0.8925, B=0.1863). The *Alphaflexiviridae* is largely (but not completely) restricted to this group whereas it was also found in the core eudicots. Lastly, the *Caulimoviridae* were characteristic of all plant clades except for the ferns and fern allies where they were not detected (p=0.003, A=1, B=0.478). The caulimovirids were the most frequently detected viruses, appearing in 15% of all plant species surveyed and 43% of species in which a virus was detected.

Additionally, growth form was a strong predictor of total virus abundance in the Spermatophytes (gymnosperms and angiosperms) (χ^2=1612, df=3, p=0.005), with climbers exhibiting higher viral abundance relative to all other growth forms. Despite this, there was no significant difference detected between growth forms in a Tukey's post hoc analysis (herbs (Tukey: z=-4.9861, p=0.519), shrubs (Tukey: z=-9.590, p=0.072), trees (Tukey: z=-9.7610, p=0.063) (Figure 3)). An indicator species analysis found that viruses belonging to the families *Chrysoviridae and Rhabdoviridae* were characteristic of climbers (p=0.026, p=0.021, respectively). While 76% of *Chrysoviridae* and 62% of *Rhabdoviridae* sequences were found in climbers these viruses were not frequently detected throughout our samples. In fact, among the climbers, *Chrysoviridae* and *Rhabdoviridae* were only detected in two and three plant species, respectively.

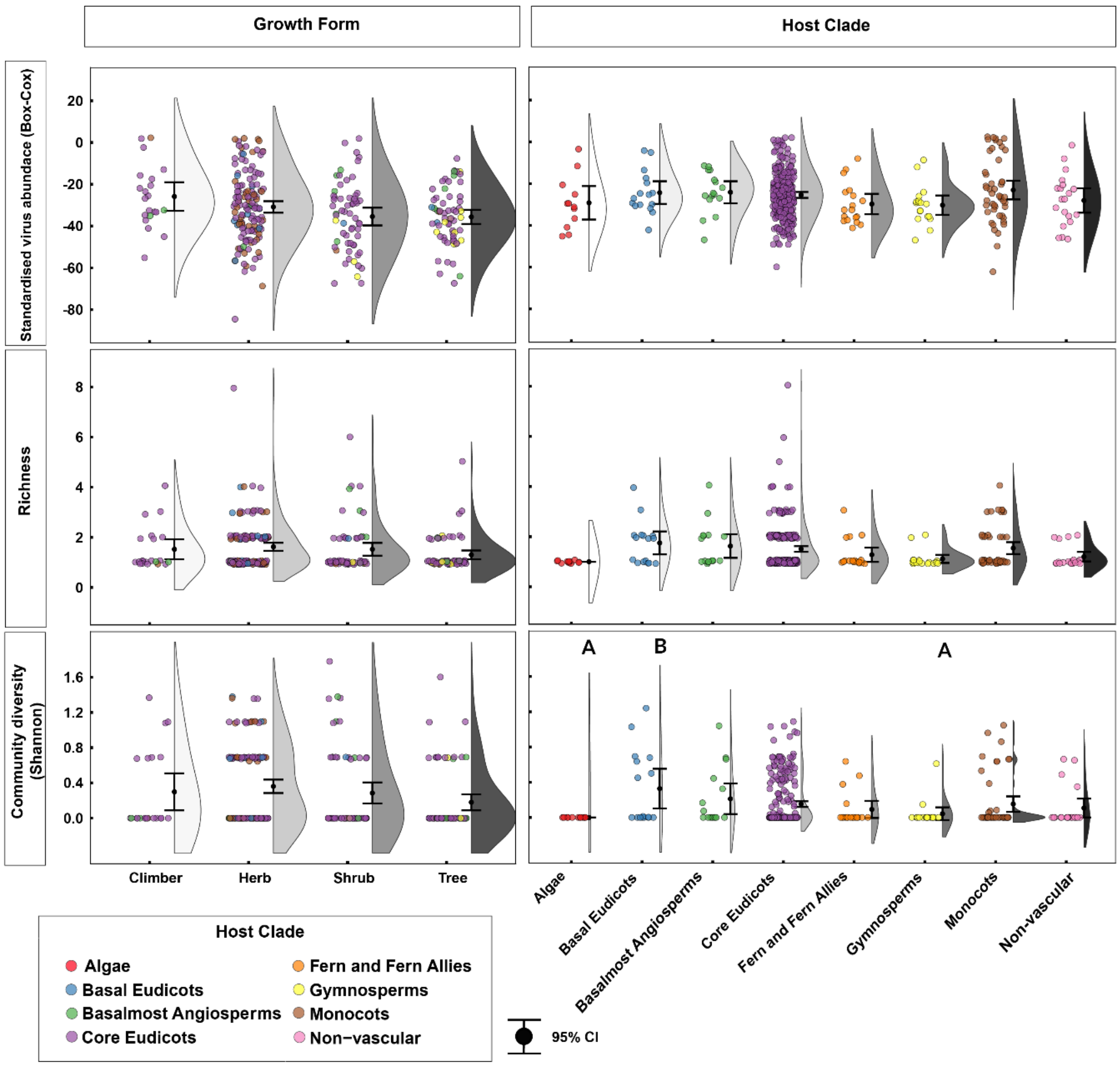


Figure 3. Raincloud plots of the relationship between two explanatory variables (growth form, host clade) and each of three measures of virus diversity and abundance (total standardised abundance of viral transcripts, richness and community diversity (Shannon diversity)). A Box–Cox transformation was applied to total standardized virus abundance. Each point represents a plant species and is coloured by its host clade. Different letters (i.e. A vs B) indicate significant differences between groups in generalized linear models as determined by post-hoc Tukey tests. 95% confidence intervals around the mean are displayed. Vertical and horizontal jitter has been applying to each data point to aid in visualisation. The R package RainCloudPlots (49) was used to construct the plots.

Phylogenetic description of viruses found

To infer the phylogenetic relationships between the viruses identified, order and family-level phylogenetic trees were estimated using the highly conserved viral region that comprises the RdRp. In total, we assembled 57 RdRp contigs (Table 2). Sequences that represent potentially novel virus species or markedly extend the known host range of a virus group are discussed below. While we detected plant virus-like transcripts belonging to 21 different virus families we did not analyse the RdRp/polymerase contigs for eight of these families as the alignments generated from RdRp-like fragments were of poor quality.

Positive-sense single-stranded RNA ((+)ssRNA) viruses

Benyviridae We identified four beny-like sequences which may represent the first benyvirids found in lower plants including the bird's-nest fern benyvirus (BnfBV) (*Asplenium nidus*), tomato fern benyvirus (TomfBV) (*Lonchitis hirsuta*), Leucodon julaceus benyvirus (LjBV) (*Leucodon julaceus*) and Wallace's spikemoss benyvirus (WasBV) (*Selaginella wallacei*). BnfBV and TomfBV share 61% and 60% amino acid identity respectively to wheat stripe mosaic virus (50). Together with wheat stripe mosaic virus, BnfBV and TomfBV represent a well-supported clade separate from the remaining plant benyviruses (Figure 4, SI figure 2).

The triple gene block (TGB) is a hallmark gene module of the *Benyviridae* among several other alpha-supergroup viruses (51). An additional ORF (111 amino acids) was assembled for the TomfBV which shared similarities to the TGB protein 1 of other benyvirids including burdock mottle virus (e-value = 4e-13, amino acid identity = 36%) and other members of the alpha-like supergroup. This may represent the first TGB protein found outside of flowering plants, although the full protein was unable to be assembled. Phylogenetic analysis placed this sequence basal to the *Benyviridae* (SI figure 3).

The remaining two viruses, LjBV and WasBV share 58% and 53% amino acid identity to Diabrotica undecimpunctata virus 2, an unclassified virus recently identified from the southern corn rootworm (52). Together with several unclassified invertebrate, fungi, and soil-derived viruses, LjBV and WasBV form a well-supported group basal to all plant benyvirids and potentially constitute a novel virus group (SI figure 2).

Due to the phylogenetic placement of LjBV and WasBV around viruses infecting distant hosts (e.g. invertebrates and fungi) we investigated the potential of contamination from other eukaryotes as the source of these viruses. Of note, the Wallace's spikemoss metatranscriptome contained reads which matched to various fungi orders (7% of all reads) as well as reads which matched to the plant-parasitic oomycete *Albugo laibachii* (7%) (SI Figure 4).

Alphaflexiviridae We identified 25 virus sequences that fell within the order *Tymovirales*. Seven viruses clustered with known plant viruses within the *Alphaflexiviridae*. Of particular interest was a divergent virus sequence found in a blue agave (*Agave tequilana*) tentatively named blue agave alphaflexivirus (BlaAV). BlaAV shares 42% amino acid identity with its closest relative vanilla virus X. This virus is phylogenetically positioned with Lolium latent virus, the sole member of the genus *Lolavirus* (Figure 4). Three additional ORFs were assembled which resemble the TGB. TGB protein 1 shared similarities to the potexvirus Phaius virus X (e-value = 7e-36, amino acid identity = 39%) and in a phylogram of the TGB protein 1 sequence was placed with the potexviruses (SI figure 3).

Betaflexiviridae Ten virus transcripts were associated with the *Betaflexiviridae* including a sequence found in sea beet (*Beta vulgaris* subsp. *maritima*), tentatively named sea beet betaflexivirus (SbBV). SbBV shared 66% amino acid identity with its closest relative maize-associated trichovirus 1 – an unclassified trichovirus. SbBV, maize-associated trichovirus and maize-associated betaflexivirus, together with the citriviruses form a weakly supported sister group to the other known trichoviruses (Figure 4).

Capillo-like virus sequences were found in Iranian poppy (*Papaver bracteatum)* and *Linum macraei* samples. The first sequence, Iranian poppy betaflexivirus (IpBV) shares 62% amino acid identity with cherry virus A while the second sequence Linum macraei betaflexivirus (LimBV) shares 59% amino acid identity with Hobart betaflexivirus 1 — an unclassified betaflexivirus found in the European honey bee (*Apis mellifera*) (53). Both sequences phylogenetically cluster with known capilloviruses and potentially represent novel virus species (Figure 4).

Tymoviridae Four virus-like sequences identified clustered within the *Tymoviridae*. Ishige okamurae tymovirus (IoTV) was detected in the brown alga *Ishige okamurae* and likely represents the first virus in the order *Tymovirales* to be found in algae. IoTV shared 44% amino acid identity with an unclassified Riboviria species (QDH90244.1) detected in a soil metagenome sample. A related virus sequence was detected in Wallace's spikemoss and tentatively named Wallace's spikemoss tymovirus (WasTV). This sequence represents the first plant tymo-like virus detected in the bryophytes. WasTV shares 60% amino acid identity with an unclassified Riboviria species (QDH87807.1) detected in a soil metagenome sample. Together with unclassified Riboviria species (QDH90244.1, QDH87807.1) IoTV and WasTV form a well-supported sister group to the recently discovered Sclerotinia sclerotiorum mycotymovirus (54) (Figure 4, SI Figure 5)

We assembled two tymo-like virus sequences tentatively named Oxera neriifolia tymovirus (OnTV) found in the climber *Oxera neriifolia* and bloodroot tymovirus (BloTV) found in *Sanguinaria canadensis*. OnTV clusters with the unclassified tymovirus poinsettia mosaic virus as a sister group to the tymoviruses. BloTV is placed between the marafaviruses and maculaviruses – although this position is not particularly well supported in our phylogeny (SI Figure 5). OnTV shares 71% amino acid identity with the tymovirus, andean potato latent virus while BloTV shares 45% amino acid identity with the marafavirus, alfalfa virus F.

Deltaflexiviridae/Gammflexiviridae We assembled two sequences which clustered within the mycotymovirus families; *Gammaflexiviridae* and the recently proposed *Deltaflexiviridae* (55). We detected a sequence in the liverwort *Calypogeia fissa*, tentatively named Calypogeia fissa associated virus (CafAV) which appears distantly related to delta- and gammaflexiviruses. CafAV shared 60% amino acid identity with the polymerase of an unclassified Riboviria species (QDH87810.1) detected in a soil metagenome sample. An additional ORF (99 amino acids) was assembled which was distantly homologous to the hypothetical ORF2 of Fusarium graminearum deltaflexivirus 1 (33% amino acid identity) (56). In a phylogenetic analysis with members of the Tymovirales, CafAV together with two unclassified Riboviria species forms a well-supported group between the delta and gammaflexiviruses (SI Figure 5). The *C. fissa* library contained numerous containments including fungi and bacteria representing 15% and 33% of total reads respectively which make discerning the host association for CafAV difficult (SI Figure 4).

A gammaflexivirus-like sequence tentatively named Pinguicula agnata associated gammaflexivirus (PaAGV) was detected in a *Pinguicula agnata* sample. This virus potentially represents the first plant gammaflexivirus. PaAGV shares 32% amino acid identity with mycoflexivirus, botrytis virus F. Phylogenetic analysis placed PaAGV within the *Gammaflexiviridae* (Figure 4). It is unclear whether the source of these virus sequences is from plants or contamination from other eukaryotes. Interestingly, no fungi-associated reads were found in the *P. agnata* library suggesting a potential plant origin (SI Figure 4)

Endornaviridae An endorna-like virus sequence was detected in the green algae species *Staurastrum sebaldi* and tentatively named Staurastrum sebaldi endornavirus (SsEV). SsEV shared 48% amino acid identity with its closest relative Persea americana alphaendornavirus 1. In a phylogenetic analysis, SsEV was situated with alphaendornaviruses which infect both land plants and fungi. To our knowledge, this is the first endornavirus infecting green algae. There was little to no evidence of the detection of SsEV being as a result of fungal contamination as <1% of all reads were found to be fungi-associated (SI Figure 4)

Potyviridae We identified six virus-like sequences that clustered with plant viruses in the family *Potyviridae*. Of particular interest are two sequences, the first was assembled from a *Traubia modesta* sample and subsequently named Traubia modesta potyvirus (TramPV) and the second, assembled from a salt wort sample (*Batis maritima*), was named salt wort potyvirus (SawPV). TramPV shared 80% amino acid identity with potato virus Y strain N while SawPV shared 77% amino acid identity with sunflower ring blotch virus. Both sequences cluster with known potyviruses in a phylogenetic analysis of the NIb gene (Figure 4).

Secoviridae We detected seven sequences which shared sequence similarity to members of the *Secoviridae*. Of particular interest was a sequence which likely represented the first secovirus detected in the bryophytes. This sequence was assembled from a common water moss (*Fontinalis antipyretica)* tentatively named common water moss secovirus (CwmSV). The polymerase region of CwmSV shared 46% amino acid identity to its closest relative blackcurrant reversion virus. The "GDD" polymerase motif was present in CwmSV but the Pro-Pol region of RNA-1 could not be assembled. Phylogenetic analysis of the RdRp region placed CwmSV with viruses in the genera Nepovirus (Figure 4). A putative RNA2 ORF (987 amino acids) was assembled for CwmSV containing a partial movement protein (MP) and a complete single coat protein (CP). CwmSV RNA2 shared 40% amino acid identity with blackcurrant reversion virus. CwmSV CP groups with nepovirus subgroup C in a phylogeny with nepovirus CP sequences (SI Figure 6).

We found little evidence that CwmSV was detected due to contamination by land plants or other eukaryotes. The *F. antipyretica* metatranscriptome was largely composed of feather moss (Hypnales) reads (34%) to which *F. antipyretica* belongs as well as reads matching an uncultured eukaryote 18S rRNA gene (54%) (HG421124.1) which is identical to the *F. antipyretica* 18S rRNA (AF023714.1) among other bryophyte 18S rRNA genes in a blastn search (e-value = 2e-102, nucleotide identity = 100%) (SI Figure 4).

A highly divergent virus sequence was assembled in the plant *Salix dasyclados* and tentatively named Salix dasyclados secovirus (SadSV). SadSV shared 30% amino acid identity with peach leaf pitting-associated virus. Phylogenetic analysis placed SadSV in a well-supported group of waikaviruses (Figure 4)

Negative-sense single-stranded RNA ((-)ssRNA) viruses

Rhabdoviridae We identified five sequences that clustered with plant viruses in the family *Rhabdoviridae*. A nucleorhabdovirus-like sequence was detected in common ivy (*Hedera helix*) tentatively named common ivy rhabdovirus 1 (CoiRV1). CoiRV1 shares 66% amino acid identity to Datura yellow vein nucleorhabdovirus and clusters with known plant nucleorhabdovirus (Figure 4)

Two cytorhabdovirus-like sequences were detected in common ivy and Canadian violet (*Viola canadensis*) tentatively named common ivy rhabdovirus 2 (CoiRV2) and Canadian violet rhabdovirus (CvRV) respectively. CoiRV2 shared 64% amino acid identity to lettuce necrotic yellows virus and CvRV shared 80% amino acid identity to persimmon virus A. Together CoiRV2 and CvRV form a well-supported clade with known plant cytorhabdovirus (Figure 4).

Lastly, two novel varicosavirus-like sequences were detected in *Goodyera pubescens* and Indian pipe (*Monotropa uniflora*) tentatively named Goodyera pubescens rhabdovirus (GopRV) and Indian pipe rhabdovirus (InpRV). GopRV shared 50% amino acid identity to red clover varicosavirus, InpRV shared 42% amino acid identity with black grass varicosavirus-like virus. Together GopRV and InpRV form a well-supported clade with known plant varicosavirus (Figure 4).

Double-stranded RNA (dsRNA) viruses

Partitiviridae We detected five sequences that share a resemblance with members of *Partitiviridae*. All sequences were found in eudicots and cluster with known partitiviruses. Of particular interest is an alphapartitivirus-like sequence detected in the great lobelia (*Lobelia siphilitica*) and tentatively named great lobelia partitivirus (GrlPV). GrlPV shared 68% amino acid identity to cannabis cryptic virus. Phylogenetic analysis places GrlPV in a well-supported branch of alphapartitiviruses (Figure 4).

Table 2. Summary of the plant-associated viral sequences assembled in this study.

Virus name, (abbreviation)	**Virus family**	**Host**	**Contig length (aa)**	**Relative abundance (%)**	**Closest match (GenBank accession number)**	**Amino acid similarity (%)**
Blue agave alphaflexivirus (BlaAV)	*Alphaflexiviridae*	Blue agave (*Agave tequilana*)	1359	0.3205%	Vanilla virus X (YP_009389479.1)	41.79%*
Cymbidium mosaic virus (CymMV) (1)	*Alphaflexiviridae*	String-of-pearls (*Senecio rowleyanus*)	204	0.0005%	Cymbidium mosaic virus (ABO41877.1)	98.04%
Cymbidium mosaic virus (CymMV) (2)	*Alphaflexiviridae*	Papyrus (*Cyperus papyrus*)	428	0.2527%	Cymbidium mosaic virus (ALJ56061.1)	97.67%
Cymbidium mosaic virus (CymMV) (3)	*Alphaflexiviridae*	*Oncidium sphacelatum*	1196	10.5537%	Cymbidium mosaic virus (AAS87218.1)	95.69%
Garlic virus B (GarV-B)	*Alphaflexiviridae*	Garlic (*Allium sativum*)	1532	0.5245%	Garlic virus B (QED43533.1)	92.04%
Schlumbergera virus X (SchVX)	*Alphaflexiviridae*	Leaf cactus (*Pereskia aculeata*)	841	0.0269%	Schlumbergera virus X (AJF19167.1)	98.81%
Winter's bark alphaflexivirus (WbAV)	*Alphaflexiviridae*	Winter's bark (*Drimys winteri*)	199	0.00%	Schlumbergera virus X (YP_002341559.1)	78.17%*
Bird's-nest fern benyvirus (BnfBV)	*Benyviridae*	Bird's-nest fern (*Asplenium nidus*)	302	0.00%	Wheat stripe mosaic virus (AYD38100.1)	60.96%*

Leucodon julaceus benyvirus (LjBV)	*Benyviridae*	Leucodon moss *(Leucodon julaceus)*	61	0.0001%	Diabrotica undecimpunctata virus 2 (QIT20101.1)	58.18%*
Tomato fern benyvirus (TomfBV)	*Benyviridae*	Tomato fern (*Lonchitis hirsuta*)	104	0.00%	Wheat stripe mosaic virus (QII15619.1)	59.62%*
Wallace's spikemoss benyvirus (WasBV)	*Benyviridae*	Wallace's spikemoss (*Selaginella wallacei*)	289	0.0001%	Diabrotica undecimpunctata virus 2 (QIT20101.1)	53.79%*
Aconitum latent virus (AcLV)	*Betaflexiviridae*	*Corydalis linstowiana*	318	0.0014%	Aconitum latent virus (NP_116487.1)	95.91%
Iranian poppy betaflexivirus (IpBV)	*Betaflexiviridae*	Iranian poppy (*Papaver bracteatum*)	292	0.00%	Cherry virus A (ATJ05023.1)	61.89%*
Lily symptomless virus (LSV) (1)	*Betaflexiviridae*	*Allium commutatum*	179	0.006%	Lily latent virus (CAB57958.1)	93.53%
Lily symptomless virus (LSV) (2)	*Betaflexiviridae*	Sargent's lily (*Lilium sargentiae*)	396	0.00%	Lily symptomless virus (BAT32749.1)	95.45%
Linum macraei betaflexivirus (LimBV)	*Betaflexiviridae*	*Linum macraei*	513	0.00%	Hobart betaflexivirus 1 (AWK77906.1)	58.67%*
Nerine latent virus (NeLV)	*Betaflexiviridae*	Belladonna lily (*Amaryllis belladonna*)	813	0.4447%	Nerine latent virus (AFJ92914.1)	99.25%
Passiflora latent virus (PLV) (1)	*Betaflexiviridae*	Katsura tree (*Cercidiphyllum japonicum*)	181	0.0004%	Passiflora latent virus (AXL95764.1)	97.79%

Passiflora latent virus (PLV) (2)	*Betaflexiviridae*	Spotted laurel (*Aucuba japonica*)	185	0.0027%	Passiflora latent virus (AXL95764.1)	97.30%
Potato virus S (PVS)	*Betaflexiviridae*	Bluecrown passionflower (*Passiflora caerulea*)	1292	0.9363%	Potato virus S (CRK77055.1)	94.12%
Sea beet betaflexivirus (SbBV)	*Betaflexiviridae*	Sea beet (*Beta maritima*)	201	0.00%	Maize-associated trichovirus 1 (QJC70224.1)	65.78%*
Cucumber mosaic virus (CMV)	*Bromoviridiae*	Beaked triggerplant (*Stylidium adnatum*)	324	0.0128%	Cucumber mosaic virus (BAD15370.1)	99.37%
Calypogeia fissa associated deltaflexivirus (CafAV)	*Deltaflexiviridae*	*Calypogeia fissa*	552	0.0011%	Riboviria sp. (QDH87810.1)	59.61%*
Staurastrum sebaldi endornavirus (SsEV)	*Endornaviridae*	*Staurastrum sebaldi*	131	0.00%	Persea americana alphaendornavirus (YP_005086952.1)	48.06%*
Pinguicula agnata associated gammaflexivirus (PaAGV)	*Gammaflexiviridae*	*Pinguicula agnata*	178	0.0001%	Botrytis virus F (NP_068549.1)	32.34%*
Cannabis sativa partitivirus (CasPV)	*Partitiviridae*	*Cannabis sativa*	340	0.0004%	Beet cryptic virus 1 (YP_002308574.1)	73.47%*
Great lobelia partitivirus (GrlPV)	*Partitiviridae*	Great lobelia (*Lobelia siphilitica*)	525	0.00%	Cannabis cryptic virus (YP_009293586.1)	67.99%*
Pittosporum cryptic virus-1 (PiCV1)	*Partitiviridae*	Tie bush (*Wikstroemia indica*)	124	0.0001%	Pittosporum cryptic virus-1 (ADE34113.1)	95.97%

String-of-pearls partitivirus (SopPV)	*Partitiviridae*	String-of-pearls (*Senecio rowleyanus*)	232	0.00%	Rose partitivirus (ANQ45203.1)	83.02%*
White campion partitivirus (WcPV)	*Partitiviridae*	White campion (*Silene latifolia*)	381	0.00%	Beet cryptic virus 1 (YP_002308574.1)	87.50%*
Asclepias yellow vein virus (AYVV)	*Potyviridae*	Common milkweed (*Asclepias syriaca*)	198	0.0004%	Asclepias yellow vein virus (QBZ81841.2)	96.97%
Henbane mosaic virus (HMV)	*Potyviridae*	Sticky nightshade (*Solanum sisymbriifolium*)	436	0.00%	Henbane mosaic virus (AZL49328.1)	97.93%
Lettuce mosaic virus (LMV)	*Potyviridae*	Common sneezeweed (*Helenium autumnale*)	334	0.0008%	Lettuce mosaic virus (AIB00279.1)	98.19%
Lily mottle virus (LMoV)	*Potyviridae*	Sargent's lily (*Lilium sargentiae*)	2133	0.00%	Lily mottle virus (BAJ10467.1)	98.83%
Salt wort potyvirus (SawPV)	*Potyviridae*	Salt wort (*Batis maritima*)	597	5.3578%	Sunflower ring blotch virus (YP_009351870.1)	77.14%*
Traubia modesta potyvirus (TramPV)	*Potyviridae*	*Traubia modesta*	285	0.0002%	Potato virus Y strain N (CAB57887.1)	79.36%*
Canadian violet rhabdovirus (CvRV)	*Rhadboviridae*	Canadian violet (*Viola canadensis*)	142	0.0001%	Persimmon virus A (YP_006576506.2)	80.28%*
Common ivy rhadbovirus 1 (CoiRV1)	*Rhadboviridae*	Common ivy (*Hedera helix*)	215	0.00%	Datura yellow vein nucleorhabdovirus (AGN98125.1)	65.58%*

Common ivy rhadbovirus 2 (CoiRV2)	*Rhadboviridae*	Common ivy (*Hedera helix*)	317	0.00%	Lettuce necrotic yellows virus (YP_425092.1)	63.80%*
Goodyera pubescens rhabdovirus (GopRV)	*Rhadboviridae*	*Goodyera pubescens*	114	0.0001%	Red clover varicosavirus (AUD57853.1)	50.00%*
Indian pipe rhabdovirus (InpRV)	*Rhadboviridae*	Indian pipe (*Monotropa uniflora*)	1068	0.00%	Black grass varicosavirus-like virus (YP_009130620.1)	42.32%*
Arabis mosaic virus (ArMV)	*Secoviridae*	Stemless gentian (*Gentiana acaulis*)	767	0.0908%	Arabis mosaic virus (ADJ39329.1)	94.47%
Broad bean wilt virus 1 (BBWV1)	*Secoviridae*	Common bugle (*Ajuga reptans*)	303	0.0001%	Broad bean wilt virus 1 (NP_951030.1)	96.65%
Broad bean wilt virus 2 (BBWV2) (1)	*Secoviridae*	African violet (*Saintpaulia ionantha*)	504	0.0437%	Broad bean wilt virus 2 (BAA34928.1)	98.98%
Broad bean wilt virus 2 (BBWV2) (2)	*Secoviridae*	*Rehmannia glutinosa*	528	0.00%	Broad bean wilt virus 2 (AFW04233.1)	99.14%
Common water moss secovirus (CwmSV)	*Secoviridae*	Common water moss (*Fontinalis antipyretica*)	414	0.0008%	Blackcurrant reversion virus (NP_734045.1)	55.31%*
Salix dasyclados secovirus (SadSV)	*Secoviridae*	*Salix dasyclados*	185	0.00%	Peach leaf pitting-associated virus (ATD53314.1)	30.46%*
Tobacco ringspot virus (TRSV)	*Secoviridae*	*Poliomintha bustamanta*	468	0.904%	Tobacco ringspot virus (QIC89926.1)	98.72%

Ihi tombusvirus (IhiTV)	*Tombusviridae*	Ihi (*Portulaca molokiniensis*)	517	0.0131%	Honeysuckle ringspot virus (YP_004191788.1)	65.90%*
Tomato spotted wilt tospovirus (TSWV) (1)	*Tospoviridae*	*Tragopogon castellanus*	108	0.00%	Tomato spotted wilt tospovirus (ACO72718.1)	98.11%
Tomato spotted wilt tospovirus (TSWV) (2)	*Tospoviridae*	*Tragopogon dubius*	181	0.00%	Tomato spotted wilt tospovirus (QAU55720.1)	96.69%
Tomato spotted wilt tospovirus (TSWV) (3)	*Tospoviridae*	*Tragopogon porrifolius*	152	0.00%	Tomato spotted wilt tospovirus (ACO72671.1)	100.00%
Bloodroot tymovirus (BloTV)	*Tymoviridae*	Bloodroot (*Sanguinaria canadensis*)	1578	0.0149%	Alfalfa virus F (YP_009551972.1)	44.52%*
Ishige okamurae tymovirus (IoTV)	*Tymoviridae*	*Ishige okamurae*	769	0.066%	Riboviria sp. (QDH90244.1)	44.03%*
Oxera neriifolia tymovirus (OnTV)	*Tymoviridae*	*Oxera neriifolia*	159	0.0001%	Andean potato latent virus (AAC25015.1)	71.07%*
Wallace's spikemoss tymovirus (WasTV)	*Tymoviridae*	Wallace's spikemoss (*Selaginella wallacei*)	269	0.00%	Riboviria sp. (QDH87807.1)	59.32%*

* Virus likely represent a novel species.

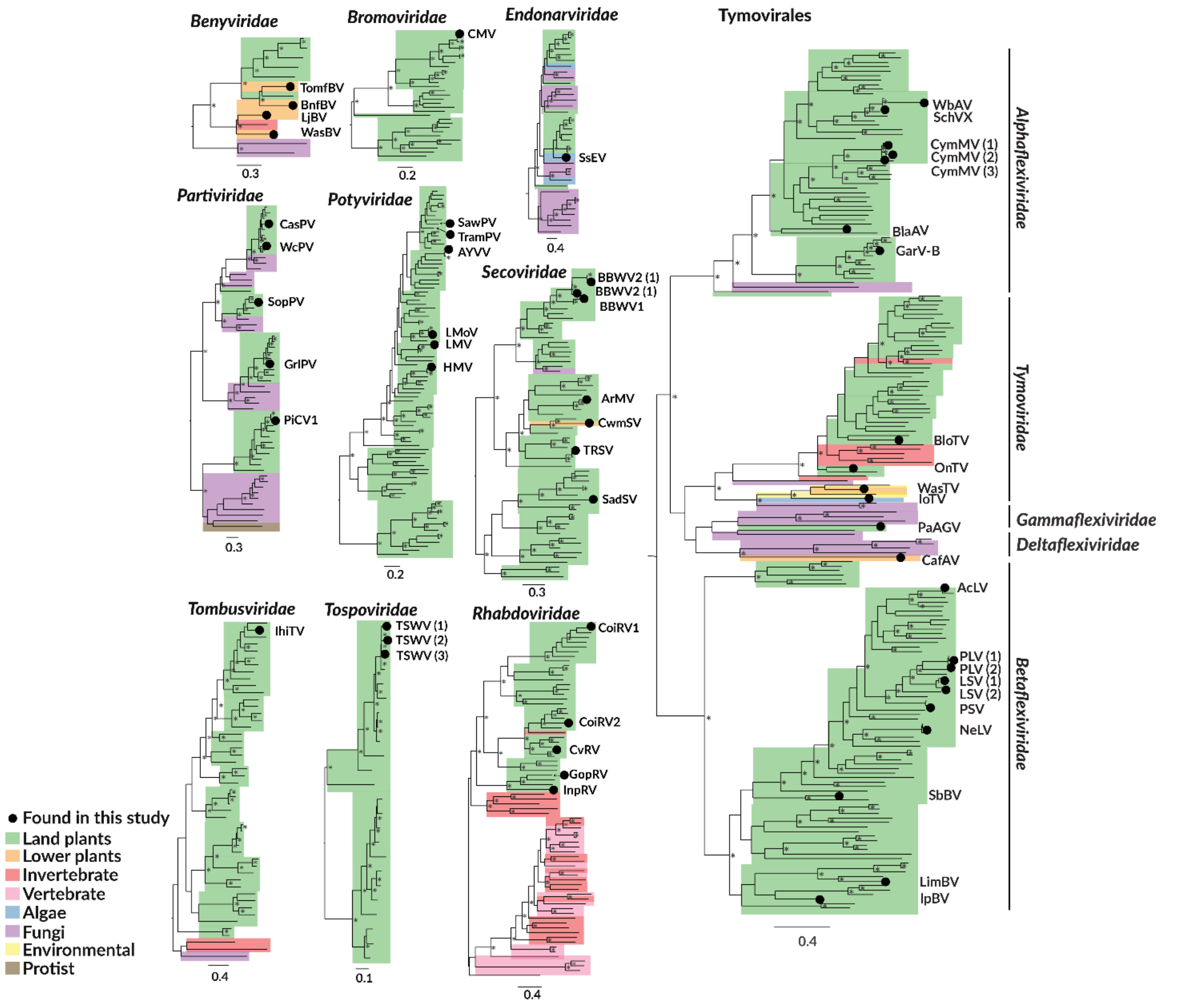


Figure 4. Phylogram of the virus RNA-dependent RNA polymerase containing contigs assembled in this study. Maximum likelihood phylogenetic trees show the topological position of discovered virus-like sequences (black circles) from this study in the context of their closest relatives. Branches are highlighted to represent host clade (land plants = green, lower plants = orange, invertebrate = red, vertebrate = pink, algae = blue, fungi = purple, yellow = environmental, brown = protist). Here "Land plants" encompasses both angiosperms and gymnosperms while "Lower plants" includes the bryophytes, lycophytes, and ferns. An abbreviation for each virus-like sequence identified in this study is provided. All branches are scaled to the number of amino acid substitutions per site and trees were mid-point rooted for clarity only. An asterisk indicates node support of >70% bootstrap support. Full names and accession numbers for protein sequences used in the alignment are in SI Table 5. A list of SRA IDs associated with each virus RdRp sequence assembled in this study is provided in SI Table 6. Where two viruses of the same species were found they are assigned a numeric identifier in brackets.

Discussion

Our ability to reconstruct the evolutionary history of plant viruses and understand the drivers of their emergence has been constrained by inadequate sampling across the wide diversity of plant species. Here, we have conducted a large scale virus discovery project in plants by mining transcriptomes from across the entire breadth of the plant kingdom. In doing so we have identified 57 viruses, of which 29 are potentially novel virus species, including the first algal virus in the tymovirids and several divergent beny-like viruses of moss and ferns. We also show that virome composition is associated with particular plant functional traits (i.e. the climbing growth form) and that viral diversity varies markedly between host clades. Collectively, this new knowledge advances our understanding of the plant virosphere and provides a strong foundation for the integration of disease dynamics into fields such as ecology and conservation management.

The diversity of the plant virome is shaped by host clade. That is, the more ancient plant lineages (algae, gymnosperms) had significantly lower virus diversity (Shannon index) relative to more recently evolved groups, like the basal eudicots. This was to be expected as host species richness is predicted to be a key determinant of virome diversity (7). Indeed, species richness in algae and gymnosperms is approximately eight and 330 times lower, respectively, than in angiosperms (57) and this lower species richness may translate to marked differences in virome composition. Given that the host range of plant viruses is commonly determined by the breadth of their biological vectors (58) the lower virus richness in gymnosperms and lower plants may also be related to the diversification of aphids associated with the rise of the angiosperms. Evidence suggests that present-day aphid diversity likely emerged after one or several shifts from gymnosperms to angiosperms driven by the rapid diversification of plants during the Cretaceous (59, 60). Another contributing factor may be that the proportion of detectable viruses is lower outside of cultivated species due to biases within our reference databases and the reliance on virus sequence homology. As little to nothing is known about the viruses of many plant lineages (e.g. red algae and glaucophytes) our ability to detect divergent viruses is lessened. Indeed, no viruses were detected in the glaucophytes while a single marnavirus transcript was detected in a red alga host. Similar difficulties with identifying viruses in uncultivated or environmental samples has been described elsewhere (11, 61, 62)

Notably, we also found clear associations between key plant functional traits (host growth form) and total virus abundance. Climbing plants exhibited higher viral abundance relative to all other growth forms examined (trees, shrubs, herbs). Potex- and carlavirus sequences appear highly abundant in several climber libraries. In the case of the climbing bluecrown passionflower (*P. caerulea*), carlaviruses account for 13.4% of all reads in the library. Understanding the mechanisms that drive high virus abundance in climbing plants is beyond the scope of our exploratory analysis. However, one promising area of future study may be to investigate the relationship between leaf nutrient chemistry – specifically phosphorus (P) and nitrogen (N) – and virus abundance. Climbers are well known to exhibit higher concentrations of N and P relative to co-occurring free-standing species (63, 64). Given that viruses rely on host nutrients to complete their epidemiological cycle, infection and proliferation can be strongly impaired by N and P-limitation. The high leaf nutrient contents of climbers may provide a particularly favourable abiotic environment for virus replication (65, 66).

Discovery of viruses in the ferns, lower plants and algae

To date, viral surveys in basal plant lineages (namely ferns, bryophytes, lichens and algae) have revealed only the minimal occurrence of (+)RNA viruses (7, 67-70), supporting the idea that much of the extant RNA virome of angiosperms evolved as they diversified during the Cretaceous (71). Yet our results may challenge this paradigm; we detected the first evidence of several (+)ssRNA families in the lower plants and algae implying that these groups may have evolved much earlier. Below we explore the specific virus families involved and the implications of this novel finding.

Divergent benyviruses in the lower plants. The sequences detected in this study represent the first beny-like viruses identified in ferns and mosses. Benyviruses are typically transmitted by the root-infecting plasmodiophorids *Polymyxa betae* and *Polymyxa graminis* (50, 72). The Phytomyxids (plasmodiophorids and phagomyxids) are parasites of plants, diatoms, oomycetes and brown algae and have been shown to demonstrate cross-kingdom host shifts (e.g. between angiosperms and oomycetes) (73). As such the plasmodiophorids may be a vehicle for cross-species transmission between aquatic protists and land plants (7) Two beny-like viruses identified in this study, TomfBV and BnfBV form a clade along with wheat stripe mosaic virus, distinct from members of the genus *Benyvirus*. Whether TomfBV and BnfBV are a result of cross-species transmission from vectors such as plasmodiophorids or have co-diverged from a virus infecting the ancestors of higher plants (Charophyte algae) is not known (67). Although, no plasmodiophorid-associated contigs were detected in any of the libraries from which we assembled a beny-like virus. LjBV and WasBV appear distantly related to the benyviruses. Interestingly, these viruses group with unclassified viruses assembled from a soil metatranscriptome study suggesting that, like the benyviruses this larger group of unclassified viruses may involve soil-borne parasites like the plasmodiophorids (74)

Expanding the Tymovirales Our detection of tymovirid-like sequences in the lycophytes, liverworts and brown algae dramatically expands the known host-range of the Tymovirales. These viruses (CafAV, IoTV and WasTV) were similar to several unclassified Riboviria species assembled from a recent survey of common wild oat soil rhizosphere and detritosphere (74) (SI Figure 5). The metatranscriptome of the sequenced soil samples was largely composed of Viridiplantae, fungi, Amoebozoa, protists, nematodes, and other eukaryotes. As such, using phylogenetic clusters to infer host associations of our viruses remains challenging. Indeed, these viruses may result from cross-contamination from other eukaryotes (e.g. fungi or invertebrates) although we found no clear evidence for this (SI Figure 4). Assuming these viruses are plant-associated, their phylogenetic pattern may suggest that they have resulted from cross-kingdom transmission events that frequent the alpha-like superfamily (5).

The gammaflexi-like virus (PaAGV) we detected in *P. agnata* is particularly noteworthy. The gammaflexiviruses are only known to infect fungi, although no fungi associated reads were found in the *P. agnata* metatranscriptome (SI Figure 4). The mycovirus families *Delta*- and *Gammaflexiviridae* are thought to have been derived from the plant alpha- and betaflexivirids through cross-species transmission (7, 75). As such PaAGV could potentially represent an intermediate between the plant and fungi flexiviruses or perhaps a more recent fungus to plant transmission. As only a fragment of the polymerase gene was assembled for this virus further work is needed to confirm the presence of PaAGV and its phylogenetic position.

The first report of secoviruses in moss We detected the first secovirid-like sequence to be found in bryophytes. As all extant secoviruses fall within larger invertebrate virus groups (e.g. *Iflaviridae*), it is thought that an ancestral secovirus was first transmitted to plants from arthropods followed by co-evolution in flowering plants (7, 76). The emergence of the extant secovirids is estimated to have recently occurred in the past 500 years coinciding with both modern agriculture and the rise of marine trade (77). CwmSV could represent an ancient cross-species transmission event from invertebrates to the common ancestor of mosses and vascular plants which may have existed about 475 million years ago. Although the detectable sequence similarity between CwmSV and the extant secoviruses in angiosperms suggests a more recent host switch (78). Generally, these finding suggests that there is likely more undiscovered secoviruses in the lower plants which undoubtedly will help clarify the origins of this group.

Endornaviruses in green algae Lastly, we reported the first endornavirus (SsEV) in green algae (Charophyta). As previously shown, the topology of the *Endornavirdae* tree shares little resemblance with the phylogeny of its hosts (79). Indeed, SsEV shares little resemblance to the two-known algal endonarviruses; Phytophthora alphaendornavirus 1 and diatom colony associated dsRNA virus 15 (albeit the latter is not an officially recognized endornavirus). Like Phytophthora alphaendornavirus 1, SsEV is phylogenetically placed between a mycovirus (Ceratobasidium endornavirus A) and numerous land plant alphaendornaviruses. Thus, SsEV may provide further evidence for the presence of cross-kingdom transmission of endornaviruses between fungi and plants/algae, although the direction of transmission is not currently known (79). Considering that the abundance of this sequence was low and the small length of the contig assembled (131 amino acids), further confirmatory work is required to definitively confirm the presence of endorna-like viruses in green algae.

Caveats

It is important to note that the data generated under the 1KP were not done so for virome analysis. It is, therefore, crucial to discuss several caveats associated with our methods and the metatranscriptomic data we have mined for virus contigs. Firstly, among the 1KP samples, there was variation in the tissues sampled, phenological stage of the plant, environmental conditions and extraction protocols used, which together may influence the abundance and diversity of viruses that we're able to be detected (80-83). Secondly, the average sequencing depth of the 1KP libraries was 1.99 gigabases of sequence per sample (range 1.3-3.0) which is lower than other virus discovery studies (76, 84, 85). As sequencing depth has been shown to correlate with the ability to detect viruses present at low abundances this may have influenced our conclusions about virome composition (80, 86) Thirdly, a large proportion of the virus transcripts detected were from viruses whose full-length genomic or subgenomic mRNAs were polyadenylated at the 3′ end (Table 2, SI Figure 1). Although this was anticipated (i.e. the libraries generated by the 1KP initiative were prepared from polyA+ RNA), this may have biased our conclusions about virome composition by limiting the detection of non-polyadenylated viruses (e.g. dsRNA, dsDNA) (86). For instance, this may have contributed to the lack of phycodnavirus sequences detected in algae. Lastly, to reduce the computational burden of assembly, we attempted to remove host-associated reads before contig assembly by mapping them to the host genomes provided by the 1KP initiative. While potentially reducing the occurrence of false-positive virus detection this procedure also risks removing a small number of virus reads. Of particular concern is the removal of reverse-transcribing plant viruses which abundantly colonise genomes across the plant kingdom (87). While we frequently detected transcripts associated with the retro-transcribing family *Caulimoviridae,* no members of the *Metaviridae* or *Pseudoviridae* were detected.

Conclusions

A common goal of many plant virus discovery studies is to determine the causative agent for emergent crop diseases. By contrast, the central aim of this thesis was to use plant virus discovery techniques to gain insights into the large-scale evolutionary biology of plant viruses. The focus of plant virology has traditionally been on pathogenic viruses in species of economic importance and as a result, we know little about the viruses inhabiting uncultivated hosts, particularly in lower plants and algae. This lack of exploration across the plant kingdom has limited our ability to reconstruct the evolutionary history of plant viruses and understand the factors that shape their composition and emergence. As such, this thesis deliberately focuses on sampling viromes across the breadth of the plant kingdom. Transcriptome sequencing projects focusing on hundreds or thousands of species remain large, global collaborative initiatives requiring substantial funding and commitment. Fortunately, the rise of the open science movement has fostered a willingness to share raw data and bypass many of the resource barriers that would prevent a virus discovery project of this size. The 1KP initiative will continue to provide a foundational resource for plant science.

Across our diverse sampling of host plants, there is enormous variation in ecological strategies and functional traits. This raised the question: how much of the variation in virome composition could be explained by key traits such as growth form and longevity, or by evolutionary lineage? Utilising plant trait databases we were able to accumulate trait information for many of our plant species and showed that high virus abundance is associated with the climbing habit. This finding hints at several potential mechanistic explanations – such as differences in leaf chemistry and biomass allocation in climbers relative to free-standing species – which may shape their virome composition.

A crucial finding of this thesis is the discovery of novel sets of (+)ssRNA viruses in lower plants and algae. This result may suggest that a number of these virus families are associated with older evolutionary lineages of plants than previously thought. Many of these viruses form deep clades in phylogenetic trees and occupy ambiguous positions between established plant virus families. Further explorations into the phytovirosphere are undoubtedly needed to clarify our findings. Such efforts should seek to understand how the divergent viruses identified in lower plants relate to the large diversity of land plant, fungi, and invertebrate viruses.

More broadly, the use of metagenomics coupled with a diverse multi-host, genomic dataset has enabled us to reveal associations between plant ecological factors and virome composition. The associations found between virome composition, plant growth form and phylogenetic positioning have highlighted numerous avenues of research into the mechanisms underlying these associations. Research of this nature will certainly uncover novel plant-virus interactions and aid in revealing the ecological role of viruses in natural systems. While transcriptome mining efficiently reveals insights into the virome composition of many samples, investigations in this area will benefit from metagenomic sampling in one system, where sampling and sequencing methodologies can be selected to recover viruses at a great completeness and larger depth.

References

1. Anderson JT. Plant fitness in a rapidly changing world. New Phytol. 2016;210(1):81-7.

2. Wren JD, Roossinck MJ, Nelson RS, Scheets K, Palmer MW, Melcher U. Plant virus biodiversity and ecology. PLoS Biol. 2006;4(3):80.

3. Mifsud JCO, Geoghegan JL, Gallagher RV. Examining the diversity of the phytovirosphere. Funct Ecol. 2020 (Manuscript under review).

4. Roossinck MJ, Martin DP, Roumagnac P. Plant virus metagenomics: advances in virus discovery. Phytopathology. 2015;105(6):716-27.

5. Dolja VV, Koonin EV. Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. Virus Res. 2018;244:36-52.

6. Coy SR, Gann ER, Pound HL, Short SM, Wilhelm SW. Viruses of eukaryotic algae: diversity, methods for detection, and future directions. Viruses. 2018;10(9):487.

7. Dolja VV, Krupovic M, Koonin EV. Deep roots and splendid boughs of the global plant virome. Annu Rev Phytopathol. 2020;58:23-53.

8. Díaz S, Kattge J, Cornelissen JH, Wright IJ, Lavorel S, Dray S, et al. The global spectrum of plant form and function. Nature. 2016;529(7585):167-71.

9. Thapa V, McGlinn DJ, Melcher U, Palmer MW, Roossinck MJ. Determinants of taxonomic composition of plant viruses at the Nature Conservancy's Tallgrass Prairie Preserve, Oklahoma. Virus Evol. 2015;1(1).

10. Seabloom EW, Borer ET, Mitchell CE, Power AG. Viral diversity and prevalence gradients in North American Pacific Coast grasslands. Ecology. 2010;91(3):721-32.

11. Bernardo P, Charles-Dominique T, Barakat M, Ortet P, Fernandez E, Filloux D, et al. Geometagenomics illuminates the impact of agriculture on the distribution and prevalence of plant viruses at the ecosystem scale. ISME J. 2018;12(1):173-84.

12. Rodríguez-Nevado C, Gavilán R, Pagán I. Host abundance and identity determine the epidemiology and evolution of a generalist plant virus in a wild ecosystem. Phytopathology. 2020;0(1):94-105.

13. Greninger AL. A decade of RNA virus metagenomics is (not) enough. Virus Res. 2018;244:218-29.

14. Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, et al. One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 2019;574(7780):679-85.

15. Carpenter EJ, Matasci N, Ayyampalayam S, Wu S, Sun J, Yu J, et al. Access to RNA-sequencing data from 1,173 plant species: The 1000 Plant transcriptomes initiative (1KP). GigaScience. 2019;8(10).

16. Johnson MT, Carpenter EJ, Tian Z, Bruskiewich R, Burris JN, Carrigan CT, et al. Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. PLoS One. 2012;7(11):e50226.

17. Leinonen R, Sugawara H, Shumway M, Collaboration INSD. The sequence read archive. Nucleic Acids Res. 2010;39:19-21.

18. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9.

19. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8(8):1494-512.

20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10.

21. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59-60.

22. Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, et al. Linking virus genomes with host taxonomy. Viruses. 2016;8(3):66.

23. RStudio T. RStudio: integrated development for R. 2020.

24. Team RC. R: A language and environment for statistical computing. Vienna, Austria; 2013.

25. Wickham H, Averick M, Bryan J, Chang W, McGowan LDA, François R, et al. Welcome to the Tidyverse. Journal of Open Source Software. 2019;4(43):1686.

26. Marcelino VR, Clausen PTLC, Buchmann JP, Wille M, Iredell JR, Meyer W, et al. CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. Genome Biol. 2020;21(1):103.

27. Clausen PT, Aarestrup FM, Lund O. Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics. 2018;19(1):1-8.

28. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.

29. Maitner BS, Boyle B, Casler N, Condit R, Donoghue J, Durán SM, et al. The bien r package: A tool to access the Botanical Information and Ecology Network (BIEN) database. Methods Ecol Evol. 2018;9(2):373-9.

30. Tavşanoğlu Ç, Pausas JG. A functional trait database for Mediterranean Basin plants. Sci Data. 2018;5:180135.

31. Kattge J, Bönisch G, Díaz S, Lavorel S, Prentice IC, Leadley P, et al. TRY plant trait database–enhanced coverage and open access. Global Change Biol. 2020.

32. USDA N. The PLANTS Database [Internet]. Greensboro, NC: National Plant Data Team; 2015 [updated 2018 May 1; cited 2020 Feb 20]. Available from: http://plants.usda.gov

33. Chamberlain SA, Szöcs E. taxize: taxonomic search and retrieval in R. F1000Res. 2013;2:191.

34. Wiersema JH, León B. World economic plants: a standard reference: CRC press; 2016.

35. Rogers HS, Fricke EC. Maternal microbes complicate coexistence for tropical trees. 2019:201902736.

36. Cipollini ML, Stiles EW. Fungi as biotic defense agents of fleshy fruits: alternative hypotheses, predictions, and evidence. Am Nat. 1993;141(4):663-73.

37. Larcher W. Physiological plant ecology: ecophysiology and stress physiology of functional groups: Springer Science & Business Media; 2003.

38. Hily JM, Garcia A, Moreno A, Plaza M, Wilkinson MD, Fereres A, et al. The relationship between host lifespan and pathogen reservoir potential: an analysis in the system Arabidopsis thaliana-Cucumber mosaic virus. PLoS Pathog. 2014;10(11):e1004492.

39. Postman J, Hummer K, Ayala-Silva T, Bretting P, Franko T, Kinard G, et al., editors. GRIN-Global: An international project to develop a global plant genebank information management system. International Symposium on Molecular Markers in Horticulture 859; 2009.

40. Lagkouvardos I, Fischer S, Kumar N, Clavel T. Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons. PeerJ. 2017;5:e2836.

41. Wille M, Shi M, Klaassen M, Hurt AC, Holmes EC. Virome heterogeneity and connectivity in waterfowl and shorebird communities. ISME J. 2019;13(10):2603-16.

42. Hothorn T, Bretz F, Westfall P. Simultaneous inference in general parametric models. Biom J. 2008;50(3):346-63.

43. Cáceres MD, Legendre P. Associations between species and groups of sites: indices and statistical inference. Ecology. 2009;90(12):3566-74.

44. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772-80.

45. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972-3.

46. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268-74.

47. Rambaut A, Drummond A. FigTree: Tree figure drawing tool, version 1.4.0. Institute of Evolutionary Biology, University of Edinburgh; 2012.

48. SankeyMATIC (BETA): a sankey diagram builder for everyone [Internet]. Steve Bogart; 2015 [cited 2020 Sep 23]. Available from: http://sankeymatic.com/.

49. Allen M, Poggiali D, Whitaker K, Marshall TR, Kievit RA. Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Res. 2019;4.

50. Valente JB, Pereira FS, Stempkowski LA, Farias M, Kuhnem P, Lau D, et al. A novel putative member of the family *Benyviridae* is associated with soilborne wheat mosaic disease in Brazil. Plant Pathol. 2019;68(3):588-600.

51. Morozov SY, Solovyev AG. Triple gene block: modular design of a multifunctional machine for plant virus movement. J Gen Virol. 2003;84(6):1351-66.

52. Liu S, Valencia-Jiménez A, Darlington M, Vélez AM, Bonning BC. Diabrotica undecimpunctata virus 2, a Novel Small RNA Virus Discovered from Southern Corn Rootworm, *Diabrotica undecimpunctata howardi* Barber (Coleoptera: Chrysomelidae). Microbiol Resour Announc. 2020;9(26).

53. Roberts JMK, Anderson DL, Durr PA. Metagenomic analysis of Varroa-free Australian honey bees (*Apis mellifera*) shows a diverse Picornavirales virome. J Gen Virol. 2018;99(6):818-26.

54. Wang Q, Cheng S, Xiao X, Cheng J, Fu Y, Chen T, et al. Discovery of two mycoviruses by high-throughput sequencing and assembly of mycovirus-derived small silencing RNAs from a hypovirulent strain of Sclerotinia sclerotiorum. Front Microbiol. 2019;10:1415-.

55. Li K, Zheng D, Cheng J, Chen T, Fu Y, Jiang D, et al. Characterization of a novel Sclerotinia sclerotiorum RNA virus as the prototype of a new proposed family within the order Tymovirales. Virus Res. 2016;219:92-9.

56. Chen X, He H, Yang X, Zeng H, Qiu D, Guo L. The complete genome sequence of a novel Fusarium graminearum RNA virus in a new proposed family within the order Tymovirales. Arch Virol. 2016;161(10):2899-903.

57. Willis KJ. State of the world's plants report - 2017. Kew: Royal Botanic Gardens; 2017. 96 p.

58. Morris CE, Moury B. Revisiting the concept of host range of plant pathogens. Annu Rev Phytopathol. 2019;57(1):63-90.

59. Heie OE. Studies on Fossil Aphids (Homoptera: *Aphidoidea*): Stiftsbogtr.; 1967.

60. Peccoud J, Simon J-C, von Dohlen C, Coeur d'acier A, Plantegenest M, Vanlerberghe-Masutti F, et al. Evolutionary history of aphid-plant associations and their role in aphid diversification. C R Biol. 2010;333(6):474-87.

61. Brum JR, Ignacio-Espinoza JC, Kim E-H, Trubl G, Jones RM, Roux S, et al. Illuminating structural proteins in viral "dark matter" with metaproteomics. Proc Natl Acad Sci U S A. 2016;113(9):2436-41.

62. Rosario K, Breitbart M. Exploring the viral world through metagenomics. Curr Opin Virol. 2011;1(4):289-97.

63. Salzer J, Matezki S, Kazda M. Nutritional differences and leaf acclimation of climbing plants and the associated vegetation in different types of an Andean montane rainforest. Oecologia. 2006;147(3):417-25.

64. Tang Y, Kitching RL, Cao M. Lianas as structural parasites: A re-evaluation. Chin Sci Bull. 2012;57(4):307-12.

65. Whitaker BK, Rúa MA, Mitchell CE. Viral pathogen production in a wild grass host driven by host growth and soil nitrogen. New Phytol. 2015;207(3):760-8.

66. Maat DS, Brussaard CP. Both phosphorus-and nitrogen limitation constrain viral proliferation in marine phytoplankton. Aquat Microb Ecol. 2016;77(2):87-97.

67. Vlok M, Gibbs AJ, Suttle CA. Metagenomes of a freshwater charavirus from British Columbia provide a window into ancient lineages of viruses. Viruses. 2019;11(3).

68. Mushegian A, Shipunov A, Elena SF. Changes in the composition of the RNA virome mark evolutionary transitions in green plants. BMC Biol. 2016;14(1):68.

69. Charon J, Marcelino VR, Wetherbee R, Verbruggen H, Holmes EC. Metatranscriptomic identification of diverse and divergent RNA viruses in green and Chlorarachniophyte algae cultures. Viruses. 2020;12(10).

70. Rousvoal S, Bouyer B, López-Cristoffanini C, Boyen C, Collén J. Mutant swarms of a totivirus-like entities are present in the red macroalga *Chondrus crispus* and have been partially transferred to the nuclear genome. J Phycol. 2016;52(4):493-504.

71. Kenrick P, Crane PR. The origin and early evolution of plants on land. Nature. 1997;389(6646):33-9.

72. Tamada T, Schmitt C, Saito M, Guilley H, Richards K, Jonard G. High resolution analysis of the readthrough domain of beet necrotic yellow vein virus readthrough protein: a KTER motif is important for efficient transmission of the virus by Polymyxa betae. J Gen Virol. 1996;77(7):1359-67.

73. Neuhauser S, Kirchmair M, Bulman S, Bass D. Cross-kingdom host shifts of phytomyxid parasites. BMC Evol Biol. 2014;14(1):33.

74. Starr EP, Nuccio EE, Pett-Ridge J, Banfield JF, Firestone MK. Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. Proc Natl Acad Sci U S A. 2019;116(51):25900.

75. Ghabrial SA, Caston JR, Jiang D, Nibert ML, Suzuki N. 50-plus years of fungal viruses. Virology. 2015;479-480:356-68.

76. Shi M, Lin X-D, Tian J-H, Chen L-J, Chen X, Li C-X, et al. Redefining the invertebrate RNA virosphere. Nature. 2016;540(7634):539.

77. Thompson JR, Kamath N, Perry KL. An evolutionary analysis of the *Secoviridae* family of viruses. PLoS One. 2014;9(9):e106305.

78. Magallón S, Hilu KW, Quandt D. Land plant evolutionary timeline: gene effects are secondary to fossil constraints in relaxed clock estimation of age and substitution rates. Am J Bot. 2013;100(3):556-73.

79. Roossinck MJ, Sabanadzovic S, Okada R, Valverde RA. The remarkable evolutionary history of endornaviruses. J Gen Virol. 2011;92(11):2674-8.

80. Maclot F, Candresse T, Filloux D, Malmstrom CM, Roumagnac P, van der Vlugt R, et al. Illuminating an ecological blackbox: using high throughput Sequencing to characterize the plant virome across scales. Front Microbiol. 2020;11:2575.

81. Lacroix C, Renner K, Cole E, Seabloom EW, Borer ET, Malmstrom CM. Methodological guidelines for accurate detection of viruses in wild plant species. Applied and environmental microbiology. 2016;82(6):1966-75.

82. Zotto AD, Nome SF, Di Rienzo JA, Docampo DM. Fluctuations of Prunus necrotic ringspot virus (PNRSV) at various phenological stages in peach cultivars. Plant disease. 1999;83(11):1055-7.

83. Nachappa P, Culkin CT, Saya PM, Han J, Nalam VJ. Water stress modulates soybean aphid performance, feeding behavior, and virus transmission in soybean. Front Plant Sci. 2016;7:552.

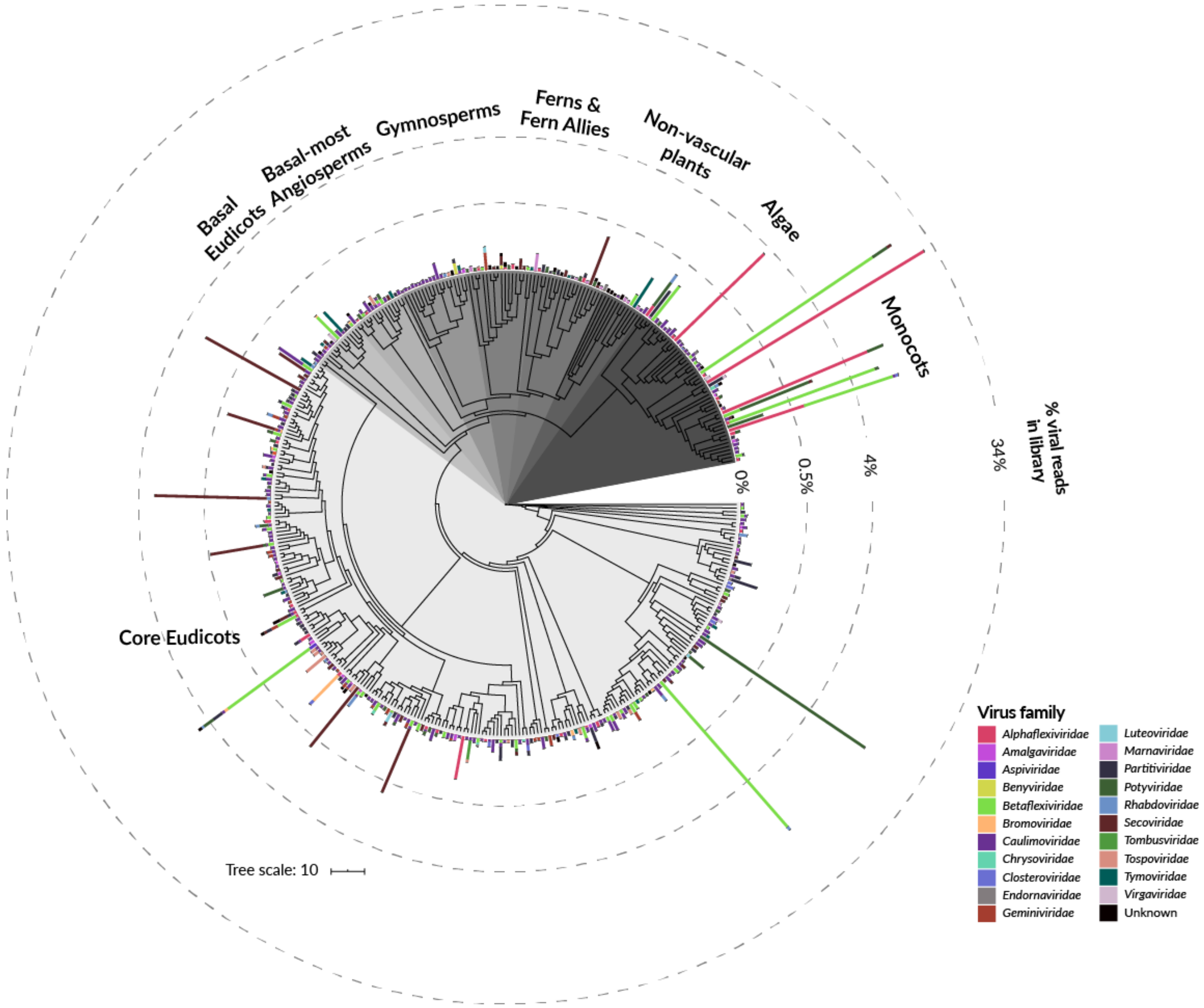
84. Shates TM, Sun P, Malmstrom CM, Dominguez C, Mauck KE. Addressing research needs in the field of plant virus ecology by defining knowledge gaps and developing wild dicot study systems. Front Microbiol. 2019;9:3305.

85. Hao X, Zhang W, Zhao F, Liu Y, Qian W, Wang Y, et al. Discovery of plant viruses from tea plant (*Camellia sinensis* (L.) O. Kuntze) by metagenomic sequencing. Front Microbiol. 2018;9:2175.

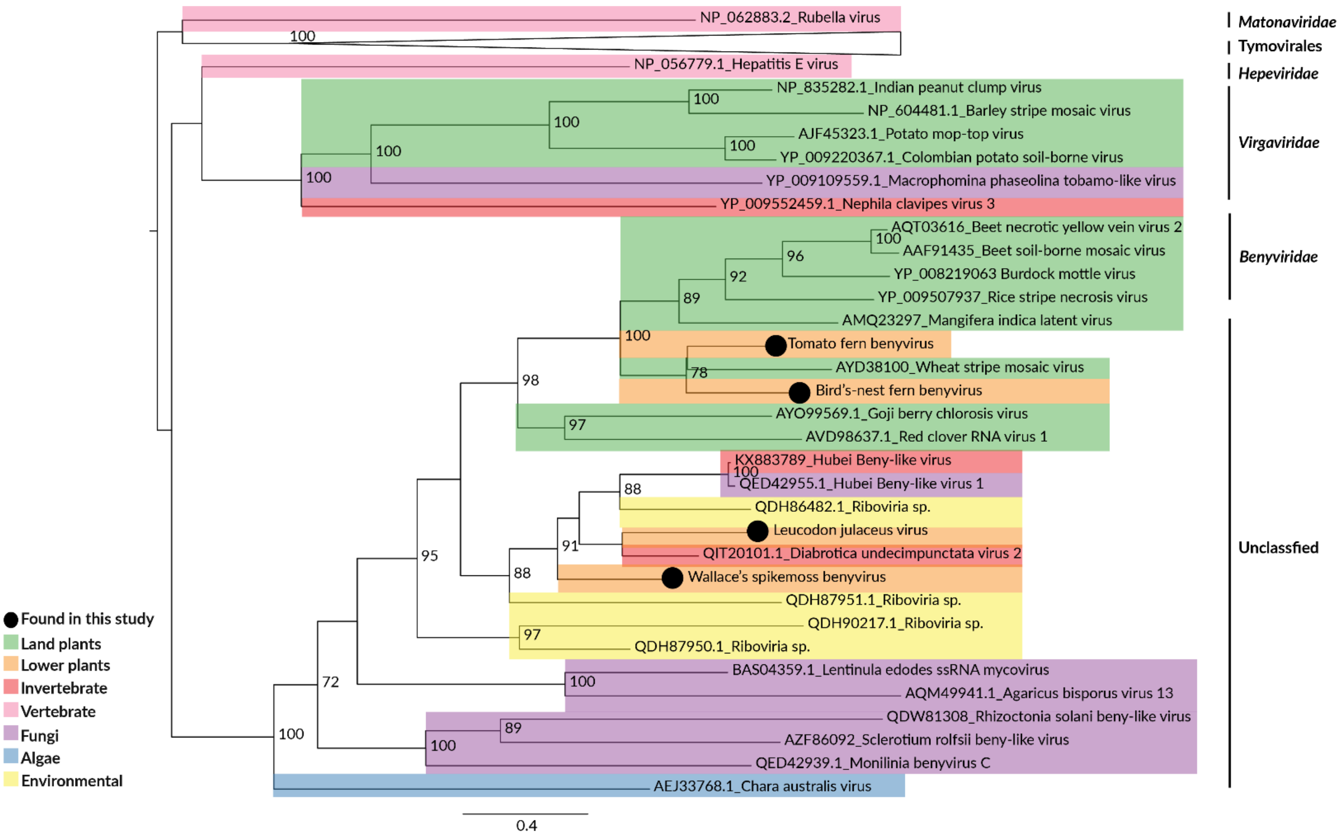
86. Visser M, Bester R, Burger JT, Maree HJ. Next-generation sequencing for virus detection: covering all the bases. Virol J. 2016;13(1):85.

87. Llorens C, Muñoz-Pomer A, Bernad L, Botella H, Moya A. Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol Direct. 2009;4(1):41.

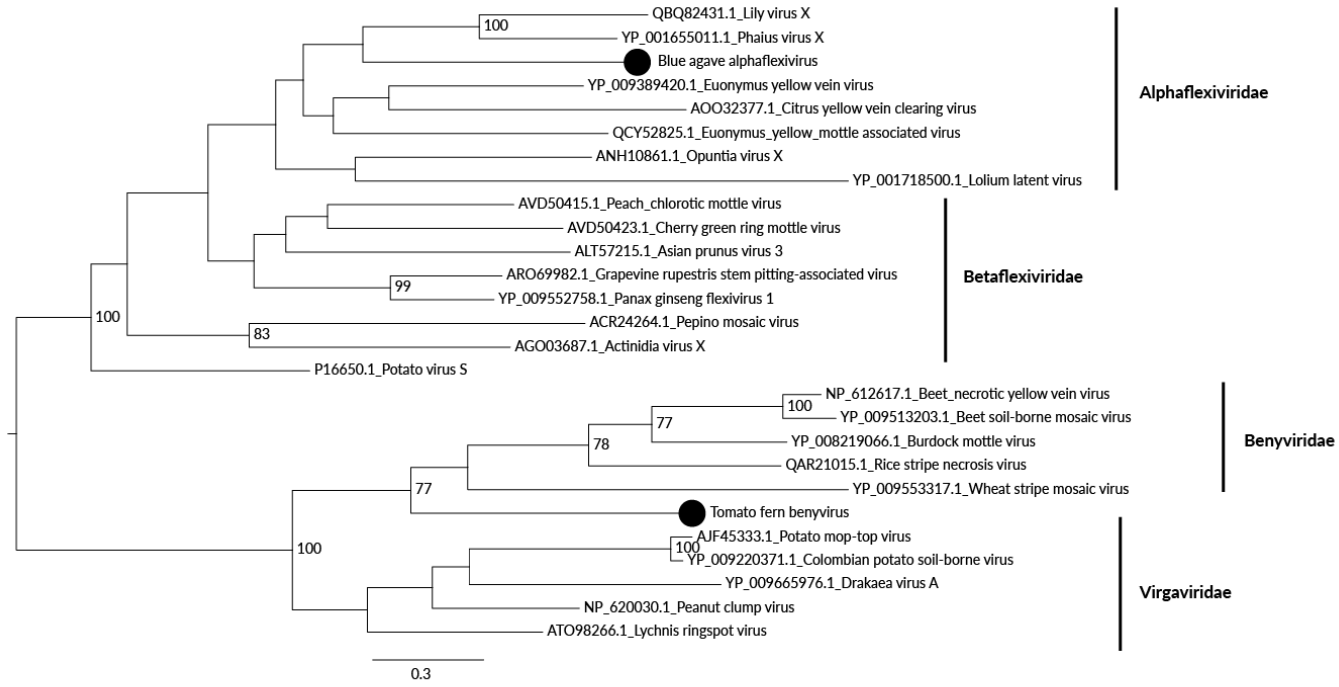
Supplementary Information



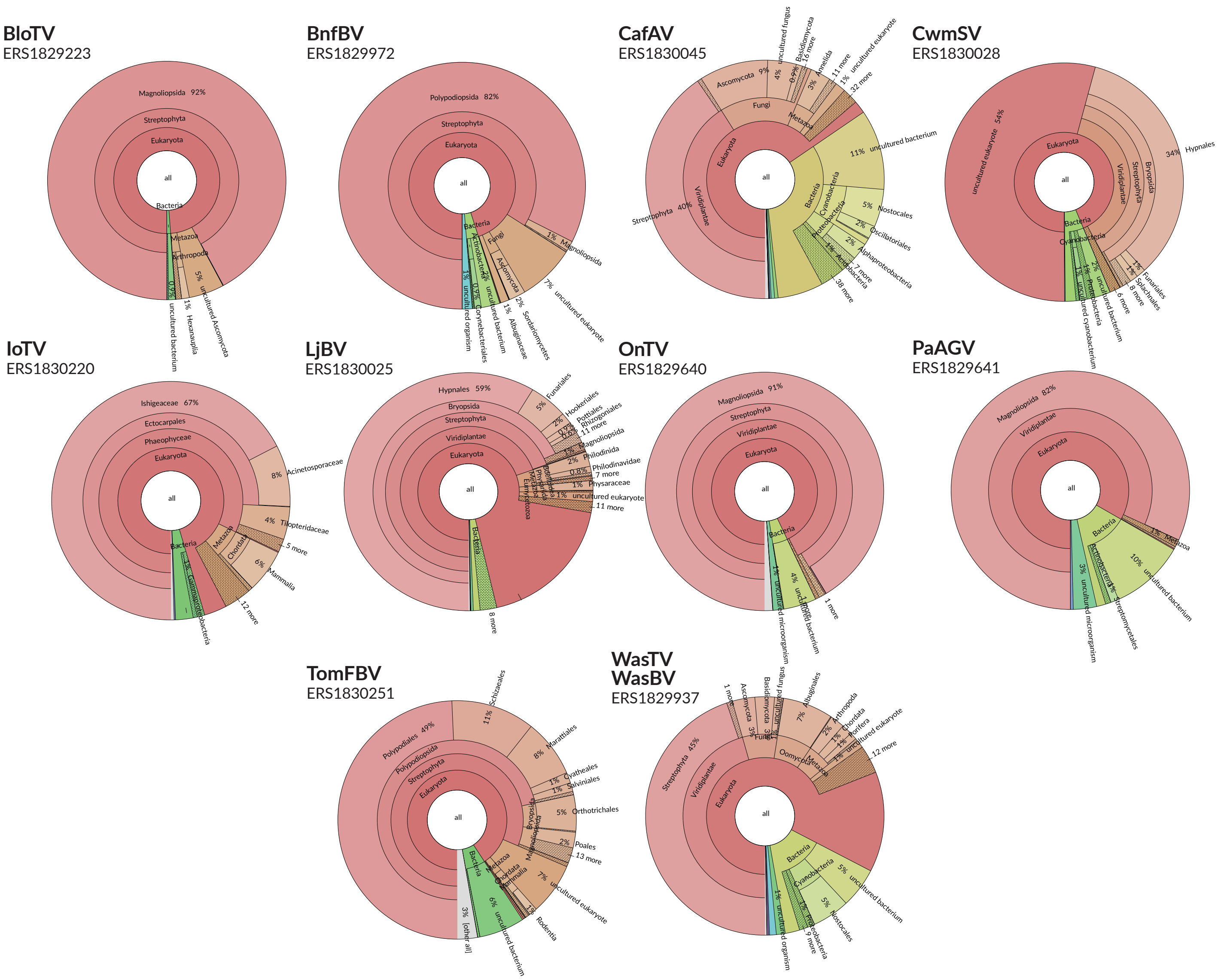
Supplementary Figure 1. Phylogram of virus composition across the One Thousand Plant Transcriptomes Initiative (1KP) samples. Virus abundance was summarised at the virus family level for each plant species and normalised using a Box-Cox transformation. The height of each bar represents the percentage of virus reads detected in each plant species. Plant clades are labelled and differentiated by shades of grey. The 1KP ASTRAL tree was used as the basis for this tree (1). Clade and abundance annotations were added using the Interactive Tree of Life (iTOL) web-based tool (2).


NP_062883.2_Rubella virus
100
NP_056779.1_Hepatitis E virus
NP_835282.1_Indian peanut clump virus
100
NP_604481.1_Barley stripe mosaic virus
100
AJF45323.1_Potato mop-top virus
100
100
YP_009220367.1_Colombian potato soil-borne virus
100
YP_009109559.1_Macrophomina phaseolina tobamo-like virus
YP_009552459.1_Nephila clavipes virus 3
AQT03616_Beet necrotic yellow vein virus 2
100
AAF91435_Beet soil-borne mosaic virus
96
92
YP_008219063 Burdock mottle virus
89
YP_009507937_Rice stripe necrosis virus
AMQ23297_Mangifera indica latent virus
100
Tomato fern benyvirus
AYD38100_Wheat stripe mosaic virus
78
98
Bird's-nest fern benyvirus
AYO99569.1_Goji berry chlorosis virus
97
AVD98637.1_Red clover RNA virus 1
KX883789_Hubei Beny-like virus
100
QED42955.1_Hubei Beny-like virus 1
88
QDH86482.1_Riboviria sp.
91
Leucodon julaceus virus
95
QIT20101.1_Diabrotica undecimpunctata virus 2
88
Wallace's spikemoss benyvirus
QDH87951.1_Riboviria sp.
QDH90217.1_Riboviria sp.
97
QDH87950.1_Riboviria sp.
BAS04359.1_Lentinula edodes ssRNA mycovirus
72
100
AQM49941.1_Agaricus bisporus virus 13
QDW81308_Rhizoctonia solani beny-like virus
89
100
100
AZF86092_Sclerotium rolfsii beny-like virus
QED42939.1_Monilinia benyvirus C
AEJ33768.1_Chara australis virus
0.4
Matonaviridae
Tymovirales
Hepeviridae
Virgaviridae
Benyviridae
Unclassfied
● Found in this study
🟩 Land plants
🟧 Lower plants
🟥 Invertebrate
▨ Vertebrate
🟪 Fungi
🟦 Algae
🟨 Environmental

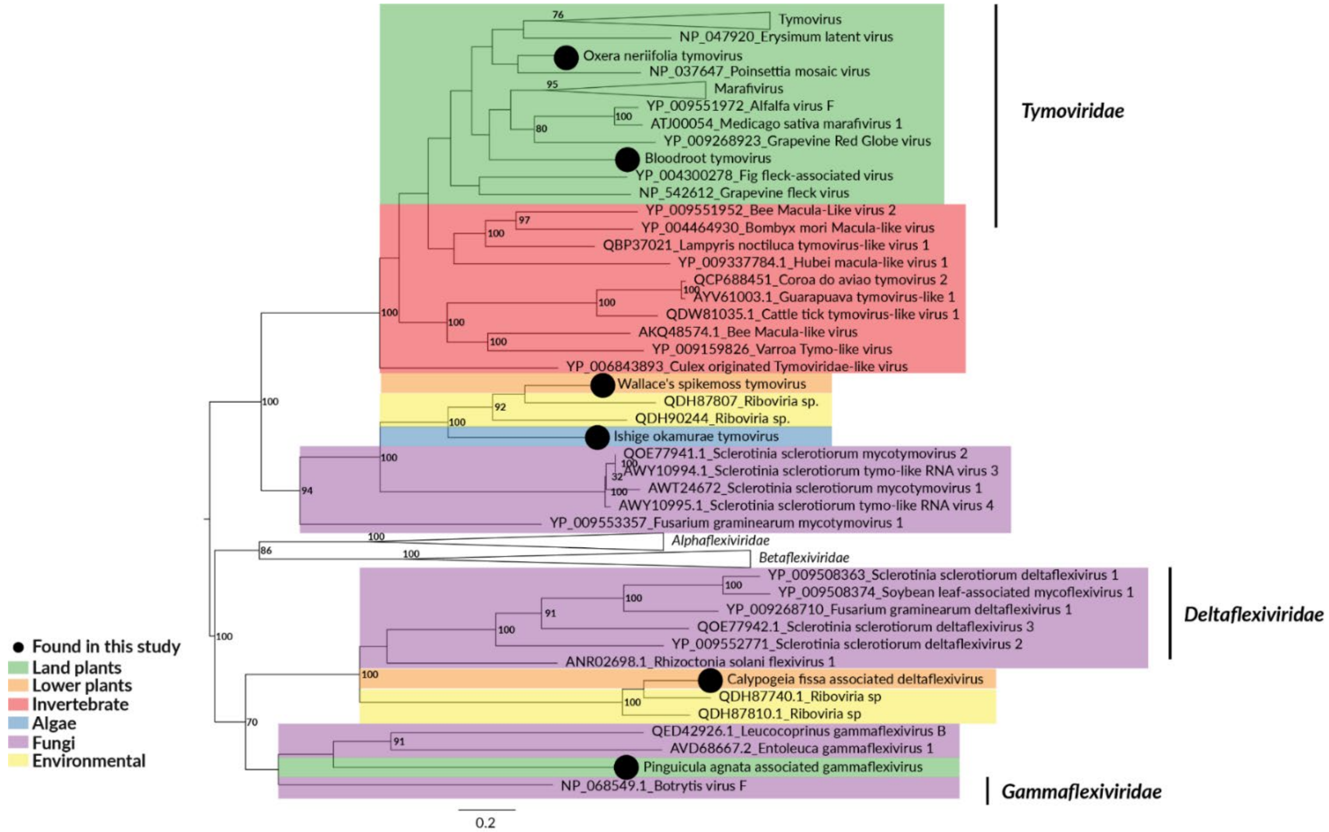
Supplementary Figure 2. Phylogram of the beny-like virus assembled in this study. Maximum likelihood phylogenetic trees show the topological position of the newly discovered virus-like sequences (black circles) in the context of their closest relatives. Branches are highlighted to represent host clade (land plants = green, lower plants = orange, invertebrate = red, vertebrate = pink, algae = blue, fungi = purple, yellow = environmental). Here "Land plants" encompasses both angiosperms and gymnosperms while "Lower plants" includes the bryophytes, lycophytes, and ferns. All branches are scaled to the number of amino acid substitutions per site and trees were mid-point rooted for clarity only. Numbers at the nodes indicate bootstrap support over 70% (1000 replicates).


QBQ82431.1_Lily virus X
100
YP_001655011.1_Phaius virus X
Blue agave alphaflexivirus
YP_009389420.1_Euonymus yellow vein virus
AOO32377.1_Citrus yellow vein clearing virus
QCY52825.1_Euonymus_yellow_mottle associated virus
ANH10861.1_Opuntia virus X
YP_001718500.1_Lolium latent virus
Alphaflexiviridae
AVD50415.1_Peach_chlorotic mottle virus
AVD50423.1_Cherry green ring mottle virus
ALT57215.1_Asian prunus virus 3
ARO69982.1_Grapevine rupestris stem pitting-associated virus
99
YP_009552758.1_Panax ginseng flexivirus 1
Betaflexiviridae
100
ACR24264.1_Pepino mosaic virus
83
AGO03687.1_Actinidia virus X
P16650.1_Potato virus S
NP_612617.1_Beet_necrotic yellow vein virus
100
YP_009513203.1_Beet soil-borne mosaic virus
77
YP_008219066.1_Burdock mottle virus
78
QAR21015.1_Rice stripe necrosis virus
Benyviridae
77
YP_009553317.1_Wheat stripe mosaic virus
Tomato fern benyvirus
AJF45333.1_Potato mop-top virus
100
100
YP_009220371.1_Colombian potato soil-borne virus
YP_009665976.1_Drakaea virus A
Virgaviridae
NP_620030.1_Peanut clump virus
ATO98266.1_Lychnis ringspot virus
0.3

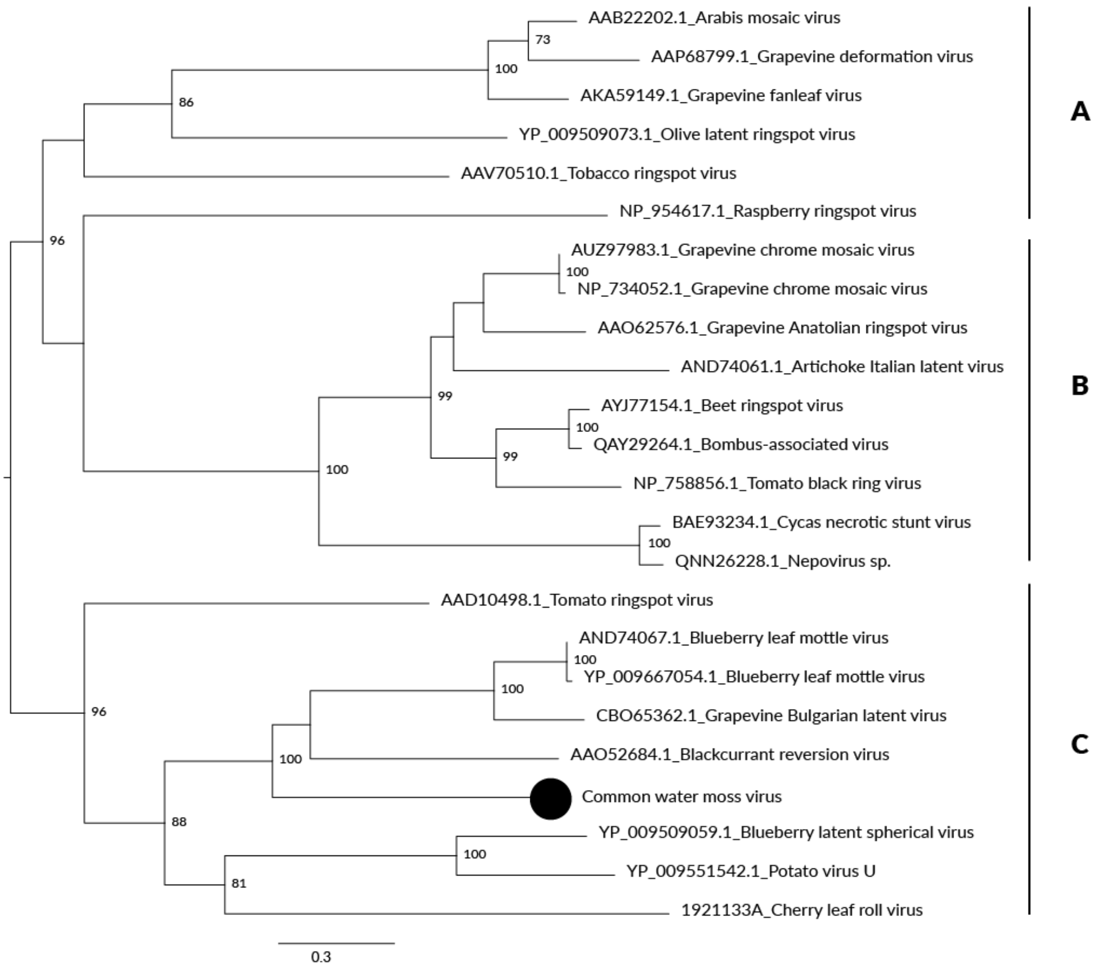
Supplementary Figure 3. Phylogram of the triple gene block (TGB) protein 1. Maximum likelihood phylogenetic trees show the topological position of the newly discovered TGB sequence in the tomato fern (black circle) in the context of the closest relatives. All branches are scaled to the number of amino acid substitutions per site and trees were mid-point rooted for clarity only. Numbers at the nodes indicate bootstrap support over 70% (1000 replicates).



Supplementary Figure 4. Taxonomic assignments of reads in select One Thousand Plant Transcriptomes Initiative (1KP) libraries. Each Krona graph illustrates the relative abundance of taxa in a metatranscriptome at varying taxonomic levels. For clarity, a max depth of five taxonomic levels was chosen for each graph. The library Sequence Read Archive accession number and the corresponding virus of interest are annotated above each graph. Reads without any match in the nt database are not shown. Krona graphs were created using the KMA and CCMetagen methods (3, 4).


76
Tymovirus
NP_047920_Erysimum latent virus
Oxera neriifolia tymovirus
NP_037647_Poinsettia mosaic virus
95
Marafivirus
YP_009551972_Alfalfa virus F
100
ATJ00054_Medicago sativa marafivirus 1
80
YP_009268923_Grapevine Red Globe virus
Bloodroot tymovirus
YP_004300278_Fig fleck-associated virus
NP_542612_Grapevine fleck virus
Tymoviridae
YP_009551952_Bee Macula-Like virus 2
97
YP_004464930_Bombyx mori Macula-like virus
100
QBP37021_Lampyris noctiluca tymovirus-like virus 1
YP_009337784.1_Hubei macula-like virus 1
QCP688451_Coroa do aviao tymovirus 2
100
AYV61003.1_Guarapuava tymovirus-like 1
100
100
QDW81035.1_Cattle tick tymovirus-like virus 1
100
AKQ48574.1_Bee Macula-like virus
100
YP_009159826_Varroa Tymo-like virus
YP_006843893_Culex originated Tymoviridae-like virus
Wallace's spikemoss tymovirus
100
QDH87807_Riboviria sp.
92
QDH90244_Riboviria sp.
100
Ishige okamurae tymovirus
QOE77941.1_Sclerotinia sclerotiorum mycotymovirus 2
100
100
AWY10994.1_Sclerotinia sclerotiorum tymo-like RNA virus 3
32
94
AWT24672_Sclerotinia sclerotiorum mycotymovirus 1
100
AWY10995.1_Sclerotinia sclerotiorum tymo-like RNA virus 4
YP_009553357_Fusarium graminearum mycotymovirus 1
100
Alphaflexiviridae
86
100
Betaflexiviridae
YP_009508363_Sclerotinia sclerotiorum deltaflexivirus 1
100
YP_009508374_Soybean leaf-associated mycoflexivirus 1
100
YP_009268710_Fusarium graminearum deltaflexivirus 1
91
QOE77942.1_Sclerotinia sclerotiorum deltaflexivirus 3
100
Deltaflexiviridae
YP_009552771_Sclerotinia sclerotiorum deltaflexivirus 2
100
ANR02698.1_Rhizoctonia solani flexivirus 1
100
Calypogeia fissa associated deltaflexivirus
QDH87740.1_Riboviria sp
100
QDH87810.1_Riboviria sp
70
QED42926.1_Leucocoprinus gammaflexivirus B
91
AVD68667.2_Entoleuca gammaflexivirus 1
Pinguicula agnata associated gammaflexivirus
NP_068549.1_Botrytis virus F
Gammaflexiviridae
0.2
● Found in this study
Land plants
Lower plants
Invertebrate
Algae
Fungi
Environmental

Supplementary Figure 5. Phylogram of the various tymo-like viruses isolated in this study. Maximum likelihood phylogenetic trees show the topological position of newly discovered tymovirus (black circles) sequences in the context of the closest relatives. All branches are scaled to the number of amino acid substitutions per site and trees were mid-point rooted for clarity only. Branches are highlighted to represent host clade (land plants = green, lower plants = orange, invertebrate = red, vertebrate = pink, algae = blue, fungi = purple, yellow = environmental).


AAB22202.1_Arabis mosaic virus
73
AAP68799.1_Grapevine deformation virus
100
AKA59149.1_Grapevine fanleaf virus
86
YP_009509073.1_Olive latent ringspot virus
AAV70510.1_Tobacco ringspot virus
NP_954617.1_Raspberry ringspot virus
96
AUZ97983.1_Grapevine chrome mosaic virus
100
NP_734052.1_Grapevine chrome mosaic virus
AAO62576.1_Grapevine Anatolian ringspot virus
AND74061.1_Artichoke Italian latent virus
99
AYJ77154.1_Beet ringspot virus
100
QAY29264.1_Bombus-associated virus
99
NP_758856.1_Tomato black ring virus
100
BAE93234.1_Cycas necrotic stunt virus
100
QNN26228.1_Nepovirus sp.
AAD10498.1_Tomato ringspot virus
AND74067.1_Blueberry leaf mottle virus
100
YP_009667054.1_Blueberry leaf mottle virus
100
CBO65362.1_Grapevine Bulgarian latent virus
96
AAO52684.1_Blackcurrant reversion virus
100
Common water moss virus
88
YP_009509059.1_Blueberry latent spherical virus
100
YP_009551542.1_Potato virus U
81
1921133A_Cherry leaf roll virus
0.3
A
B
C

Supplementary Figure 6. Phylogram of the common water moss virus coat protein (CP). Maximum likelihood phylogenetic trees show the topological position of the newly discovered CP sequence in the common water moss (black circle) in the context of the closest relatives. All branches are scaled to the number of amino acid substitutions per site and trees were mid-point rooted for clarity only. Nepovirus subgroups A, B and C are indicated. Numbers at the nodes indicate bootstrap support over 70% (1000 replicates).

Online supplementary

The following are available online at: https://cloudstor.aarnet.edu.au/plus/s/s5uTDqvKfJu8TxZ
SI Table 1: Botanical definitions of each trait state, SI Table 2: Clade assignment for all One Thousand Plant Transcriptomes Initiative (1KP) species for which a virus was detected, SI Table 3: Summary information for each One Thousand Plant Transcriptomes Initiative (1KP) libraries analysed. SI Table 4: Proportion of transcripts and abundance assigned to each plant virus family. SI Table 5: Name and accession IDs of the virus sequences used in each phylogenetic tree. SI Table 6: List of Sequence Read Archive (SRA) IDs associated with each virus assembled in this study. SI Table 7: Summary of the number of families, species and libraries that make up each host clade. SI Document 1. Commentary article: Examining the diversity of the phytovirosphere. High-definition images of all plots and trees are available in SI Manuscript plots.

References

1. Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, et al. One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 2019;574(7780):679-85.
2. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47(W1):W256-W9.
3. Clausen PT, Aarestrup FM, Lund O. Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics. 2018;19(1):1-8.
4. Marcelino VR, Clausen PTLC, Buchmann JP, Wille M, Iredell JR, Meyer W, et al. CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. Genome Biol. 2020;21(1):103.