Genome evolution, taxonomy, and transmission of potexviruses in cacti (*Alphaflexiviridae*)

Alexa Tyszka[1,†], Karolis Ramanauskas[1], and Boris Igić[1]

[1]**Department of Biological Sciences, University of Illinois at Chicago, 840 West Taylor St. MC067, Chicago, IL 60607, United States of America**
[†]**Author for correspondence.**

ABSTRACT

Potexvirus is a group of positive-sense single-stranded RNA viruses known to infect many flowering plants, including cacti (Cactaceae). The current viral taxonomic naming schemes in this group often employ informal or outdated host plant names (synonyms), which complicate systematic study. One such group, often named with a suffix "Virus X," presents a further complication—nearly all of its published sequences are from infections of cultivated plants, in which infections may dramatically affect yield. Because their host-specificity is broad, the source of infections, the natural distribution of this group, and the significance of infections in wild species of cacti all remain unclear. The lack of clarity is partly related to low sampling across the Potexviruses that infect cacti. And yet, the availability of sampled plant transcriptomes, all of which are practically metatranscriptomes, has recently exploded, along with the decreasing expense and difficulty of conducting RNAseq experiments. Here, we harness these new tools and perform phylogenetic analyses aimed at clarifying taxonomic diversity, quantifying patterns of tissue expression, diversity, and examining selective pressures across viral genomes. The results suggest a novel mode of transmission by sex (pollination) for this viral group, based on significant expression in pollen. We examine and discuss the implications of our key results for the taxonomy of *Potexviruses* that infect Cactaceae, noting their vastly understudied ecological significance.

INTRODUCTION

Molisch's (1885) discovery of "protein bodies" on several species of cacti was one of the first documented descriptions of viruses. For nearly a century, subsequent comparative study of viruses remained limited to direct observational data of gross morphology, augmented with clever experimental approaches, such as filtration and inoculation (Mettenleiter, 2017). A transformative advancement in virology—and all of biology—has been the advent of massively parallel DNA and RNA sequencing. The rapidly improving sequencing tools enable rapid identification of organisms from seemingly any sampled surface of the Earth. One common thread is that virtually every macro-organism genome study uncovers a micro-organismal metagenome, composed of both targeted host sequences and those from myriad co-existing organisms. Metagenomic studies have yielded an enormous number of genomes and have vastly expanded the global viriome (Gregory et al., 2019; Lefeuvre et al., 2019; Shi et al., 2016). The unprecedented amount of data resulting from metagenomic studies has also caused significant policy changes and revisions by the International Committee on Taxonomy of Viruses (ICTV) policy (International Committee on Taxonomy of Viruses Executive Committee, 2020; Simmonds et al., 2017), but nearly all viruses remain named by their original description of host, location, and/or symptoms.

Historic naming conventions are ill-suited for host plants whose own taxonomic placement is uncertain, which has been particularly true for rapidly diversified groups such as Cactaceae (cacti). Molisch's "protein bodies" are now widely understood to be comprised of plant-infecting potexviruses (*Tymovirales*, family *Alphaflexiviridae*). Their positive-sense, single-stranded RNA genomes consist of 5.9-7.0 kb of positive-sense single-stranded RNA (Martelli et al., 2007). Generally presenting as elongated, rod-shaped filamentous viruses, they express five primary open reading frames (ORFs): Replicase (Rep), Triple gene block (TGB), Coat protein (CP), coded in the 5' direction as well as two smaller overlapping ORFs

coded in the 3' direction: ORF6 and ORF7 (Martelli et al., 2007). Members of this group produce variably symptomatic infections in cacti, and many infected plants show no external signs of viral infection (Bos, 1977; Liou et al., 2004). Neither the significance of their infections in nature, nor relative modes of transmission are clear. Reports of symptomatic plants range from 0%-5.5% in wild species in the southwestern United States (Attathom et al., 1978) to 44% in agricultural fields on Hainan Island, China (Peng et al., 2016). The most commonly recognized symptoms of the disease are mosaic, mottling, stunted growth, and distortion (Attathom et al., 1978; Maliarenko and Mudrak, 2013; Peng et al., 2016). Infection through grafting and mechanical contact, particularly following stem injury and human-mediated or hemipteran insect-mediated sap inoculation, is well-documented (Liou et al., 2004; Maliarenko and Mudrak, 2013; Park et al., 2018). Grafting is a primary means of propagation among crop cacti (Park et al., 2018), and *Selenicereus* is a commonly chosen graft stock. However, there are reports of other members within the family *Alphaflexiviridae* transmitting via insect and seed vectors (Martelli et al., 2007), and pre-DNA studies tentatively suggest that in the wild, pollen may transmit CVX (Attathom et al., 1978).

Viral taxonomy is complicated by many aspects of biology and taxonomic practices. *Schlumbergera truncata* (Haworth) Moran has undergone a number of name changes, including *Epiphyllum truncatum* Haworth in 1819, *Cactus truncatus* (Haworth) Link in 1822, and *Zygocactus truncatus* (Haworth) K. Schumann in 1890, dramatically confusing subsequent viral taxonomy. Thus, currently accepted names in the *Potexvirus* group include *Cactus Virus X* (CVX), *Zygocactus Virus X* (ZyVX), and *Schlumbergera Virus X* (SchVX), each of which was likely characterized on the same host genus (and possibly species). The Baltimore classification system standardizes viral classification by intrinsic morphological characteristics of a virus' replication machinery. It has been integrated into the ICTV guidelines to better reflect viral evolutionary relationships (International Committee on Taxonomy of Viruses Executive Committee, 2020). The term "plant virus" in itself is problematic since there is strong evidence to suggest that many viruses have transitioned from fungal or invertebrate hosts to plant hosts (Lefeuvre et al., 2019). Additionally, many plant viruses that infect agriculturally important species are named using the common name of a plant, which carries its own problems, for example: *Pitaya Virus X* is named for the common name "Pitaya" which can refer to as many as thirty-one species within the genus *Selenicereus* (Korotkova et al., 2017; Guerrero et al., 2019; Le Bellec and Vaillant, 2011). The matter is further complicated by basic viral ecology, because one virus may infect many hosts, and one host may be co-infected by many viruses. Single-stranded RNA viruses have faster rates of evolution than their host plants. There is no guarantee that viral evolution and speciation follow linearly behind plant evolution and speciation—especially due to viral host-switching. These problems persist throughout the genus *Potexvirus* and are especially prominent in cactus-infecting *Potexvirus* species. We suggest a phylogeny-based approach to remedy some prominent taxonomic issues within this specific clade.

Knowledge about cactus-infecting *Potexviruses* contributes to a growing yet biased study of plant viruses. Human-assisted dispersal, grafting, and cultivation obscure the evolutionary history of these viruses, which parallels the disproportionate sampling representation of plants raised in greenhouses or for agricultural production. However, *Cactus Virus X* and associated viruses seem restricted to cactaceous hosts for unknown reasons—every sample of CVX or CVX-related viruses has come from cacti. The few studies that have investigated wild *Potexviruses* of cacti predate DNA methods and have yet to identify the origin. Recent sequencing efforts have revealed multiple inconsistent virus-host pairs on cacti. Although many metagenomic studies capture environmental, genetic information that allows for virus identification, tissue type may bias expression rates of viruses (Lacroix et al., 2016). The pursuit of wild cactus-infecting *Potexviruses* expands our evolutionary knowledge of viral evolution, host selection, and transmission mechanics. The relationships of the virus can be investigated with a thorough phylogenetic approach, using available virus samples. In this study we present the largest to date phylogeny of cactus-infecting *Potexviruses*. We attempt to use this expanded phylogeny to answer relevant questions about Potexvirus evolutionary relationships and revisit the utility of decades-old taxonomy in current virus research.

MATERIALS AND METHODS

Host Study Species and Sampling

We relied on two types of sequencing data for all analyses: original sequences obtained from tissues we collected and sequences deposited in public sequence data archives. We recovered original viral sequence data from tissues of *Schlumbergera truncata* (Haworth) Moran, commonly known as "crab cactus" or "false Christmas cactus," a widely cultivated species. Although there are dozens of named varieties of

this species, nearly all commercially grown plants are of uncertain provenance. They almost certainly trace to a handful of plants collected in their native Atlantic forests of Brazil and brought to England in the early 1800s (Boyle, 2003). Plants are easily grown from cuttings and the species has been extensively hybridized across Western Europe and exported across the world, prized for their showy winter (short-day) displays.

Our host plant samples were sourced from a haphazardly collected personal collection (B.I.), purchased or found abandoned around the city of Chicago. Most of the plants were either apparently asymptomatic or weakly symptomatic at the time of tissue collection. All of our accessioned host plants are independent genets (unique genotypes) (Ramanauskas and Igić, 2021).

We searched the NCBI Sequence Read Archive (SRA) database (www.ncbi.nlm.nih.gov/sra) for RNA-sequencing (RNA-seq) data within the flowering plant order Caryophyllales (NCBI:txid3524) that had been sequenced using an Illumina library sequencing platform. For each identified SRA run accession (SRR), viral RNA that matched sample cactus-infecting Potexvirus RNA (accession numbers provided in Supplemental Information) was identified, extracted, and assembled using the kakapo 0.7.3-dev pipeline (http://flightless.one) with Kraken2 viral filters disabled. The search returned 59 sequences aligned to members of Potexvirus within PRJNA608981 (https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA608981).

Additional publicly available partial or complete viral genomes, gene annotations, and available metadata including host information from the genera Potexviruses (NCBI:txid12176) were downloaded from the NCBI genome browser (NCBI: https://www.ncbi.nlm.nih.gov/genome/) (Table S2). These genomes will be referred to as "GenBank" genomes.

RNA Sequencing

Pistils (without ovaries), pollen, leaf, and root tissues were removed and submerged in 1.5 ml of *RNAlater*^{TM} solution (Invitrogen). Samples were held at room temperature for 30-60 minutes and then moved to a -80°C freezer for storage. Approximately 100 mg of tissue was ground to a fine powder in 1.5 ml tubes submerged in liquid nitrogen. Total RNA was isolated using Total RNA Mini Kit (Plant kit; IBI Scientific, Cat. No. IB47341) following manufacturer's instructions. We assessed RNA concentration and purity with a NanoDrop^{TM} Lite Spectrophotometer (Thermo Scientific). The twenty three samples used in this study were sequenced as part of a larger sequencing effort which consisted of four total separate sequencing runs and included additional samples from other plant species.

Sequencing libraries were prepared using KAPA Stranded mRNA-Seq (Roche). These libraries were sequenced on a single lane of Illumina HiSeq 4000 or Illumina NovaSeq 6000 platform (paired-end 150 bp reads) at the Duke University Center for Genomic and Computational Biology. The number of resulting read pairs (for the twenty-three samples presented here) ranged from 4,148,932 to 9,618,084 with a median of 6,363,556 and average of 6,293,553 (Table S1).

RNAseq Assemblies

Raw paired-end Illumina reads were first processed using Rcorrector v1.0.4 (Song and Florea, 2015) to infer and correct sequencing errors. Reads were next trimmed with Trimmomatic v0.39 (Bolger et al., 2014) to remove any read containing bases with Phred scores lower than 20, low quality reads less than 50 bp long, and any adapter or other Illumina-specific sequences that were still present. The remaining reads were filtered with Kraken 2 (Wood et al., 2019) to remove small and large subunit ribosomal RNA (using the SILVA database; Quast et al. 2013) and contaminating reads (minikraken2_v2 database). We used custom-built databases, derived from RefSeq libraries: UniVec_Core, viral, mitochondrion, plastid, plasmid, archaea, bacteria, protozoa, human, and fungi to minimize the number of contaminating and non-nuclear reads (Ramanauskas and Igić, 2021). Filtered reads were combined across all samples into a single RNA-seq data set including *S. truncata* and *CVX* RNA.

We conducted a *de novo* transcriptome assembly to assemble *S. truncata* and GenBank accessed RNA-seq data to reference genomes NC_002815, NC_006059, NC_011659, and NC_024458 (Table X, Table S3).

Sequence Alignment and Phylogenetic Analyses

The untranslated regions (UTRs) were trimmed from the sequences for consistency. Sequence alignments were performed through MAFFT v7.490 (Katoh, 2002) using the full dataset of RNA sequences and automatic strategy detection. Each aligned sequence was annotated using the Geneious annotationR11 11.0.5 (https://www.geneious.com). The aligned sequences were divided by ORF using annotations to

produce sequence alignments for each of the five genes, along with the whole-genome alignment. The individual proteins were exported to FASTA files, then gaps at the start of the sequence and stop codons were removed manually.

Phylogenetic relationships, including those used for assessing bootstrap support, were inferred using `IQ-Tree` v2.0.3. Maximum likelihood inference for the whole genome sequences—as well as for each gene region, separately—relied on a model of sequence evolution (GTR+F+I+Γ4) favored by both AIC- and BIC-based selection procedure implemented in IQ-Tree's model selection module *ModelFinder* (Kalyaanamoorthy et al., 2017). Akaike and Bayesian weights exceeded 0.99. Branch support was assessed with IQ-Tree's *UFBoot*, an ultrafast bootstrap implementation (Hoang et al., 2018).

Species Delimitation Metrics

ICTV guidelines state that species within *Potexvirus* are delineated by 72% shared nucleotide identity, or 80% shared amino acid identity within the coat protein or replication genes (ICTV, 2022). Raw pairwise distance calculation was conducted on gene sequence alignments in R using `ape` v5.5.

Automated delimitation was also preformed using mPTP (Kapli et al. 2017; http://mptp.h-its.org/#/tree) and bPTP servers (Zhang et al. 2013; http://species.h-its.org/ptp/). bPTP was run using 100,000 MCMC generations and 0.1 burn-in. Outgroups were removed for both delimitation analyses. Gene trees were compared to the full genome sequences manually under the Phylogenetic Species Concept, based on previously named species genomes with maximum clade inclusivity. Gene to genome relationships were also compared in *R* using the function *cophylo* from phytools v 2.0.3.

Detection and Estimation of Molecular Selection

The strength and direction of selection pressure across genomes—measured with a relative ratio of silent and protein-altering mutations per available site—may vary. We estimated molecular selection with the Fast, Unconstrained Bayesian AppRoximation (FUBAR) method (Murrell et al., 2013), which uses a Bayesian approach to infer nonsynonymous (dN or beta) and synonymous (dS or alpha) substitution rates on a per-site basis for a given coding alignment and corresponding phylogeny. FUBAR reports evidence for positive selection using posterior probabilities (which range 0 to 1). Posterior probabilities greater than 0.9 are generally considered to be strongly suggestive of positive selection (Murrell et al., 2013). The method makes an important assumption that the selection pressure for each site is constant along the entire phylogeny.

Data Accessibility

This article contains a Supplementary Information Appendix containing Supplemental Tables S1–S3 and Supplemental Figures S1–S5. All sequence data associated with *S. truncata* is deposited in GenBank within project accession number PRJNA705387 (https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA705387). Scripts and data are accessible at github.com/alexatyszka/cactusvirusx.

RESULTS

Sequence Assembly and Approach

In an attempt to characterize the infection patterns of cactus-infecting potexviruses, we assembled 83 viral sequences from the cactus samples analyzed. 24 of these sequences were from *S. truncata* samples, and 59 of the sequences were from *Hylocereus (now Selenicereus) spp.* in PRJNA608981 (https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA608981)(Fan et al., 2020). Genome sizes of 7 kb within newly assembled sequences were consistent with previously reported 5.5-9.0 kb genome lengths within *Alphaflexiviridae* (Kreuze et al., 2020; ICTV, 2022)(Table S2). We add these 83 sequences to an existing database of 38 related *Potexvirus* samples Table S2) and demonstrate the utility of the *kakapo* pipeline within a phylogenetic and transcriptomic workflow. Our assembly recovered viral sequences with high coverage, representing millions of viral reads for each sample Table S1.

Viral Detection

Assembly of a virus or its proteins from a host plant metatranscriptome represents presence of the virus within host tissues. In externally sequenced samples, none of the hosts had been noted as symptomatic (Fan et al., 2020). However, we broadly recovered diverse potexviruses from many (of plant samples within the project. Within newly-sequenced Schlumbergera samples, viruses were generally found in high amounts on pollen and style tissue.

Diversity and Phylogeny

We offer multiple metrics in an attempt to guide viral identification and placement within the group. These metrics include phylogeny-based grouping, automated phylogenetic delimitation, sequence similarity, gene-specific comparisons, host-based delimitation, or a combination of these. We will briefly review our findings from each. A well-supported phylogenetic tree was recovered using available sequences (Figure 1). Newly assembled sequences from this study nest closely with previously described viruses on well-supported distinct branches. These sequences greatly expand the cactus-infecting clade of potexviruses and nearly triple the amount of available sequences within the clade.

Comparing Gene to Genome

Each phylogeny, from whole genome to each of the five genes, recovered currently delimited viruses together in monophyletic clades. Gene phylogenies did not recover different topologies when comparing only monophyletic named groups (Figure S1 through Figure S5). Gene to genome and gene to gene tree topologies were largely similar (Figure S13).

Automated Delimitation

Five named species currently define cactus-infecting potexviruses. Two automated species delimitation methods, mPTP and bPTP (Kapli et al. 2017; http://mptp.h-its.org/#/tree; Zhang et al. 2013; http://species.h-its.org/ptp/), delimited 11 and 16 species respectively when given the same 94-tip full sequence tree Figure 1, Figure S6, Figure S7. The two delimitation methods agreed on species delimitations in all but one case, where one mPTP-delimited species consisting of newly assembled sequences was divided into 7 separate species by bPTP (Figure S8, Figure S9). The divided clade included all but one sample from Schlumbergera_truncata_15H03 and Schlumbergera_truncata_15H02.

Sequence Similarity

Sequence similarity is an alternative method that has been used to delimit viral species. Due to the two-dimensional nature of sequence similarity data, we present a matrix heatmap of sequence similarity emphasizing the 72 percent delimitation cutoff often used for potexviruses. We also display heatmaps for two genes, RdRp and coat protein, as well as phylogenetic trees for each of the five genes composing the viral genome.

DISCUSSION

We use an effective, pipeline-based assembly approach to contribute 83 new potexvirus sequences to the literature. Novel viral discovery has implications for plant reproduction and immune defense and is vital for agriculture. We found ample viral reads in both stigma and style of *Schlumbergera sp.*, which could represent sexual transmission of the virus. We find regions of increased selective pressure within viral genes, although viral genome structure and reverse coding regions may render these measures inefficient. A phylogenetic analysis found that existing viral clades are monophyletic; the new viral sequences were placed on extremely short terminal branches in a clade with one or more previously named viruses. Our results imply an expansion of presently known viral species as well as an expansion of host ranges for some species. Host-based species delimitation has been inefficient in the face of mixed viral infections (Li et al. 2015), so we also present species delimitation from a phylogenetic species concept; sequence similarity is a related measure but did not return identical results. Analysis of these new potexvirus sequences has elucidated the disagreements between different species concepts when it comes to viruses in general, and we present a small viral clade as a case study for viral species delimitation efforts.

Reproductive and Immune Implications

Viral infections are known to spread to nearly all tissues within a plant (Hipper et al. 2013). However, little is known about cross-tissue infections, or the full extent of viral infection a plant may endure, which could vary by life stage or species. Plants possess some defense mechanisms to prevent viral spread through tissues, most notable RNAi gene silencing (Reviewed in Hipper et al. 2013). Plant immune responses are dependent on viral recognition, which may impose selective pressure differentially on viral genes depending on their function. Most notably, the Coat Protein gene may be under selective pressure due to the fitness advantages for avoiding plant immune response. We find some evidence of selective pressure

across the viral genome Figure S10, although we cannot discern whether the higher dN/dS values within overlapping regions is due to selection.

Plant reproductive tissue is susceptible to viral infection in at least some cases, the most famous perhaps being tulips (*Tulipa*) displaying different floral coloration due to Tulip breaking virus. The presence of virus in reproductive tissue implies that sexual transmission of a virus may be possible (Kim et al. 2015). We recover viruses from pollen and style within samples of *Schlumbergera truncata* Table S1, Table S4, which is the first reported instance of viral reads on reproductive tissue of this species. The viruses recovered from *Schlumbergera truncata* phylogenetically are similar to the taxa*Cactus virus X*, which may represent an avenue for*Cactus virus X* to be transmitted sexually from plant to plant. Further, we can confirm that the *Schlumbergera truncata* plant samples sequenced as a part of this study were all housed in the same location and were frequently the subjects of pollination experiments (Ramanauskas and Igić 2021), making it likely that infected pollen was transmitted from plant to plant. The manuscript describing the plants sampled from SRR samples did not mention symptoms of viral infection, although viral infections are often asymptomatic. A viral infection has the potential to cause stress to a plant, and infections can spread quickly through contact with equipment, which could bias gene expression levels or other measurements collected during the course of study. Although viral contaminants can be filtered out of RNA-seq data, we caution that undetected viral infections could potentially bias data in unexpected ways.

Host-based Delimitation

We approached the problem of species delimitation with a variety of methods. One common rudimentary approach to viral classification has been description firstly based on identified host species. However, this concept quickly loses usefulness in the face of reports of multiple infections within a single host plant (Li et al. 2015), or reports that a certain virus is not constrained to infecting a singular plant species. As more hosts are discovered for cactus-infecting potexviruses, the question shifts from *which* hosts a virus may infect to *why* precisely the virus may infect those hosts and not others, if exposed equally to many potential hosts. Further, potexviruses are perfectly able to infect certain phylogenetically distant hosts *ex situ*, such as *Chenopodium* (Pleše and Miličič 1966, Attathom et al. 1978) and *Nicotiana* (Casper and Brandes 1969). The discovery of novel hosts and novel viruses will surely continue, although more conclusive measures are needed to investigate viral host specificity. The problem is exacerbated when plant viruses do not recapitulate the evolutionary patterns of their host in a logical manner, such as the cactus-infecting potexvirus clade (Figure 4), where the formal name of a species often disagrees with the actual host range. Especially because plant genus names are prone to change, viral species names such as *Cactus Virus X* are not particularly informative. We advise more sampling of hosts, particularly wild host plants, as the true ranges of many cactus-infecting potexviruses are yet unknown. Our study represents a near-tripling of the amount of sequences available for this small potexvirus clade, from a relatively narrow sampling of plant transcriptomes.

Sequence Similarity Delimitation

We used a sequence similarity-based delimitation method to determine the percentage of similarity a sample shared with another. The ICTV suggests that potexviruses with more than 72% similarity between their RdRp or CP genes should be considered a species. The sequence similarity method becomes ambiguous when confronted with multiple sequences, which may share more or less similarity with an unrelated sequence where its sister taxa do not. We recovered cases where distinct species emerged according to one of the two ICTV guidelines (Figure 2, Figure 3), but guidelines are less clear for edge cases, where the RdRp delimitation may disagree with the CP delimitation, or vice versa. For short sequences, sequence similarity delimitation is an efficient delimitation method, but may suffer due to incidental biases of evolutionary convergence. Particularly in cases where multiple infections are present, RNA viruses could hypothetically also receive genes from distinct species. Sequence similarity delimitation based on ICTV guidelines rapidly becomes imprecise and impractical when considering more than a handful of clades.

Phylogenetic Delimitation

Using a phylogenetic species concept, we recover the five clades that have already been described (Figure 1). Newly assembled sequences nest squarely within and around monophyletic clades, which we have described using formally described species. Of note is a single genome from *Mytcor Virus 1* (MG210801), which is described as present on a bivalve host. We can only postulate about the placement of this virus, but

it is recovered within the putative *Pitaya virus X* clade. Multiple explanations exist for its discovery on a bivalve host, but perhaps the most likely is accidental human contamination during sampling or RNA extraction. *Pitaya Virus X* has been reported to infect *Selenicereus spp*, and all members of the putative *Pitaya Virus X* clade were reported on plants within the genus *Selenicereus* (Figure S11, Figure 4). *Selenicereus* is an important crop fruit, which lends credibility to the possibility of contamination.

We delimited the putative clades based on the full genome sequences and inclusion of a previously named viral sequence, and each clade was marked inclusion of a basal named species. No sequences fell between clades. The groups putative *Cactus virus X* and putative *Schlumbergera Virus X* were marked by longer branches splitting the group into two distinct subclades. Further discussion is needed as to whether these subclades necessitate distinct species, but we err on the side of previously established naming conventions for this study. Gene trees (Figure S1 through Figure S5) did not display markedly different clade-level topologies when compared (Figure S13). This may imply that the genes are inherited faithfully with regard to the full genome, although we acknowledge that longer ($\sim$5000bp) genes such as RdRp contribute more to the full genome than smaller genes.

CONCLUSION

A recent uptick in available transcriptome data has paved the way for metatranscriptomic research. We present a strategy for obtaining, mining, and processing viral sequence data from multiple sources using the kakapo pipeline. Our sources included plant samples sequenced by our group fora separate project, samples from a large sequencing project of a related group of cacti, and official genomes for species as confirmed by the ICTV. We placed new viral sequences, representing a nearly threefold increase for the small viral clade, on a phylogeny and found that the new sequences were closely related to previously reported potexviruses, representing multiple recoveries of the same or similar species. Close phylogenetic placement, coupled with low levels of topological discordance between genes, indicate that currently defined viral species adequately delimit viral diversity, with a few outlier cases which we present for further discussion (Table 1). We also present evidence that cactus-infecting potexviruses, specifically putative sequences of *Cactus virus X* found on *Schlumbergera truncata*, are present on reproductive tissue; we postulate that this may represent sexual transmission of the virus. The impact of viral infections on plants is not well-known for any group, and questions remain unanswered regarding the true distribution or infection dynamics for any given plant-host pairing, which might be ameliorated by broader sequencing of potential hosts.

ACKNOWLEDGMENTS

This work was supported by the Award for Graduate Research from the Graduate College, University of Illinois at Chicago (to K.R.) and the National Science Foundation grant NSF-DEB-1655692 (to B.I.).

REFERENCES

Attathom, S., Weathers, L. G., and Gumpf, D. J. (1978). Occurrence and distribution of a virus induced disease of barrel cactus in California. *Plant disease reporter*.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

Bos, L. (1977). *SYMPTOMS OF VIRUS DISEASES IN PLANTS*. Research Institute for Plant Protection.

Boyle, T. H. (2003). Identification of self-incompatibility groups in *Hatiora* and *Schlumbergera* (Cactaceae). *Sexual Plant Reproduction*, 16(3):151–155.

Casper, R. and Brandes, J. (1969). A New Cactus Virus. *J. gen. Virol.*, 5:155–156.

Fan, R., Sun, Q., Zeng, J., and Zhang, X. (2020). Retracted article: Contribution of anthocyanin pathways to fruit flesh coloration in pitayas. *BMC Plant Biology*, 20(1):1–12.

Gregory, A. C., Zayed, A. A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., Dimier, C., Domínguez-Huerta, G., Ferland, J., Kandels, S., Liu, Y., Marec, C., Pesant, S., Picheral, M., Pisarev, S., Poulain, J., Tremblay, J.-É., Vik, D., Acinas, S. G., Babin, M., Bork, P., Boss, E., Bowler, C., Cochrane, G., de Vargas, C., Follows, M., Gorsky, G., Grimsley, N., Guidi, L., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Karp-Boss, L., Karsenti, E., Not, F., Ogata, H., Pesant, S., Poulton, N., Raes, J., Sardet, C., Speich, S., Stemmann, L., Sullivan, M. B., Sunagawa, S., Wincker, P., Babin, M., Bowler, C., Culley, A. I., de Vargas, C., Dutilh, B. E., Iudicone, D., Karp-Boss, L., Roux, S., Sunagawa, S., Wincker, P., and Sullivan, M. B. (2019). Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell*, 177(5):1109–1123.e14.

Guerrero, P. C., Majure, L. C., Cornejo-Romero, A., and Hernández-Hernández, T. (2019). Phylogenetic Relationships and Evolutionary Trends in the Cactus Family. *Journal of Heredity*, 110(1):4–21.

Hipper, C., Brault, V., Ziegler-Graff, V., and Revers, F. (2013). Viral and cellular factors involved in phloem transport of plant viruses. *Frontiers in Plant Science*, 4.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*, 35(2):518–522.

ICTV (2022). The International Code of Virus Classification and Nomenclature.

International Committee on Taxonomy of Viruses Executive Committee (2020). The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nature Microbiology*, 5(5):668.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6):587–589.

Kapli, P., Lutteropp, S., Zhang, J., Kobert, K., Pavlidis, P., Stamatakis, A., and Flouri, T. (2017). Multi-rate poisson tree processes for single-locus species delimitation under maximum likelihood and markov chain monte carlo. *Bioinformatics*, page btx025.

Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066.

Kim, J., Kil, E., Kim, S., Seo, H., Byun, H., Park, J., Chung, M., Kwak, H., Kim, M., Kim, C., Yang, J., Lee, K., Choi, H., and Lee, S. (2015). Seed transmission of sweet potato leaf curl virus in sweet potato (ipomoea batatas). *Plant Pathology*, 64(6):1284–1291.

Korotkova, N., Borsch, T., and Arias, S. (2017). A phylogenetic framework for the Hylocereeae (Cactaceae) and implications for the circumscription of the genera. *Phytotaxa*, 327(1):1.

Kreuze, J. F., Vaira, A. M., Menzel, W., Candresse, T., Zavriev, S. K., Hammond, J., Hyun Ryu, K., and Report Consortium, I. (2020). ICTV Virus Taxonomy Profile: Alphaflexiviridae. *Journal of General Virology*, 101(7):699–700.

Lacroix, C., Renner, K., Cole, E., Seabloom, E. W., Borer, E. T., and Malmstrom, C. M. (2016). Methodological guidelines for accurate detection of viruses in wild plant species. *Applied and environmental microbiology*, 82(6):1966–1975.

Le Bellec, F. and Vaillant, F. (2011). Pitahaya (pitaya) (Hylocereus spp.). In *Postharvest Biology and Technology of Tropical and Subtropical Fruits*, pages 247–273e. Elsevier.

Lefeuvre, P., Martin, D. P., Elena, S. F., Shepherd, D. N., Roumagnac, P., and Varsani, A. (2019). Evolution and ecology of plant viruses. *Nature Reviews Microbiology*, 17(10):632–644.

Li, Y.-S., Mao, C.-H., Kuo, T.-Y., and Chang, Y.-C. (2015). VIRAL DISEASES OF PITAYA AND OTHER CACTACEAE PLANTS. *Improving Pitaya Production and Marketing*, page 9.

Liou, M. R., Chen, Y. R., and Liou, R. F. (2004). Complete nucleotide sequence and genome organization of a Cactus virus X strain from Hylocereus undatus (Cactaceae). *Archives of Virology*, 149(5):1037–1043.

Maliarenko, V. M. and Mudrak, T. P. (2013). Cactus viruses in fasciated plants. *Biologija*, 59(2).

Martelli, G. P., Adams, M. J., Kreuze, J. F., and Dolja, V. V. (2007). Family *Flexiviridae* : A Case Study in Virion and Genome Plasticity. *Annual Review of Phytopathology*, 45(1):73–100.

Mettenleiter, T. C. (2017). The First "Virus Hunters". *Advances in Virus Research*, 99:1–16.

Molisch, H. (1885). Über merkwürdige geformte Proteinkörper in den Zweigen von *Epiphylium*. *Berichte der Deutschen botanischen Gesellschaft in Berlin*, 3:195–202.

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., and Scheffler, K. (2013). Fubar: a fast, unconstrained bayesian approximation for inferring selection. *Molecular Biology and Evolution*, 30(5):1196–1205.

Park, C. H., Song, E. G., and Ryu, K. H. (2018). Detection of Co-Infection of Notocactus leninghausii f. cristatus with Six Virus Species in South Korea. *The Plant Pathology Journal*, 34(1):65–70.

Peng, C., Yu, N. T., Luo, Z. W., Fan, H. Y., He, F., Li, X. H., Zhang, Z. L., and Liu, Z. X. (2016). Molecular Identification of Cactus virus X Infecting Hylocereus polyrhizus (Cactaceae) in Hainan Island, China. *Plant Disease*, 100(9):1956.

Pleše, N. and Miličič, D. (1966). Vergleichende Untersuchungen an Isolaten des Kakteen-X-Virus mit Testpflanzen. *Journal of Phytopathology*, 55(3):197–210.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596.

Ramanauskas, K. and Igić, B. (2021). Rnase-based self-incompatibility in cacti. *New Phytologist*, 231(5):2039–2049.

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

Simmonds, P., Adams, M. J., Benkő, M., Breitbart, M., Brister, J. R., Carstens, E. B., Davison, A. J., Delwart, E., Gorbalenya, A. E., Harrach, B., et al. (2017). Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, 15(3):161–168.

Song, L. and Florea, L. (2015). Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience*, 4(1):48.

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1):257.

Zhang, J., Kapli, P., Pavlidis, P., and Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, 29(22):2869–2876.

This table is unclear, unless linked with the supplementary figure showing current taxonomy and PTP.
The column names are also long, which makes it sparsely populated andunwieldy to display

Table 1. Summary statistic table for many metrics reported in this manuscript. Generally, the existing species definitions are well-supported.

Formal species	Monophyletic under PSC	Number of tips under PSC	Number of formally named tips	Supported by mPTP	Supported by bPTP	Supported by 72% RdRp	Supported by 72% CP
Cactus virus X	Y	34	3	N, two subsp supported	N, two subsp supported	Y	Y
Schlumbergera virus X	Y	13	3	N, two subsp supported	N, two subsp supported	Y	Y
Zygocactus virus X	Y	13	2	Y	Y	Y	Y
Opuntia virus X	Y	3	3	Y	Y	Y	Y
Pitaya virus X (Including Mytcor Virus 1)	Y	21	1	Y	Y	Y	Y

- Viral taxon color-coding is inconsistent (we'll pick a scheme and stick to it for the remainder of the paper)

- GenBank ID sequences are now missing the host taxon (e.g. see vNov2023)

- Column names are confusing and try to provide explanation. It's better to keep them very short (equally confusing), but visually clean and explain in caption

- Remove NA from m/bPTP legend

- I don't think that "Formal taxonomy" is quite the right name here, especially for Mytcor

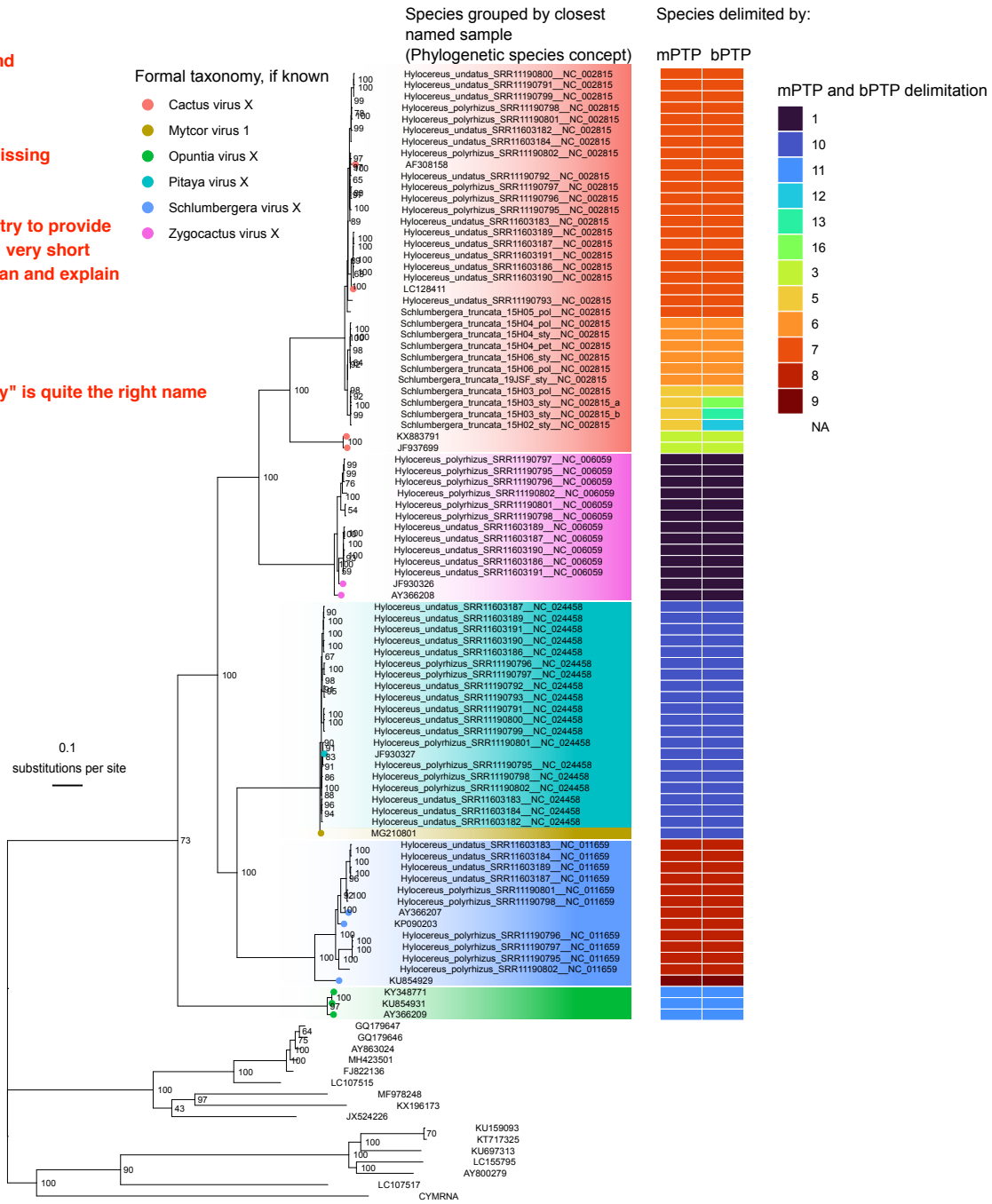

Figure 1. Maximum likelihood tree for full genome sequences from selected potexviruses. UFboot scores are represented on nodes. Tips are characteristically arranged in large groups with short branches, subtended by a longer branch. This subset of cactus-infecting potexviruses displays narrow sampling across related cohort plants which may represent local viral transmission from plant to plant. The outgroup clade consists of closely related but non-cactus infecting potexviruses.

- Move to Supplementary Figure S1 or S7 (after the individual gene phylogenies and upper matrix of co-phylo plots))

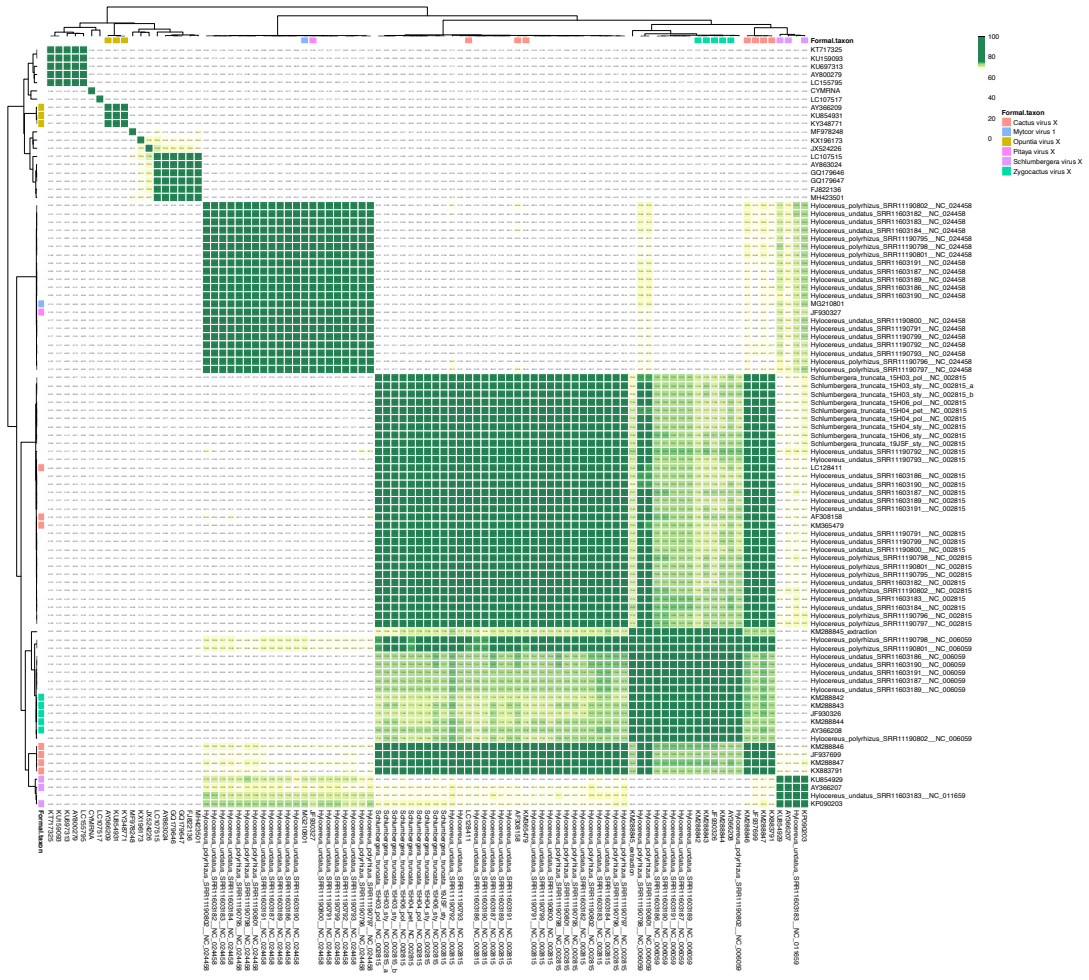

Figure 3. Heatmap displaying sequence similarity for Coat Protein (CP) sequences, using the same coloration as Figure 2. CP in potexviruses is typically located in ORF5 and this alignment represents 93 sequences each 714 nt in length.

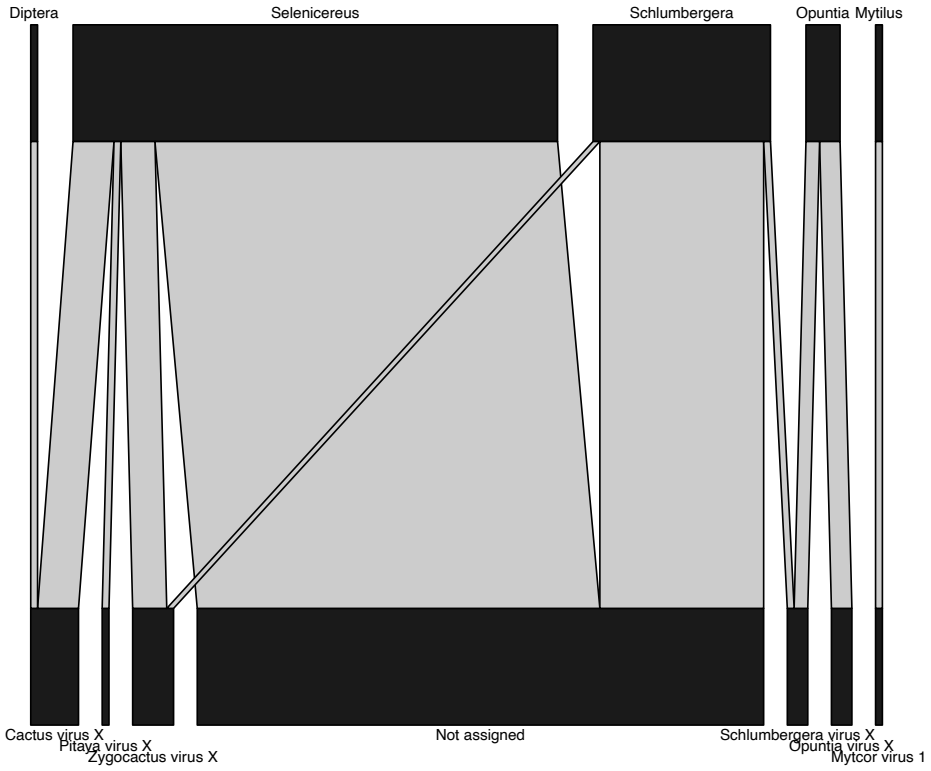

Figure 4. Comparing host information to formal viral species name for all samples (n=120) included in our dataset reveals the inconsistencies of viral naming schemes. Delimiting a viral species from its host plant is often an informative first step in viral discovery but may eventually lead to confusion.