Soil- and plant-associated viral ecology in natural and managed systems

By

Anneliek M. ter Horst

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Plant Pathology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Joanne B. Emerson

Amélie C. M. Gaudin

Johan H. J. Leveau

Committee in Charge

2023

Acknowledgements

First, I would like to thank dr. Joanne Emerson for giving me the opportunity to work in her lab. I am extremely grateful for all her support, patience, guidance and time, and I could not have imagined a better mentor for my PhD. Thank you to all the lab members for creating a friendly and collaborative work environment, but especially Christian Santos-Medellín, Jane Fudyma and Grant Gogul, who have been amazing friends as well. Thank you to my dissertation committee, Amélie Gaudin and Johan Leveau, for their insightful comments and encouragement. Thanks to Katie Simpson-Johnson, Jess Sorensen, Sara Geonczy and Devyn Durham for all their sampling help in the mud. And last but not least, thank you to my family and friends for all their support, love and encouragement over the years.

Contents

Dissertation abstract

Viruses are abundant in soil, and by infecting other soil biota, viruses have the potential to impact soil food webs, and carbon and nutrient cycling. While viruses are an important component of the soil microbiome, they are relatively understudied. To explore viral diversity and ecology in plants and natural and agricultural soil systems, in this thesis, I used a combination of total soil metagenomics, viral size-fraction metagenomics (viromics) and dsRNA metatranscriptomics to investigate the viral communities in Minnesotan peatland soils, various oak and conifer plant species, California wetland soils and agricultural bulk and rhizosphere soils. In chapter one of this thesis, we determined that viral communities are mainly structured by depth and water content in a Minnesota peatland. In Chapter two we found that the viral communities of oak and conifer species are predominantly structured by host tree phylogeny. In Chapter three, we uncovered that habitat characteristics, such as soil salinity and plant community, play an important role in structuring the soil virome in a California wetland ecosystem with a salinity gradient. In Chapter four, we determined that viruses are abundant in the rhizosphere microbiome, and that soil compartment, moisture content, and spatial location of the field all have significant impact on viral community composition. Moreover, we created a database for reference-based viral genome recovery, named Phages and Integrated Genomes Encapsidated Or Not (PIGEON), in order to explore viral biogeographical patterns. In conclusion, viruses are an important, diverse and understudied component of the soil microbiome, and here, we explored viral diversity and community structuring in a variety of habitats.

CHAPTER 1

Introduction

Viruses are the most abundant biological entities on Earth, harboring a substantial reservoir of genetic diversity [**1**]. In soil, current estimates suggest that viruses of bacteria (bacteriophages) are as abundant as, or even more abundant than their bacterial hosts [**2**], at an estimated amount of 10^7 - 10^9 viruses per gram of soil [**3**]. Viruses impact food webs, nutrient and carbon cycling, and host mortality [**2**, **4**]. Through infection and lysis of their hosts, viruses influence host metabolic function and soil chemistry [**4**, **5**]. For example, upon lysis, cells release their carbon and nutrient contents into the soil, which then become available for other members of the soil community [**4**, **6**, **7**]. Recent methodological improvements have made it possible to investigate soil and plant-associated viral communities in detail. Although shotgun metagenomic data can be mined to retrieve viral sequences bioinformatically, most sequences in total soil metagenomes are from bacterial (and occasionally eukaryotic) genomes [**8**, **9**], which dilute the viral signal. By purifying the viral size fraction through a 0.22 μm filter prior to DNA extraction and deep metagenomic sequencing (defined as viromics) [**10**], a much higher viral diversity can be recovered than via shotgun metagenomics [**8**, **11**]. To investigate soil viruses with RNA genomes, metatranscriptomic (shotgun RNA sequencing) data mining has been used to identify virus-encoded RNA-dependent RNA polymerase (RdRP) genes, for example, revealing significant differences in viral communities between soil compartments (bulk, rhizosphere and detritusphere) in grassland microcosms [**12**]. However, similar to shotgun metagenomes, total metatranscriptomes are dominated by non-viral sequences, such that targeted approaches to enrich the RNA viral signal are needed to advance the field, particularly in host-associated environments like plant tissues, from which the majority of sequences are host-derived. Such advances include RNA viromics, which has been successfully applied to grassland and peatland soils [**13**], and here (Chapter 3) we report dsRNA extraction and sequencing of oak and conifer leaves to explore viral communities associated with asymptomatic trees. In this dissertation, we report a new laboratory protocol for recovering DNA viromes from rhizosphere soils (Chapter 5), and we apply a suite of cutting-edge approaches (total metagenomics, viromics, total metatranscriptomics, and dsRNA sequencing) to assess soil and plant-associated viral biogeography across a range of ecosystems from the field to global scale.

To comprehensively investigate soil and plant-associated viral ecology, the studies reported here span a variety of ecosystems, from peatlands to fresh- and saltwater wetlands to agricultural soils, tomato rhizospheres, and oak and conifer phyllospheres (leaf surface and endophytic viruses). Wetlands are important ecosystems that are estimated to store between 20 and 30% of the global soil carbon [**14**], and microorganisms play key roles in carbon cycling and the emission of greenhouse gasses from these ecosystems [**15**]. Given the evidence for viral impacts on microbial ecology and biochemistry in other ecosystems [**16**, **17**], viruses are likely to play key roles in wetland ecosystem dynamics as well, but little is known about viral wetland ecology [**4**, **15**]. Forest trees, such as oaks and conifers, are of economic importance and have a broad ecological distribution, but there is a relative lack of knowledge about their associated viral diversity [**18**, **19**, **20**, **21**]. By expanding our knowledge of the natural tree virome, we might be able to better predict tree responses to emerging pathogens. The rhizosphere microbiome is an important factor for plant growth, health and nutrient acquisition [**22**, **23**, **24**], and viruses are likely impacting the rhizosphere microbiome by infecting rhizosphere microbes [**4**, **25**, **26**]. However, very little is known about the rhizosphere virome [**4**, **25**, **26**], and it remains to be seen if rhizosphere viral communities display similar patterns to those of their presumed host bacteria and fungi.

Prior to this dissertation, local and global distributions of soil viral species, as well as their habitat preferences, were unknown. Recent studies, including Chapter two of this dissertation, have shown that soil viral communities differ substantially at local-to-regional scales, with few viral species shared even in the same habitat meters apart [**7**, **11**, **27**, **28**]. To expand these studies to the global scale, we developed the viral population genomic reference database, Phages and Integrated Genomes Encapsidated Or Not (PIGEON), introduced in chapter two. In chapters two, four, and five of this dissertation, we leverage PIGEON to show that the same viral 'species' (viral operational taxonomic units, vOTUs, $\geq$ 10 kbp, $\geq$ 95% average nucleotide identity [**29**]) can be recovered on different continents, usually in the same type of habitat [**11**]. This dissertation demonstrates that, at both local and global scales, soil viral communities tend to differ most significantly by habitat type, for example, with freshwater wetland viruses differing from those in saltwater, agricultural soil viruses differing from those in natural soils, and viruses in peatlands differing from those in most other environments.

In chapter two, we sought to explore local and global peatland viral biogeography in climate-vulnerable peatlands, as well as compare methods for analyzing peat viral communities. We leveraged total soil metagenomes from the Spruce and Peatland Responses Under Changing Environments (SPRUCE) whole ecosystem warming experiment in the Marcell Experimental Forest (MEF) in northern Minnesota, and we generated viromes from the bog surrounding the SPRUCE experiment [30, 31]. Viral species (vOTU) recovery was 32 times higher from viromes compared to total soil metagenomes [11], indicating that a viromic approach vastly improved resolution of soil viral diversity and thus is more appropriate for soil viral ecological investigations. Whole ecosystem warming treatments did not significantly affect soil viral community composition during the first two years of the experiment, but viral communities differed significantly by peat depth, water content, and carbon chemistry, indicating local habitat characteristics as important drivers of viral biogeography [11]. Evidence for strong viral species boundaries between terrestrial and aquatic ecosystems was found at both local and global scales, with viral species shared in similar habitats. These results suggest that there may be specific niches for viruses in similar habitats, presumably partially driven by host niche preferences.

In chapter three, we explored viral diversity in 16 healthy oak and conifer tree species, as the diversity and role(s) of phyllosphere viruses are virtually unknown in asymptomatic trees [18, 19, 20, 21]. While most research on plant viruses has been focused on viruses that cause disease in economically important crops [32, 33], trees and other wild plants can harbor viruses as well. Some of these can cause disease within wild plant communities, and others can be reservoirs of emerging diseases in crops [34, 35, 36]. Some plant-associated viruses have positive, mutualistic, or neutral interactions with the plant host and can play important roles in the phytobiome [32, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46]. Here, we found that oak and conifer phyllosphere viral communities were significantly correlated to host plant phylogeny, suggesting that these viruses were highly host lineage-specific, potentially adapted to the host's physiological environment and/or its phytobiome. Many of these viruses were phylogenetically related to viruses with known plant and/or fungal hosts, suggesting that their primary hosts were plants or fungi, suggesting persistent, asymptomatic infection of the host plant and/or infection of members of the host plant microbiome. Interestingly, we recovered the greatest diversity of putative mycoviruses (fungal viruses) in an oak tree with senescing leaves (known to support saprobic fungi that feed on the decaying tissue [47]), suggesting increased mycoviral infection with increased host activity and hinting at a dynamic role for viruses in phyllosphere microbiomes that bears further exploration.

In chapter four, we investigated viral community biogeography across wetland habitats at a local scale (within one field site), and how habitat, plant community composition, and soil salinity affected viral community composition. We generated 63 viromes and total soil metagenomes from a California freshwater and saltwater wetland ecosystem. Viral communities were distinct at each wetland site, but they were secondarily structured by habitat characteristics, such as soil salinity and plant community composition, indicating environmental filtering. Globally, 1.5% of the vOTUs were previously recovered elsewhere in the world, and global biogeographical patterns were largely linked to habitat characteristics, suggesting wetland habitat-specific niches for these viruses. Together, these results suggest that environmental filtering, dispersal and dispersal limitation are likely drivers of both local and global wetland viral biogeographical patterns.

In chapter five, we investigated viral community biogeography and dynamics in rhizosphere and adjacent bulk soil microbiomes of tomato plants over one growing season. While the structure and function of microbial and fungal communities in rhizosphere microbiomes is better understood (for example, plant species, temporal (plant developmental stage) and spatial scales are known to shape the rhizosphere prokaryotic and fungal communities [**48, 49, 50, 51, 52**]), rhizosphere viral community composition and its underlying drivers are virtually unknown [**4, 23, 25, 26, 53**]. Viromic processing of rhizosphere samples had not been done previously, and here we report that viromics can be successfully applied to rhizosphere samples. Results showed that viral community composition was primarily structured by soil compartment (differing between rhizosphere and bulk soils) and, within bulk soils, by soil moisture contents. Plot location secondarily structured both bulk and rhizosphere viral communities, counter to the observed trends in host prokaryotic community composition, which most significantly separated by soil compartment, and, within each compartment by time. As of now, explanations for these differences are unknown, but there could be differences in dispersal limitation, or viruses may represent more active members of the host microbiome. 15% of the vOTUs were previously detected at the same field site in a different sampling year, suggesting stable or recurring viruses in these agricultural soilsover time. An additional 10% of the vOTUs had been previously detected elsewhere, nearly all from other agricultural sites, suggesting habitat specificity. Taken together, results suggest that tomato rhizosphere viral communities are a dynamic part of the rhizosphere microbiome, that respond to changes in the environment (such as soil moisture level and other members of the soil microbiome) and have greater dispersal limations compared to prokaryotes and fungi.

Overall, this dissertation demonstrates that applying viromic methods to plant ecosystems increased the knowledge on phyllosphere viral ecology in asymptomatic trees, where host plant phylogeny seems to play an important role. Applying viromic methods to soil ecosystems has resulted in a better understanding of host-virus interactions and the ecology and local and global biogeography of soil viral communities.

Bibliography

[1] David Paez-Espino, Emiley A Eloe-Fadrosh, Georgios A Pavlopoulos, Alex D Thomas, Marcel Huntemann, Natalia Mikhailova, Edward Rubin, Natalia N Ivanova, and Nikos C Kyrpides. Uncovering earth's virome. *Nature*, 536(7617):425–430, August 2016.

[2] Janet K Jansson. Soil viruses: Understudied agents of soil ecology. *Environ. Microbiol.*, 25(1):143–146, January 2023.

[3] Kurt E Williamson, Jeffry J Fuhrmann, K Eric Wommack, and Mark Radosevich. Viruses in soil ecosystems: An unknown quantity within an unexplored territory. *Annu Rev Virol*, 4(1):201–219, September 2017.

[4] Joanne B Emerson. Soil viruses: A new hope. *mSystems*, 4(3), May 2019.

[5] Yakov Kuzyakov and Kyle Mason-Jones. Viruses in soil: Nano-scale undead drivers of microbial life, biogeochemical turnover and ecosystem functions. *Soil Biol. Biochem.*, 127:305–317, December 2018.

[6] Joanne B Emerson, Simon Roux, Jennifer R Brum, Benjamin Bolduc, Ben J Woodcroft, Ho Bin Jang, Caitlin M Singleton, Lindsey M Solden, Adrian E Naas, Joel A Boyd, Suzanne B Hodgkins, Rachel M Wilson, Gareth Trubl, Changsheng Li, Steve Frolking, Phillip B Pope, Kelly C Wrighton, Patrick M Crill, Jeffrey P Chanton, Scott R Saleska, Gene W Tyson, Virginia I Rich, and Matthew B Sullivan. Host-linked soil viral ecology along a permafrost thaw gradient. *Nature Microbiology*, 3(8):870–880, July 2018.

[7] Christian Santos-Medellín, Katerina Estera-Molina, Mengting Yuan, Jennifer Pett-Ridge, Mary K Firestone, and Joanne B Emerson. Spatial turnover of soil viral populations and genotypes overlain by cohesive responses to moisture in grasslands. *Proc. Natl. Acad. Sci. U. S. A.*, 119(45):e2209132119, November 2022.

[8] Christian Santos-Medellin, Laura A Zinke, Anneliek M Ter Horst, Danielle L Gelardi, Sanjai J Parikh, and Joanne B Emerson. Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *ISME J.*, 15(7):1956–1970, July 2021.

[9] Gareth Trubl, Paul Hyman, Simon Roux, and Stephen T Abedon. Coming-of-age characterization of soil viruses: A user's guide to virus isolation, detection within metagenomes, and viromics. *Soil syst.*, 4(2):23, April 2020.

[10] Pauline C Göller, Jose M Haro-Moreno, Francisco Rodriguez-Valera, Martin J Loessner, and Elena Gómez-Sanz. Uncovering a hidden diversity: optimized protocols for the extraction of dsDNA bacteriophages from soil, 2020.

[11] Anneliek M Ter Horst, Christian Santos-Medellín, Jackson W Sorensen, Laura A Zinke, Rachel M Wilson, Eric R Johnston, Gareth Trubl, Jennifer Pett-Ridge, Steven J Blazewicz, Paul J Hanson, Jeffrey P Chanton, Christopher W Schadt, Joel E Kostka, and Joanne B Emerson. Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. *Microbiome*, 9(1):233, November 2021.

[12] Evan P Starr, Erin E Nuccio, Jennifer Pett-Ridge, Jillian F Banfield, and Mary K Firestone. Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. *Proc. Natl. Acad. Sci. U. S. A.*, 116(51):25900–25908, December 2019.

[13] Luke S Hillary, Evelien M Adriaenssens, David L Jones, and James E McDonald. RNA-viromics reveals diverse communities of soil RNA viruses with the potential to affect grassland ecosystems across multiple trophic levels. *ISME Commun*, 2:34, April 2022.

[14] A M Nahlik and M S Fennessy. Carbon storage in US wetlands. *Nat. Commun.*, 7:13835, December 2016.

[15] Paula Dalcin Martins, Robert E Danczak, Simon Roux, Jeroen Frank, Mikayla A Borton, Richard A Wolfe, Marie N Burris, and Michael J Wilkins. Viral and metabolic controls on high rates of microbial sulfur and carbon cycling in wetland ecosystems. *Microbiome*, 6(1):138, August 2018.

[16] Bonnie L Hurwitz and Jana M U'Ren. Viral metabolic reprogramming in marine ecosystems. *Curr. Opin. Microbiol.*, 31:161–168, June 2016.

[17] Simon Roux, Jennifer R Brum, Bas E Dutilh, Shinichi Sunagawa, Melissa B Duhaime, Alexander Loy, Bonnie T Poulos, Natalie Solonenko, Elena Lara, Julie Poulain, Stéphane Pesant, Stefanie Kandels-Lewis, Céline Dimier, Marc Picheral, Sarah Searson, Corinne Cruaud, Adriana Alberti, Carlos M Duarte, Josep M Gasol, Dolors Vaqué, Tara Oceans Coordinators, Peer Bork, Silvia G Acinas, Patrick Wincker, and Matthew B Sullivan. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, 537(7622):689–693, September 2016.

[18] C Büttner, S von Bargen, M Bandte, and H P Mühlbach. Forest diseases caused by viruses. In *Infectious forest diseases*, pages 50–75. CABI, Wallingford, 2013.

[19] Preston R Aldrich and Jeannine Cavender-Bares. Quercus. In Chittaranjan Kole, editor, *Wild Crop Relatives: Genomic and Breeding Resources: Forest Trees*, pages 89–129. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[20] Aljos Farjon. The kew review: Conifers of the world. *Kew Bull.*, 73(1):8, March 2018.

[21] David M Richardson, Philip W Rundel, Stephen T Jackson, Robert O Teskey, James Aronson, Andrzej Bytnerowicz, Michael J Wingfield, and Şerban Procheş. Human impacts in pine forests: Past, present, and future. *Annu. Rev. Ecol. Evol. Syst.*, 38(1):275–297, December 2007.

[22] Joseph Edwards, Cameron Johnson, Christian Santos-Medellín, Eugene Lurie, Natraj Kumar Podishetty, Srijak Bhatnagar, Jonathan A Eisen, and Venkatesan Sundaresan. Structure, variation, and assembly of the root-associated microbiomes of rice. *Proc. Natl. Acad. Sci. U. S. A.*, 112(8):E911–20, February 2015.

[23] Rodrigo Mendes, Paolina Garbeva, and Jos M Raaijmakers. The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. *FEMS Microbiol. Rev.*, 37(5):634–663, September 2013.

[24] Corné M J Pieterse, Ronnie de Jonge, and Roeland L Berendsen. The Soil-Borne supremacy. *Trends Plant Sci.*, 21(3):171–173, March 2016.

[25] Li Bi, Dan-ting Yu, Shuai Du, Li-mei Zhang, Li-yu Zhang, Chuan-fa Wu, Chao Xiong, Li-li Han, and Ji-zheng He. Diversity and potential biogeochemical impacts of viruses in bulk and rhizosphere soils, 2021.

[26] Akbar Adjie Pratama and Jan Dirk van Elsas. The 'neglected' soil virome – potential role and impact, 2018.

[27] Ruonan Wu, Michelle R Davison, William C Nelson, Emily B Graham, Sarah J Fansler, Yuliya Farris, Sheryl L Bell, Iobani Godinez, Jason E Mcdermott, Kirsten S Hofmockel, and Janet K Jansson. DNA viral diversity, abundance, and functional potential vary across grassland soils with a range of historical moisture regimes. *MBio*, 12(6):e0259521, December 2021.

[28] Devyn M Durham, Ella T Sieradzki, Anneliek M ter Horst, Christian Santos-Medellín, C Winston A Bess, Sara E Geonczy, and Joanne B Emerson. Substantial differences in soil viral community composition within and among four northern california habitats. *ISME Communications*, 2(1):1–5, October 2022.

[29] Simon Roux, Evelien M Adriaenssens, Bas E Dutilh, Eugene V Koonin, Andrew M Kropinski, Mart Krupovic, Jens H Kuhn, Rob Lavigne, J Rodney Brister, Arvind Varsani, Clara Amid, Ramy K Aziz, Seth R Bordenstein, Peer Bork, Mya Breitbart, Guy R Cochrane, Rebecca A Daly, Christelle Desnues, Melissa B Duhaime, Joanne B Emerson, François Enault, Jed A Fuhrman, Pascal Hingamp, Philip Hugenholtz, Bonnie L Hurwitz, Natalia N Ivanova, Jessica M Labonté, Kyung-Bum Lee, Rex R Malmstrom, Manuel Martinez-Garcia, Ilene Karsch Mizrachi, Hiroyuki Ogata, David Páez-Espino, Marie-Agnès Petit, Catherine Putonti, Thomas Rattei, Alejandro Reyes, Francisco Rodriguez-Valera, Karyna Rosario, Lynn Schriml, Frederik Schulz, Grieg F Steward, Matthew B Sullivan, Shinichi Sunagawa, Curtis A Suttle, Ben Temperton, Susannah G Tringe, Rebecca Vega Thurber, Nicole S Webster, Katrine L Whiteson, Steven W Wilhelm, K Eric Wommack, Tanja Woyke, Kelly C Wrighton, Pelin Yilmaz, Takashi Yoshida, Mark J Young, Natalya Yutin, Lisa Zeigler Allen, Nikos C Kyrpides, and Emiley A Eloe-Fadrosh. Minimum information about an uncultivated virus genome (MIUViG). *Nat. Biotechnol.*, 37(1):29–37, January 2019.

[30] Richard J Norby, Joanne Childs, Paul J Hanson, and Jeffrey M Warren. Rapid loss of an ecosystem engineer: Sphagnum decline in an experimentally warmed bog. *Ecol. Evol.*, 9(22):12571–12585, November 2019.

[31] Paul J Hanson, Jeffery S Riggs, W Robert Nettles, Jana R Phillips, Misha B Krassovski, Leslie A Hook, Lianhong Gu, Andrew D Richardson, Donald M Aubrecht, Daniel M Ricciuto, Jeffrey M Warren, and Charlotte Barbier. Attaining whole-ecosystem warming using air and deep-soil heating methods with an elevated CO_2 atmosphere. *Biogeosciences*, 14(4):861–883, February 2017.

[32] Marilyn J Roossinck. Metagenomics of plant and fungal viruses reveals an abundance of persistent lifestyles. *Front. Microbiol.*, 5:767, 2014.

[33] Marilyn J Roossinck, Darren P Martin, and Philippe Roumagnac. Plant virus metagenomics: Advances in virus discovery. *Phytopathology*, 105(6):716–727, June 2015.

[34] Ian Cooper and Roger A C Jones. Wild plants and viruses: Under-Investigated ecosystems. In *Advances in Virus Research*, volume 67, pages 1–47. Academic Press, January 2006.

[35] Yuxin Ma, Armelle Marais, Marie Lefebvre, Sébastien Theil, Laurence Svanella-Dumas, Chantal Faure, and Thierry Candresse. Phytovirome analysis of wild plant populations: Comparison of Double-Stranded RNA and Virion-Associated nucleic acid metagenomic approaches. *J. Virol.*, 94(1), December 2019.

[36] James E Schoelz and Lucy R Stewart. The role of viruses in the phytobiome. *Annual Review of Virology*, 5:93:111, July 2018.

[37] Pierre Lefeuvre, Darren P Martin, Santiago F Elena, Dionne N Shepherd, Philippe Roumagnac, and Arvind Varsani. Evolution and ecology of plant viruses. *Nat. Rev. Microbiol.*, 17(10):632–644, October 2019.

[38] Marilyn J Roossinck. Move over, bacteria! viruses make their mark as mutualistic microbial symbionts. *J. Virol.*, 89(13):6532–6535, July 2015.

[39] Marilyn J Roossinck. Deep sequencing for discovery and evolutionary analysis of plant viruses. *Virus Res.*, 239:82–86, July 2017.

[40] Ping Xu, Fang Chen, Jonathan P Mannas, Tracy Feldman, Lloyd W Sumner, and Marilyn J Roossinck. Virus infection improves drought tolerance. *New Phytol.*, 180(4):911–921, September 2008.

[41] Jack H Westwood, Lucy McCann, Matthew Naish, Heather Dixon, Alex M Murphy, Matthew A Stancombe, Mark H Bennett, Glen Powell, Alex A R Webb, and John P Carr. A viral RNA silencing suppressor interferes with abscisic acid-mediated signalling and induces drought tolerance in arabidopsis thaliana. *Mol. Plant Pathol.*, 14(2):158–170, February 2013.

[42] Donald L Nuss. Hypovirulence and chestnut blight: From the field to the laboratory and back. In J W Kronstad, editor, *Fungal Pathology*, pages 149–170. Springer Netherlands, Dordrecht, 2000.

[43] Donald L Nuss. Hypovirulence: mycoviruses at the fungal-plant interface. *Nat. Rev. Microbiol.*, 3(8):632–642, August 2005.

[44] Xiaofang Wang, Zhong Wei, Keming Yang, Jianing Wang, Alexandre Jousset, Yangchun Xu, Qirong Shen, and Ville-Petri Friman. Phage combination therapies for bacterial wilt disease in tomato. *Nat. Biotechnol.*, 37(12):1513–1520, December 2019.

[45] B Balogh, Jeffrey B Jones, F B Iriarte, and M T Momol. Phage therapy for plant disease control. *Curr. Pharm. Biotechnol.*, 11(1):48–57, January 2010.

[46] Colin Buttimer, Olivia McAuliffe, R P Ross, Colin Hill, Jim O'Mahony, and Aidan Coffey. Bacteriophages and bacterial plant diseases. *Front. Microbiol.*, 8:34, January 2017.

[47] Björn D Lindahl and Anders Tunlid. Ectomycorrhizal fungi - potential organic matter decomposers, yet not saprotrophs. *New Phytol.*, 205(4):1443–1447, March 2015.

[48] Peter A H M Bakker, Roeland L Berendsen, Rogier F Doornbos, Paul C A Wintermans, and Corné M J Pieterse. The rhizosphere revisited: root microbiomics. *Front. Plant Sci.*, 4:165, May 2013.

[49] Davide Bulgarelli, Matthias Rott, Klaus Schlaeppi, Emiel Ver Loren van Themaat, Nahal Ahmadinejad, Federica Assenza, Philipp Rauf, Bruno Huettel, Richard Reinhardt, Elmon Schmelzer, Joerg Peplies, Frank Oliver Gloeckner, Rudolf Amann, Thilo Eickhorst, and Paul Schulze-Lefert. Revealing structure and assembly cues for arabidopsis root-inhabiting bacterial microbiota. *Nature*, 488(7409):91–95, August 2012.

[50] Derek S Lundberg, Sarah L Lebeis, Sur Herrera Paredes, Scott Yourstone, Jase Gehring, Stephanie Malfatti, Julien Tremblay, Anna Engelbrektson, Victor Kunin, Tijana Glavina Del Rio, Robert C Edgar, Thilo Eickhorst, Ruth E Ley, Philip Hugenholtz, Susannah Green Tringe, and Jeffery L Dangl. Defining the core arabidopsis thaliana root microbiome. *Nature*, 488(7409):86–90, August 2012.

[51] Laurent Philippot, Jos M Raaijmakers, Philippe Lemanceau, and Wim H van der Putten. Going back to the roots: the microbial ecology of the rhizosphere. *Nat. Rev. Microbiol.*, 11(11):789–799, November 2013.

[52] Thomas R Turner, Euan K James, and Philip S Poole. The plant microbiome. *Genome Biol.*, 14(6):209, June 2013.

[53] M M Swanson, G Fraser, T J Daniell, L Torrance, P J Gregory, and M Taliansky. Viruses in soils: morphological diversity and abundance in the rhizosphere. *Ann. Appl. Biol.*, 155(1):51–60, August 2009.

CHAPTER 2

Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations

Anneliek M. ter Horst[1], Christian Santos-Medellín[1], Jess W. Sorensen[1], Laura A. Zinke[1], Rachel M. Wilson[2], Eric R. Johnston[3], Gareth G. Trubl[4], Jennifer Pett-Ridge[4], Steven J. Blazewicz[4], Paul J. Hanson[5], Jeffrey P. Chanton[2], Christopher W. Schadt[3], Joel E. Kostka[6,7], and Joanne B. Emerson* [1,8]

[1] Department of Plant Pathology, University of California, Davis, Davis, CA, USA

[2] Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, FL, USA

[3] Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

[4] Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California, USA

[5] Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

[6] Schools of Biology and Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, GA, USA

[7] Center for Microbial Dynamics and Infection, Georgia Institute of Technology, Atlanta, GA, 30332, USA

[8] Genome Center, University of California, Davis, Davis, CA, USA

Abstract

Background: Peatlands are expected to experience sustained yet fluctuating higher temperatures due to climate change, leading to increased microbial activity and greenhouse gas emissions. Despite mounting evidence for viral contributions to these processes in peatlands underlain with permafrost, little is known about viruses in other peatlands. More generally, soil viral biogeography and its potential drivers are poorly understood at both local and global scales. Here, 87 metagenomes and five viral size-fraction metagenomes (viromes) from a boreal peatland in northern Minnesota (the SPRUCE whole-ecosystem warming experiment and surrounding bog) were analyzed for dsDNA viral community ecological patterns, and the recovered viral populations (vOTUs) were compared to our curated PIGEON database of 266,125 vOTUs from diverse ecosystems.

Results: Within the SPRUCE experiment, viral community composition was significantly correlated with peat depth, water content, and carbon chemistry, including CH_4 and CO_2 concentrations, but not with temperature during the first two years of warming treatments. Peat vOTUs with aquatic-like signatures (shared predicted protein content with marine and/or freshwater vOTUs) were significantly enriched in more waterlogged surface peat depths. Predicted host ranges for SPRUCE vOTUs were relatively narrow, generally within a single bacterial genus. Of the 4,326 SPRUCE vOTUs, 164 were previously detected in other soils, mostly peatlands. None of the previously identified 202,371 marine and freshwater vOTUs in our PIGEON database were detected in SPRUCE peat, but 0.4 % of 80,714 viral clusters (VCs, grouped by predicted protein content) were shared between soil and aquatic environments. On a per-sample basis, vOTU recovery was 32 times higher from viromes compared to total metagenomes.

Conclusions: Results suggest strong viral “species” boundaries between terrestrial and aquatic ecosystems and to some extent between peat and other soils, with differences less pronounced at higher taxonomic levels. The significant enrichment of aquatic-like vOTUs in more waterlogged peat suggests that viruses may also exhibit niche partitioning on more local scales. These patterns are presumably driven in part by host ecology, consistent with the predicted narrow host ranges. Although more samples and increased sequencing depth improved vOTU recovery from total metagenomes, the substantially higher per-sample vOTU recovery after viral particle enrichment highlights the utility of soil viromics.

2.1. Introduction

Peatlands store approximately one-third of the world’s soil carbon (C) and have a significant role in the global C cycle [1]. Microbial activity in peatlands plays a key role in soil C and nutrient cycling, including soil organic C mineralization to the greenhouse gases, methane (CH_4) and carbon dioxide (CO_2) [2,3,4,5]. Given the abundance of viruses in soil (107 to 1010 per gram of soil [6,7,8,9]) and evidence for viral impacts on microbial ecology and biogeochemistry in other ecosystems [10,11,12], it is likely that viral infection of soil microorganisms influences the biogeochemical and C cycling processes of their hosts [**13**, **14**, **15**]. In marine ecosystems, viruses are estimated to lyse 20-40% of ocean microbial cells daily, impacting global ocean food webs and the marine C cycle [**16**, **17**, **18**], and viral contributions to terrestrial ecosystems are presumed to be similarly important but are less well understood [**6**, **13**, **15**, **19**, **20**, **21**].

Our current understanding of soil viral ecology stems from pioneering studies on viral abundance, morphology, amplicon sequencing, and lysogeny of bacteria [**7**, **22**, **23**, **24**, **25**, **26**], along with early viral size-fraction metagenomic (viromic) investigations [**27**, **28**, **29**]. More recently, total soil and wetland metagenomic datasets have been mined for viral sequences [**10**, **14**, **30**], revealing thousands of previously unknown viral populations (vOTUs) and suggesting habitat specificity for some of these viruses. Metatranscriptomic data mining has recently been used to explore RNA viral communities, revealing differences in bulk, rhizosphere, and detritusphere (plant litter-influenced) soil compartments [**31**], along with potential viral contributions to the ecology of the Sphagnum moss microbiome [**32**]. In addition to mining omic data for viral signatures, viromics (the laboratory enrichment of viral particles prior to DNA extraction and metagenomic sequencing) has recently been paired with high-throughput sequencing to investigate viral communities in soil [**13**, **14**, **33**, **34**]. Although we now have an array of laboratory and bioinformatics methods for soil viral ecology [**8**, **14**, **22**, **30**, **33**, **35**, **36**, **37**, **38**, **39**, **40**], we lack a thorough comparative understanding of these approaches and best practices.

Thawing permafrost peatlands have been the focus of several recent studies of viral diversity and virus-host dynamics, in order to better understand the ecological patterns underlying C emissions from these climate-vulnerable ecosystems [**13**, **14**, **41**, **42**, **43**]. Thawing permafrost peat has been characterized by relatively high viral diversity (thousands of vOTUs), including viruses predicted to infect methanogens and methanotrophs responsible for CH_4 cycling [**14**]. Evidence for more direct viral impacts on ecosystem C cycling has been revealed by the recovery of putative viral auxiliary metabolic genes (AMGs) [**13**, **14**], specifically, virus-encoded glycosyl hydrolases capable of degrading complex C into simple sugars [**14**]. Although we are gaining insights into soil viral ecology within specific ecosystems, our understanding of global soil viral biogeographical patterns is limited and is thus far derived predominantly from cultivation-based efforts [**43**, **44**].

In this study, we examined peat viral communities at the southern edge of the boreal zone in the Marcell Experimental Forest (MEF) in Minnesota, USA [**45**, **46**]. MEF has been the site of numerous studies on greenhouse gas emissions, C sequestration, hydrology, biogeochemistry, and vegetation [**47**, **48**, **49**, **50**, **51**, **52**]. To investigate the response of peatlands to increasing temperature and atmospheric CO_2 concentrations, the US Department of Energy (DOE) established the Spruce and

Peatland Responses Under Changing Environments (SPRUCE) experiment in MEF. This experiment is within an intact peat bog ecosystem, consisting of Picea mariana (black spruce) and Larix laricina (larch) trees, an ericaceous shrub layer, and a predominant cover of Sphagnum with minor contributions of other mosses [**45**, **46**, **53**]. SPRUCE researchers are studying whole-ecosystem responses to temperature and elevated CO_2 (eCO_2), including the responses of plants, above- and belowground microbial communities, and whole-ecosystem processes, such as greenhouse gas emissions [**1**, **45**, **46**, **54**, **55**, **56**, **57**, **58**], but as yet, the peat viral communities in this experiment remain unexplored.

Here, we used a combination of total soil metagenomics and viromics to: 1) investigate peat viral community composition and its potential drivers in the SPRUCE experiment, 2) place the recovered vOTUs in biogeographical and ecosystem context, and 3) compare the two approaches (total metagenomics and viromics) for recovering soil viral population sequences. We are also contributing a new database for reference-based viral genome recovery: the Phages and Integrated Genomes Encapsidated Or Not (PIGEON) database of 266,125 vOTU sequences from diverse ecosystems.

2.2. Results and Discussion

2.2.1. Dataset overview and peat viral population (vOTU) recovery.

To improve our understanding of peat viral diversity, we leveraged 82 peat metagenomes from cores collected from the SPRUCE experiment in northern Minnesota, USA in 2015 and 2016, along with five paired viromes and metagenomes that we collected along a transect outside the experimental plots from the same bog in 2018 at near-surface (top 10 cm) depths. In the field experiment, deep peat heating (DPH) and whole ecosystem warming (WEW) treatments heated the peat (to a depth of 2 m) and air inside 8 chambered enclosures (two per treatment) to target temperatures of +2.25, +4.5, +6.75 and +9 °C above ambient temperature [**1**, **46**, **53**, **59**]. There were also two ambient experimental chambers and two unchambered ambient plots (Table S1). Peat samples for metagenomics were collected from four depths (10-20 cm, 40-50 cm, 100-125 cm and 150-175 cm) per year in each chamber and unchambered ambient plot (38 and 44 total soil metagenomes were successfully sequenced in 2015 and 2016, respectively), with approximate sequencing depths of 6 Gbp per metagenome in 2015 and 15 Gbp in 2016. From each of the five transect peat samples (Supplementary Figure 1), a viral size-fraction metagenome (virome) and total soil metagenome were sequenced, each to a depth of approximately 14 Gbp.

Reads from the SPRUCE experiment metagenomes (82), transect viromes (5), and transect total soil metagenomes (5) were assembled into contigs $\geq$ 10 kbp, from which viral contigs were identified [**37**, **38**] and clustered into 5,006 approximately species-level viral populations (viral operational taxonomic units, vOTUs [**60**]). These vOTUs were then clustered with 261,799 vOTUs from diverse habitats in our PIGEON database (see methods, Table S2, available on Dryad (https://datadryad.org/, by DOI of this paper) [**10**, **13**, **14**, **30**, **33**, **61**, **62**, **63**, **64**, **65**]. The resulting clustered database of 266,125 "species-level" vOTUs was used as a reference for read mapping from each of our metagenomes. In total, we detected 4,326 vOTUs through read mapping from the SPRUCE experiment and adjacent peatlands. Henceforth, "SPRUCE" refers to our data from the SPRUCE experiment and/or transect, unless otherwise specified.

2.2.2. Investigating patterns and potential drivers of peat viral community composition in the SPRUCE experimental plots.

To characterize peat viral community compositional patterns and their potential drivers, vOTU abundances from the 82 SPRUCE experiment metagenomes were compared to environmental measurements. Using the 4,326 SPRUCE vOTUs as references, we recovered 2,699 vOTUs from the SPRUCE experimental plots through read recruitment and tracked their abundances (average per bp coverage depth) across the experimental plot metagenomes. No significant differences in viral community composition were detected according to temperature treatment (Mantel $\rho = 0.0057$, p = 0.56), as discussed in more detail below. Viral community composition was significantly correlated with depth (Fig 2.1A), even across different temperature treatments and years (Mantel $\rho = 0.57$, p=0.00001), consistent with previous evidence that viral community composition varies with depth in Swedish peatlands [**14**] and other soils [**66**]. These results are also consistent with observations of microbial communities in SPRUCE peat, where depth explained the largest amount of variation in peat microbial community composition, and temperature effects have thus far (from 2015-2018) not been significant [**1**, **56**]. We also measured a significant difference in viral community composition between the two sampling years (June 2015 and June 2016, PERMANOVA p=0.009). Other factors that significantly ($p < 0.05$) correlated with viral community composition included microbial community composition, porewater CO_2 and CH_4 concentrations, and the calculated fractionation factor for carbon in porewater $\delta 13CH_4$ relative to δCO_2 (αC) [**67**] (Table S3), which can be used to infer CH_4 production and consumption pathways [**3**, **14**, **67**, **68**]. Although all of these factors also co-varied with depth, interestingly, viral community composition was more significantly correlated with αC and porewater CH_4 concentrations than with depth. Together, these results prompted further exploration of potential explanations for these compositional patterns with depth, including links between SPRUCE vOTUs and water content, peat C cycling, and microbial hosts.

To investigate potential drivers of viral community compositional patterns with depth, we identified 121 vOTUs that exhibited significant differential abundance patterns across peat depth levels (adjusted-p $<$ 0.05, Likelihood Ratio Test). We assigned these vOTUs to one of three groups via hierarchical clustering (Fig 2.1B): vOTUs abundant in the near-surface (10-20 cm) but depleted at other depths, vOTUs abundant from 40-50 cm but depleted at other depths, and vOTUs abundant in only the two deepest depth ranges (100-125 and 150-175 cm). Given that near-surface peat had significantly higher gravimetric soil moisture measurements than deeper peat (p=0.002, Student's T-test), we used a trait-based approach to assign an "aquatic-like" trait to vOTUs that were found in the same viral clusters (VCs, based on predicted protein content) as vOTUs from freshwater and/or marine environments in our PIGEON database , and then we compared the proportion of aquatic-like vOTUs in the three depth-range groups. Near-surface depths displayed the highest proportion of aquatic-like vOTUs, followed by mid-depth s, while the deepest peat had zero recognizable aquatic-like vOTUs (Fig 2.1C). The proportion of aquatic-like vOTUs in the near-surface group was significantly higher than the aquatic-like proportion of the total set of 2,699 vOTUs ($p < 0.05$, Hypergeometric Test) ,suggesting that vOTUs in the surface horizons (and/or their hosts) might be better adapted to water-rich environments. Consistent with this interpretation , we did not exclude porewater from our samples [**3**, **8**, **14**, **43**], so it is likely that some of the vOTUs were derived from the porewater directly. Also, although water table depth measurements indicated that the entire sampled peat column was saturated for each of the samples, qualitatively, there was substantially more volumetric water content (waterlogging) in the near-surface depths compared to the deeper, more compacted peat. Although peat viral community composition was significantly correlated with both depth and measured soil moisture content (Mantel $p < 1E-5$), the Mantel r value was higher for the correlation with depth ($r = 0.569$) than with soil moisture ($r = 0.298$, Table S3), suggesting that differences in aquatic-like vOTUs alone do not fully explain the patterns in viral community composition with depth. Indeed, the underlying explanation for the observed enrichment of aquatic-like vOTUs in the near surface could be due to a variety of ecological similarities between near-surface peatlands and aqueous systems beyond simply water content (e.g., redox chemistry, substrates, and dissolved oxygen content [**41**, **69**]) and warrants further exploration in the future.

Under the assumption that patterns in viral community composition were at least partially indirect, resulting from interactions with hosts, we attempted to bioinformatically link SPRUCE vOTUs to microbial host populations [**14**]. All 4,326 vOTUs and a total of 486 bacterial and archaeal metagenome-assembled genomes (MAGs, 443 from the SPRUCE experiment metagenomes (Table S4) and

43 from the transect (>60% complete, <10% contaminated, Table S5)), were considered in this analysis. A total of 2,870 CRISPR arrays were recovered from the metagenomes via Crass [**70**], and 29 CRISPR-derived virus-host linkages were made between 23 vOTUs and 21 host MAGs (Fig 2.2, Table S6). For 25 of the 29 linkages, 0 mismatches were found between the CRISPR spacers and linked viral protospacers, and four linkages had a one-nucleotide mismatch. All 21 of the MAGs were bacterial and could be taxonomically classified to at least the family level, and for each of the six vOTUs linked to more than one host, the predicted hosts were all in the same family.Where genus-level host classification was possible, all vOTUs were predicted to infect the same host genus.

To investigate potential connections between virus-host dynamics and environmental conditions, along with viral community links to carbon chemistry, we attempted to assess virus-host abundance ratios and their patterns across samples, and we explored the auxiliary metabolic gene (AMG) content of the vOTUs. Only 10 virus-host pairs (10 vOTUs linked to 9 MAGs) were identified for which both the vOTU and the MAG were detected together in at least one sample, and significant patterns in virus-host abundance were not found for any of these pairs according to any of the parameters considered, including depth, year, αC, CH_4 and CO_2 concentrations, and moisture content. To further investigate the significant correlation between αC and viral community composition, we also looked for vOTU linkages to methanogen or methanotroph MAGs. HMM searches for McrA (a methanogenesis biomarker) [**71**, **72**], sMMO, pMMO, and pXMO (methanotrophy biomarkers) [**3**] predicted proteins were performed on the 443 SPRUCE experiment MAGs. Nine MAGs were found to contain McrA-encoding genes, and evidence for methanotrophy was found in 22 MAGs, but none of these MAGs had a CRISPR linkage to a vOTU. Thus, we infer either that αC co-varies with an unmeasured variable that better explains viral community composition and/or that important virus-host linkages associated with CH_4 cycling were not identified through these approaches. Finally, consistent with potential viral roles in the soil C cycle, we identified 287 putative AMGs encoded by viral genomes predicted to be involved in 18 C-cycling processes, based on VIBRANT and DRAM-v output [**39**, **40**] (Supplementary discussion table S7, S8, S9). These results are consistent with previously identified glycosyl hydrolase genes encoded in peat viral genomes [**13**, **14**], along with other putative C-cycling AMGs from soil [**73**, **74**] (see Supplementary Discussion).

As indicated above, no significant influence of temperature on viral community composition was detected over the first two years of experimental warming. Consistent with these findings, no differences in microbial community composition were found according to temperature treatments in these samples over the first five years of whole ecosystem warming, although warming exponentially increased CH_4 emissions and enhanced CH_4 production rates throughout the entire soil profile [**56**]. These results are also consistent with prior studies that have shown that soil microbial community responses to similar temperature increases can take multiple years to manifest [**26, 75, 76**]. Warming has been shown to substantially alter the community composition, diversity, and N_2 fixation activity of peat moss microbiomes [**57**], and in microcosms of surface peat collected from the SPRUCE site, microbial diversity was negatively correlated with temperature, suggesting that prolonged exposure of the peatland ecosystem to elevated temperatures will lead to a loss in microbial diversity [**77**]. In the SPRUCE experiment, the fractional cover of Sphagnum mosses [**45**] and plant phenology (the timing of different traits throughout the growing season) [**53**] have changed in response to temperature, suggesting that differences in belowground viral and microbial community composition may follow after a longer period of warming.

2.2.3. Placing SPRUCE peat viruses in global and ecosystem context.

Of the 4,326 "species-level" vOTUs from SPRUCE, 4,162 were assembled from SPRUCE-associated metagenomes (including the viromes), and 164 were recovered through read mapping to our PIGEON database of vOTUs from diverse ecosystems (Fig 2.3A). The 164 previously recovered vOTUs were first reported from other globally distributed sites, mainly peatlands (160 of 164), including peat vOTUs from Sweden (147), Germany (5), Alaska, USA (4), Wisconsin, USA (2), and Canada (2) (Fig 2.3B). The recovery of hundreds of viral species (4% of the dataset) in geographically distant peatlands suggests that there may be a peat-specific niche for these viruses. In addition, four vOTUs recovered from SPRUCE peat were first identified in a wet tropical soil in Puerto Rico, suggesting some global species-level sequence conservation across soil habitats (Table S10). Existing deeply sequenced soil viromic datasets are predominantly from peat [**8, 13, 14, 33**], so the extent to which these patterns reflect database bias or true differences between peat and other soils will require additional sampling.

Interestingly, despite the overwhelming dominance of marine vOTUs in our database (190,502 vOTUs, 71%), zero species-level vOTUs from the oceans were recovered in the SPRUCE peatlands. Freshwater vOTUs (predominantly from freshwater lakes) have less representation in our database (11,869 vOTUs, 4.45%), but similarly, no freshwater vOTUs were recovered from SPRUCE peat (though, as described above, vOTUs that shared higher-level taxonomy with aquatic viruses were recovered in SPRUCE peatlands). No other vOTUs from our PIGEON database, including bioreactor, hot spring, non-peat wetland, human-, plant-, and other host-associated vOTUs, were recovered in SPRUCE peat. These results suggest viral adaptation to soil and/or strong viral species boundaries between terrestrial, aquatic, and other ecosystems, as previously observed for bacterial species[80,81], though data for soil viruses are limited, so further studies across diverse soils will be necessary to assess the generalizability of these results.

To further compare vOTUs from diverse soil ecosystems, we constructed a phylogenetic tree of the terminase large subunit (terL) gene from 1,045 PIGEON soil vOTUs (81 from SPRUCE, 143 from other peat, and 821 from other soil) and 1,613 RefSeq prokaryotic viral genomes from which a terL sequence could be recovered (Figure 2.4a). Overall, the tree revealed two large superclades, one with predominantly RefSeq viral sequences and one with predominantly soil viral sequences (phylogenetic dispersion, D=-0.25, with $D < 0$ indicating significant phylogenetic separation of RefSeq and soil sequences [**78**, **79**]. As expected, these results indicate that known isolates do not adequately capture soil viral diversity. A second terL tree was constructed from only the soil sequences without RefSeq (Figure 2.4b), revealing approximately even phylogenetic distributions across soil habitats and no detectable soil habitat-specific phylogenetic groupings (D=0.58 for all peat vs. other soil, D=0.41 for SPRUCE vs. all other soil).

To assign taxonomy to vOTUs and group them at higher taxonomic levels for cross-ecosystem comparisons, the 4,326 SPRUCE vOTUs were clustered according to shared predicted protein content [**80**, **81**] with all other vOTUs in our PIGEON database, including 2,305 RefSeq viral genomes (release 85) [**64**]. The SPRUCE vOTUs formed 3,114 viral clusters (VCs), 2,193 of which were singletons and 921 of which contained at least two vOTUs (Table 1, Supplementary figure 2A). We note that, although singletons are not technically clusters, each VC has been suggested to represent a distinct viral “genus” [**80**, **81**], so we include singletons in VC counts for ease of interpretation. We describe each VC as a “genus”, in accordance with previously described terminology for this approach [84,85], but viral taxonomy is in flux [**82**, **83**], and an analysis of average amino acid identity (AAI) within 100 randomly chosen PIGEON VCs revealed that most VCs represent the equivalent of bacterial family or higher taxonomy. Briefly, vOTUs within most VCs shared an average of 45-65% AAI (for bacteria, that AAI range approximates the same family but different genera [**84**]), though ∼1/3 of the VCs had average AAIs above or below this range. Only fourteen of the SPRUCE VCs, containing 61 vOTUs (1.4% of the dataset), were taxonomically classifiable (Fig 2.3C, Supplementary figure 3). This is a lower proportion than a prior study [**14**], which we attribute at least in part to differences in the size of the dataset used for clustering (for example, 17% of peat vOTUs from northern Sweden were previously taxonomically classifiable [**14**], but only 3.9% of those same vOTUs could be taxonomically classified in our analysis, which included orders of magnitude more vOTUs but was otherwise similar, apart from use of the updated vConTACT2.0 pipeline instead of vConTACT). The taxonomically classifiable vOTUs from SPRUCE included 45 Myoviridae, five Podoviridae, four Siphoviridae, and seven Tectiviridae, consistent with the more abundant viral taxa previously reported from thawing permafrost peatlands [**14**], but we note that Myo-, Podo-, and Siphoviridae have been recommended for removal as taxonomic groups [**81**]. Although most SPRUCE VCs were not taxonomically classifiable, 562 included a vOTU that was also found in another dataset in PIGEON, meaning that just under 1/3 of the SPRUCE VCs had been observed before (compared to previous detection of only 4% of SPRUCE vOTUs, or viral "species").

All 31,049 of the vOTUs from soil in our PIGEON database, including those from SPRUCE and globally distributed soils, grouped into 20,939 VCs (Table 1). Of these, 16,524 included only a single vOTU, meaning that most of the known "genus"-level soil viral sequences have only been recovered from a single study and/or location so far. In total, 12.8% of the soil VCs were exclusively found in SPRUCE peatlands , 0.7% included at least one vOTU each from SPRUCE, other peat habitats, and other soils (Fig 2.3D), and 0.9% contained a vOTU from SPRUCE and other peat sites but not other soils. Together, these data suggest that, although much of soil viral sequence space remains to be explored, species-level similarities may be relatively restricted to specific soil habitat types, while similarities at higher taxonomic levels may be more common across soil habitats.

To investigate similarities between viruses from soil and aquatic (marine and freshwater) ecosystems, 233,420 vOTUs from our PIGEON database (31,049 soil [**10, 14, 30, 34**], 190,502 marine [**30, 62, 63**], and 11,869 freshwater [**30**]) were clustered into 80,714 VCs (Table S11). Of the soil VCs, 0.4% shared a cluster with vOTUs from one or both aquatic systems, indicating a small amount of "genus"-level similarity between aquatic and soil viruses (Fig 2.3E). However, most VCs were found in only one habitat, consistent with differences in microbial community composition in aquatic compared to soil and sediment habitats and between freshwater and saltwater environments [**85**].

2.2.4. Comparing viral recovery from viromes and total soil metagenomes.

Metagenomic studies of viral community composition typically take one of two approaches: either the viral signal is mined from total metagenomic assemblies, which predominantly tend to contain bacterial sequencing data [**13, 14, 30**], or viral particles are physically separated from other microbes in the laboratory (e.g., through filtration), and then viral size-fraction enriched metagenomes (viromes) are sequenced and analyzed [**12, 13, 14, 18**]. To directly compare results from both approaches, we first analyzed the paired total soil metagenomes and viromes from the five transect samples. Considering all assembled contigs $\geq$ 10 kbp, only 0.8% of the metagenomic contigs were classified as viral after passing them through viral prediction software (see methods), relative to 16% of the virome contigs. This ∼20-fold improvement is consistent with our observed ∼30-fold improvement in viral contig recovery from viromes relative to total metagenomes in agricultural soils [**34**], and similar differences in the composition of metagenomes and viromes have been reported from grassland soils [**86**]. When accounting for read mapping to all vOTUs in the PIGEON database (including all of the SPRUCE vOTUs), 1,952 vOTUs were detected in the viromes, relative to 401 in the metagenomes from the same samples (Fig 2.5A, Supplementary figure 4A). Only 37 vOTUs were detected in the metagenomes alone. Although far more vOTUs were recovered from the viromes, vOTU accumulation curves were still climbing steeply after five samples for both viromes and metagenomes (Fig 2.5B, Supplementary figure 4B, 4C), suggesting that more viral diversity remains to be recovered . A comparison of the five viromes indicated that there was no spatial relationship between the samples (Supplementary figure 5A), but there was high variability in the number of recovered vOTUs per sample (Supplementary figure 5B).

To place these comparisons from the same samples in the context of the larger SPRUCE dataset, we compared the five viromes from 2018 to the 82 metagenomes from 2015 and 2016, again with vOTU recovery assessed through read recruitment to all vOTUs in the PIGEON database. We note that the samples in this set of comparisons differ in multiple ways beyond the extraction method, including the sampling year, depth range, location, and (in some cases) temperature treatment, all of which could contribute to the observed trends. On a per-sample basis, the viromes recovered far more vOTUs than the metagenomes, as indicated by the much steeper accumulation curve slope for viromes after only five samples (Fig 2.5B). However, the much larger number of samples in the SPRUCE experimental plot metagenomes resulted in a higher total vOTU recovery of 2,699 in the 82 metagenomes, compared to 1,952 in the five viromes (Fig 2.5A).

We next considered the metagenomes from 2015 and 2016 separately, because the sequencing throughput from 2016 was 1.4 times higher than in 2015. The first of these comparisons was based on read recruitment only to vOTUs derived from contigs that assembled from samples in the same category, considering four categories: the five transect viromes, five transect metagenomes, 38 metagenomes from 2015, and 44 metagenomes from 2016. These "self-mapped" analyses were meant to simulate a situation in which only the vOTUs from that particular dataset would have been available. The perceived viral richness per sample was 32 times higher in viromes (mean 649 vOTUs) compared to their paired metagenomes (mean 20 vOTUs) but was nine and three times higher, respectively, in viromes compared to the 2015 and 2016 metagenomes (mean 72 and 207 vOTUs) (Fig 2.5C). The perceived viral richness was 2.8 times higher in the 2016 metagenomes compared to 2015 metagenomes, indicating that a greater sequencing depth of total soil metagenomes (in this case from 6 to 15 Gbp on average) likely increased vOTU recovery, though we cannot exclude the possibility of a true difference in viral richness between the two years. A further comparison of vOTU recovery from the transect viromes and the three sets of metagenomes was based on read recruitment to all 266,125 PIGEON vOTUs from SPRUCE and other datasets. In this case, the perceived viral richness in the viromes (mean 721 vOTUs) was 5.7 times higher than in the paired metagenomes (mean 127 vOTUs), 3.5 times higher than in the 2015 metagenomes (mean 200 vOTUs), and two times higher than in the 2016 metagenomes (mean 370 vOTUs, Fig 2.5D). Thus, the availability of reference vOTUs, particularly from the SPRUCE viromes, substantially improved recovery from the total metagenomes.

Lastly, we compared the VCs formed by vOTUs from the 2018 viromes, the 2018 metagenomes, and the 2015/2016 metagenomes to determine whether there were differences in the taxonomic space recovered by the different approaches. When comparing the five paired total metagenomes and viromes, all of the metagenome vOTUs shared a VC with at least one vOTU from the viromes, whereas 1,401 vOTUs were in VCs exclusively recovered from the viromes, indicating that viromes expanded the recoverable viral taxonomic space relative to paired metagenomes (Supplementary figure 2A, 2B). However, the vOTUs recovered from the unpaired 2015/2016 metagenomes recovered substantially different VCs compared to the 2018 viromes. We suspect that these differences were largely due to the different collection years, locations, and, particularly, numbers of samples, as opposed to differences between extraction methods.

Few direct comparisons of viromes and total metagenomes from the same samples have been reported from any ecosystem. Consistent with these results from peat, agricultural and grassland soil viromes have been shown to be enriched in both viral sequences and genomes from ultrasmall cellular organisms (which would be more likely to pass through the 0.2 μm filters used for viral enrichment) but depleted in sequences from most other cellular organisms, compared to total metagenomes [**34**, **86**]. In aqueous systems, water samples are often separated into multiple size fractions (for example, 3-20 μm, 0.8-3 μm, 0.2-0.8 μm, post-0.2 µm), such that previous studies have compared viral sequences recovered across different size fractions, and generally, the viruses recovered from different size fractions seem to be distinct [**87**, **88**]. A recent meta-analysis of human gut viral data recovered from viromic and metagenomic sequences suggested that more viral contigs could be recovered from metagenomes than from viromes [**83**]. However, of the 2,017 viromes considered in that study, 1,966 were multiple-displacement amplification (MDA) treated, and, as the authors acknowledged, MDA of viromes has known methodological biases (for example, MDA preferentially recovers circular ssDNA viruses [**6**]) and thus would result in artificially lower-richness viral communities. Although differences in the environments could have contributed to the observed differences in viral recovery from viromes compared to total metagenomes in the human gut study compared to our work, the large difference in the number of total metagenomes (680) compared to non-MDA amplified viromes (51) in the human gut study could also have contributed to the greater recovery of viral sequences from total metagenomes in that study. Consistent with that interpretation, here we have shown that increasing the number of samples, in combination with deeper sequencing and the availability of relevant reference vOTU sequences, improved vOTU recovery from total soil metagenomes, which have the added advantage of accessing virus and host population sequences from the same dataset.

2.3. Conclusions

We analyzed dsDNA viral diversity in a climate-vulnerable peat bog, revealing significant differences in viral community composition at different soil depths and according to peat and porewater C chemistry. Aquatic-like SPRUCE vOTUs were significantly more abundant at near-surface depths, suggesting potential adaptation of these viruses to water-rich environments. Some viral species-level similarities were observed across large geographic distances in soil: 4% of the vOTUs found in SPRUCE peat were previously recovered elsewhere, predominantly in other peatlands. Interestingly, zero marine or freshwater vOTUs were recovered from SPRUCE peat, suggesting the potential for viral species boundaries between terrestrial and aquatic ecosystems. When comparing vOTU recovery from viromes and total soil metagenomes, increasing the dataset size through deeper sequencing and more samples improved vOTU recovery from metagenomes, but viromics was a better approach for maximizing viral recovery on a per-sample basis. Together, these results expand our understanding of soil viral communities and the global soil virosphere, while hinting at a vast diversity of soil viruses remaining to be discovered.

2.4. Materials and Methods

2.4.1. Sample collection.

In June 2018, five peat samples were collected along “Transect 4” in the S1 bog $\sim$ 150 m from the SPRUCE experimental plots in the Marcell Experimental Forest in northern Minnesota, USA (For GPS coordinates, see Table S12). Avoiding green Sphagnum moss capitula at the surface ($\sim$ 2 cm), the top 10 cm of peat (5 cm diameter) was collected for each sample with a sterile spatula and placed in 50 mL conical tubes on dry ice. Samples were stored at -80°C for 6 months prior to DNA extraction for total metagenomes and viromes.

Within the SPRUCE study, temperature treatments were applied in large (∼115 sq m) open-topped enclosures. Temperature treatments in the 10 enclosures were as follows: +0, +2.25, +4.5, +6.75 and +9, with two chambers assigned to each temperature treatment. Data were also collected from two ambient environment plots where there was no enclosure but within the treatment area on the south end of the S1 Bog. In each enclosure, warming of deep soil started in June 2014 [**46**], and aboveground warming began in August 2015 with continuous whole ecosystem warming (365 days per year) operating since late in 2015. A more detailed explanation of deep soil heating procedures and construction of the enclosures and warming mechanics can be found in Hanson et al., 2017 [**45**, **46**, **53**].

Peat samples for 82 total soil metagenomes were collected from the SPRUCE experiment in June 2015 and June 2016 from cores that were extracted using defined hand sampling near the surface and via Russian corers below 30 cm. Samples for analysis were obtained from depth ranges 10-20 cm, 40-50 cm, 100-125 cm, and 150-175 cm from a total of 10 chambers in 2015 (no samples were analyzed from the open, ambient plots that year), with the exception of only two samples collected from chamber 19 (control plot, no temperature treatment, only 10-20 cm and 40-50 cm samples collected), for a total of 38 samples from 2015. In 2016, samples were collected from the same depth ranges from all 10 chambers, plus two samples from each of the two ambient, open plots (depth ranges 10-20 cm and 40-50 cm), for a total of 44 samples from 2016. These 82 samples were used for DNA extraction and total metagenomic analysis and MAG recovery, as described below. Soil temperature, moisture content, CH_4 and CO_2 concentrations, and αC measurements (see supplementary methods) were collected from the same samples (Table S13).

2.4.2. DNA extraction.

All samples from the peatland transect were stored at -80°C until further processing. 24 hours prior to DNA extraction, samples were placed at -20 °C. For total metagenomes from the transect, DNA was extracted from 0.25 g peat per sample with the QIAGEN DNeasy Powersoil Kit (QIAGEN, Germany), according to the manufacturer's protocol. For viromes, 50 g of peat per sample was divided between two 50 mL conical tubes, and 37.5 mL of Amended Potassium Citrate Prime buffer (AKC', 0.02 μm filtered, 1% K-citrate + 10% PBS + 150 mM $MgSO_4$) [**33**] was added per tube, for a total of 75 mL buffer. Tubes were shaken at 400 rpm for 15 min, then centrifuged at 4,700 g for 20 min. Excluding the pelleted soil, the supernatant was filtered through a 0.2 μm polyethersulfone filter (Corning, USA) and ultracentrifuged in a

Beckman LE-8K ultracentrifuge with a 70 Ti rotor for 3 hours at 32,000 RPM at 4 °C under vacuum. The supernatant was decanted, and the pellet containing virions was resuspended in 200 μl UltraPure water and added to the QIAGEN DNeasy PowerSoil Kit bead tubes (QIAGEN, Germany) for DNA extraction according to the manufacturer's instructions with one exception: instead of vortexing for 10 minutes with the beads, samples in the bead tubes were incubated at 70 °C for 10 min, vortexed briefly, and incubated at 70 °C for another 5 min. Consistent with our prior work on hypersaline lake viromes, which showed that a DNase treatment of viromes stored frozen resulted in removal of all DNA [**89**], a DNase treatment was not included prior to virion lysis. For the 82 2015 and 2016 peat samples used in metagenomic analysis and MAG recovery, DNA was extracted from homogenized samples of each depth interval using the MO BIO Powersoil DNA extraction kit (QIAGEN, Germany). Six replicate 0.35 g extractions were combined and re-purified with the MO BIO PowerClean Pro kit (QIAGEN, Germany) and eluted in 50 mL of 10 mM Tris buffer.

2.4.3. Library construction and sequencing.

Library construction and sequencing for the five viromes and five total soil metagenomes from Transect 4 were conducted by the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center. Libraries were prepared with the DNA Hyper Prep library kit (Kapa Biosystems-Roche, Basel, Switzerland), as previously described [**34**]. Paired-end sequencing (150 bp) was done on the Illumina NovaSeq platform, using 4% of a lane per virome and 8% of a lane per total soil metagenome. Sequencing of the 82 metagenomes from the SPRUCE experiment and ambient plots was done by the DOE Joint Genome Institute (JGI), using standard protocols for Nextera XT metagenomic library construction. These barcoded libraries were sequenced on an Illumina HiSeq 2500 instrument in 2x150 bp mode.

2.4.4. Sequencing read processing, assembly, viral population (vOTU) recovery, and read mapping.

Raw reads from the SPRUCE experiment metagenomes (82), transect viromes (5), and transect total soil metagenomes (5) were first quality-trimmed with Trimmomatic v0.38 [**90**] with a minimum base quality threshold of 30 evaluated on sliding windows of 4 bases and minimum read length of 50. Reads mapped to the PhiX genome were removed with bbduk [**91**]. Reads were assembled into contigs $\geq$ 10 kbp in length, using MEGAHIT v 1.1.3 [**92**] with standard settings. All 92 metagenomes underwent single-sample assemblies, and two additional co-assemblies were generated from the transect, one each for the five viromes and five total soil metagenomes, respectively. For co-assemblies, the preset meta-large option was used. 82 previously existing assemblies from the SPRUCE experiment metagenomes were also used. Briefly, for those assemblies, raw metagenomic fastq sequences were quality trimmed with bbduk from the BBTools software package (options: qtrim=window,2 trimq=17 minlength=100) [**93**] and assembled with IDBA-UD [**94**] (options: -mink 43 –maxk 123 –step 4 –min_contig 300).

DeepVirFinder [**38**] and VirSorter [**62**] were used to recover viral contigs from each assembly. Briefly, DeepVirFinder is a machine-learning approach that recognizes viral sequence signatures, and VirSorter searches for viral hallmark genes in PFAM annotation. Consistent with established recommendations, contigs with DeepVirFinder scores > 0.9 and $p < 0.05$ were considered viral [**63**], and DeepVirFinder results were filtered with a custom python script (parse_dvf_results.py, all scripts are available on GitHub, see Data Availability Statement below) to only retain results in compliance with this score. VirSorter was run in regular mode for all total metagenomes and in virome decontamination mode for the viromes. Only contigs from VirSorter categories 1, 2, 4 and 5 (high-confidence) were retained, as previously recommended [**62**]. All resulting viral contigs were clustered into vOTUs using CD-HIT [**95**] at a global identity threshold of 0.95 across 85% of the length of the shorter contig [**60**]. Different sets of vOTUs were used as references for read mapping throughout the manuscript (see main text), with the most commonly used and most comprehensive reference database being PIGEON (see below). In all cases, read mapping was performed with BBMap [**96**] at $\geq$ 90% identity, following thresholds set previously [**14**, **60**, **97**], and vOTU coverage tables were generated with BamM [**98**], using the 'tpmean' setting, and bedfiles were generated using bedtools [**99**]. Custom python scripts (percentage_coverage.py, filter_coveragetable.py) were used to implement the thresholds for detecting viral populations (vOTUs) in accordance with community standards ($\geq$ 75% of the contig length covered $\geq$ 1x by reads recruited at $\geq$ 90% nucleotide identity) [**60**]. The final vOTU coverage table of per-bp vOTU abundances in each metagenome was normalized by the number of metagenomic sequencing reads for each sample [**14**].

2.4.5. Construction of the PIGEON reference database of vOTUs.

An in-house database, Phages and Integrated Genomes Encapsidated Or Not (PIGEON), was created, containing 266,125 species-level vOTUs, of which 190,502 came from marine environments, 11,869 from freshwater, 31,049 from soil (including 4,326 from SPRUCE), 2,305 RefSeq viral genomes (release 85) [**64**], and 30,400 from other environments in a meta-analysis, including human microbiomes, other animal microbiomes, plant microbiomes, and other environments). Available viral contigs were downloaded from published datasets [**10**, **13**, **14**, **30**, **33**, **37**, **61**, **63**, **64**, **65**], compiled from ongoing work in Alaskan peat soil and Puerto Rican soils (see supplementary methods), and those recovered from SPRUCE (see above). For most of the previously published datasets, viral contigs were derived from viromes, or a combination of viromes and total soil metagenomes, but two datasets only considered viral recovery from total soil metagenomes [**10**, **30**]. For all but one of the datasets, VirSorter [**62**], VirFinder [**100**], DeepVirFinder [**38**], or a combination of these programs was used for viral contig recovery (Contigs with DeepVirFinder scores > 0.9 and $p < 0.05$ were considered viral [**63**], and only contigs from VirSorter categories 1, 2, 4 and 5 were considered). The exception was the meta-analysis dataset of Paez-Espino et al. (2016), which used its own viral discovery pipeline [**30**]. From all of these datasets, viral contigs were downloaded, and those >10 kbp were retained and then clustered into vOTUs using CD-HIT [**95**] at a global identity threshold of 0.95 across 85% of the shorter contig length to generate PIGEON v1.0. PIGEON v1.0 (the version used in this manuscript) is available on Dryad (https://datadryad.org/, by DOI of this paper). We are actively improving PIGEON and expect to release a new version in the future.

2.4.6. Viral taxonomic classification and protein-based viral clustering.

Viral taxonomic classifications for the 4,326 SPRUCE vOTUs (detected in the SPRUCE dataset through read mapping) were assigned using vConTACT2 (options: –rel-mode 'Diamond' –db 'ProkaryoticViralRefSeq85-Merged' -pcs-mode MCL –vcs-mode ClusterONE) [**80**, **81**]. The vOTUs were clustered according to shared predicted protein content with the 261,799 other vOTUs in our PIGEON database, including 2,305 RefSeq viral genomes [**64**]. The viral_cluster_overview output file was used for further analysis, including to manually identify SPRUCE vOTUs that shared a viral cluster with one or more vOTUs from marine and/or freshwater (aquatic) environments. For the analysis of AAI within PIGEON VCs, a random set of 100 VCs was analyzed with CompareM (standard settings) [**101**]. For each VC, the mean pairwise AAI between vOTUs was calculated.

2.4.7. Metagenome-assembled genome (MAG) reconstruction.

MAG reconstruction from the five transect total metagenomes was done as follows: quality-trimmed reads were assembled using MEGAHITv 1.1.3 [**92**] with a minimum contig length of 2,000, using the meta-large preset. After individual assembly of each sample, quality-filtered and trimmed reads were mapped to the resulting contigs using bbmap [**96**] with standard settings, and this abundance information was used to bin the contigs into MAGs using MetaBAT [**102**], using the –veryspecific setting and the coverage depth information. Quality and identification of bins was done with CheckM [**103**], following Sorensen et al., [**104**].

From the 82 SPRUCE experiment metagenomes, metagenome assembly, recovery, and analysis of metagenome-assembled genomes (MAGs) was performed as described in Johnston et al., [**105**]. Briefly, metagenomic sequences were assembled with IDBA-UD [**94**] (options: -mink 43 –maxk 123 –step 4 –min_contig 300). Resulting contigs $\geq$ 2.5 kbp were used to recover microbial population genomes with MetaBAT2 (options: –minCVSum 10) [**102**] and MaxBin2 [**106**]. Before binning, Bowtie2 was used to align short-read sequences to assembled contigs (options: –very-fast) [**107**], and SAMtools was used to sort and convert SAM files to BAM format [**108**]. Sorted BAM files were then used to calculate the coverage (mean representation) of each contig in each metagenome. The quality of each resulting MAG was evaluated with the CheckM v1.0.3 taxonomy workflow for Bacteria and Archaea separately [**103**]. The result from either evaluation (i.e., taxonomy workflow for Archaea or Bacteria) with the highest estimated completeness was retained for each MAG. MAGs with a quality score $\geq$ 60 were retained (from Parks et al., 2017 [**109**] calculated as the estimated completeness – 5 × contamination). MAGs recovered from different metagenomes were dereplicated with dREP [**110**], and the GTDB-tk classify workflow [**111**, **112**] was used to determine MAG taxonomic affiliations. MAG gene prediction, functional annotation, and assessment of metabolic pathway completeness (e.g., for assessing methanogenesis potential) was performed as described in Johnston et al., 2019 [**105**]. Taxonomic classification, source dataset SRA ID, basic genome statistics, and CheckM summaries for each MAG can be found in Table S4.

Using the parameters described above for vOTU coverage table generation, a microbial contig coverage table was generated. From this coverage table, we calculated the coverage of each population genome as the average of all of its binned contig coverages, weighting each contig by its length in base pairs. In-house scripts for this are available on GitHub. Hmm searches were done on both MAGs and vOTUs for proteins involved in methanogenesis or methanotrophy (McrA (a methanogenesis biomarker) [**71**, **113**], sMMO, pMMO, and pXMO (methanotrophy biomarkers) [**3**]). The MAG and vOTU contigs were annotated with prodigal (standard settings) [**114**], and an HMM search was done on these annotations with hmmr [**115**], using hmmsearch (standard settings) with an e-value cutoff of 1E-5 [**72**].

2.4.8. Reconstruction of microbial CRISPR arrays and virus-host linkages.

CRISPR repeat and spacer arrays were assembled with Crass v0.3.12 [**70**], using standard settings, and BLASTn was used to match spacer sequences with vOTUs and repeats to MAGs, in order to link viruses to putative hosts. Briefly, for protospacer-spacer matches (i.e., matches between vOTUs and CRISPR spacer sequences), the BLASTn-short function was used, with $\leq$ 1 mismatch to spacer sequences, e-value threshold of 1.0×10^{-10}, and a percent identity of 95 [30, 116]. For MAG-repeat matches, the BLASTn-short function was used, with an e-value threshold of 1.0×10^{-10} and a percent identity of 100 [14].

2.4.9. Phylogenetic tree construction.

A phylogenetic tree of bacterial host MAGs with CRISPR matches to one or more vOTUs (i.e., a repeat match to a MAG and a spacer from the same CRISPR array with a match to a vOTU protospacer) was constructed with CheckM [103] via a marker-gene alignment of 43 conserved marker genes with largely congruent phylogenetic histories, defined by [103]. This alignment was used to construct a maximum-likelihood tree with MEGA [117], with the LG plus frequencies model [118]. A total of 500 bootstrap replicates were conducted under the neighbor-joining method with a Poisson model.

For the terminase large subunit (TerL) tree, we predicted proteins on all viral contigs from PIGEON soil-associated vOTUs (n=31,346) with Prokka [119], (std settings, –kingdom viruses, –norrna –notrna), resulting in 1,045 large terminase subunit predictions. We downloaded the terminase large subunits (n=2799) that were available from RefSeq and clustered the Refseq terminase sequences at 95% AAI using USEARCH, following [31, 120], resulting in 1,613 terminase sequences from RefSeq. We then aligned predicted terminase sequences from PIGEON soil vOTUs with those from RefSeq (2,658 sequences total), using MAFFT v7.471 [121] with the G-INS-1 algorithm and otherwise standard settings [122]. Ambiguous aligned regions were removed using the TrimAl v1.41 program with the ‘gappyout’ setting [121, 123]. The best model of amino acid substitution was determined using ProtTest v1.5, standard settings [124]. Phylogenetic trees were constructed with IQ-TREE v1.6.12, [125], using -st AA -m LG+I+G4+F -bb 1000 -alrt 1000 options. Trees were visualized using iTol [126]. Bootstrap support was calculated, using an approximate likelihood ratio test (aLRT) with the Shimodaira–Hasegawa-like procedure (SH-aLRT), using 1000 bootstrap replicates.

2.4.10. Data analysis (ecological statistics).

The following statistical analyses were performed in R using the Vegan [127] package: accumulation curves were calculated using the speccacum function, vOTU coverage tables were standardized using the decostand function with the Hellinger method, and Bray-Curtis dissimilarity matrices were calculated using the vegdist function. Mantel tests were performed with the mantel function, using the Pearson method, and permutational multivariate analyses of variance (PERMANOVA) were performed with the Adonis function. Venn diagrams were created with the VennDiagram package, using the draw.triple.venn function. The differential abundance analysis of vOTUs across depth levels was performed using the likelihood ratio test implemented in DESeq2 [**128**]. Hierarchical clustering of the viral abundance patterns of the five viromes was done with the hclust function (method=complete), and heatmaps were created with the pheatmap and dendextend libraries. The world map was created with the maps library.

2.4.11. Detection of putative viral auxiliary metabolic genes (AMGs).

VIBRANT [**39**] and DRAM-v [**40**] were used to identify putative AMGs in SPRUCE vOTU sequences. Briefly, these tools consider gene annotation in order to identify genes in the input contigs (in this case, our vOTUs) that have predicted functions in cellular metabolism [**39**, **40**]. Since there is no standardized approach for AMG identification, we sought to compare results from both tools. VIBRANT was run (using standard settings) on all SPRUCE viral contigs that we had previously identified by either VirSorter or DeepVirFinder (n=2,802 vOTUs). Because DRAM-v requires VirSorter output,we could not use all of the DeepVirFinder-derived vOTUs. We re-ran the 4,326 SPRUCE vOTUs through VirSorter, resulting in 3,780 vOTUs, of which 2,645 also appeared in the VIBRANT output. DRAM-v was applied (using standard settings) to these 2,645 vOTUs. VIBRANT output was manually screened to determine whether predicted AMGs had viral genes upstream and downstream [**14**], and in many cases, they did not (see supplementary discussion). DRAM-v includes an analysis to assess the presence of viral genes upstream and downstream of the putative AMG, producing an ‘auxiliary score’ as a measure of confidence in the AMG prediction. From the DRAM-v output, only putative AMGs with auxiliary scores < 4 were retained (a low auxiliary score indicates a gene that is confidently viral), and no viral flag (F), transposon flag (T), viral-like peptidase (P), or attachment flag (A) could be present. Putative AMGs that did not have a gene ID or a gene description were also discarded. See supplemental discussion for more information.

Bibliography

[1] R M Wilson, A M Hopple, M M Tfaily, S D Sebestyen, C W Schadt, L Pfeifer-Meister, C Medvedeff, K J McFarlane, J E Kostka, M Kolton, R K Kolka, L A Kluber, J K Keller, T P Guilderson, N A Griffiths, J P Chanton, S D Bridgham, and P J Hanson. Stability of peatland carbon to rising temperatures. *Nat. Commun.*, 7:13723, December 2016.

[2] Alexander T Tveit, Tim Urich, and Mette M Svenning. Metatranscriptomic analysis of arctic peat soil microbiota. *Appl. Environ. Microbiol.*, 80(18):5761–5772, September 2014.

[3] Caitlin M Singleton, Carmody K McCalley, Ben J Woodcroft, Joel A Boyd, Paul N Evans, Suzanne B Hodgkins, Jeffrey P Chanton, Steve Frolking, Patrick M Crill, Scott R Saleska, Virginia I Rich, and Gene W Tyson. Methanotrophy across a natural permafrost thaw environment. *ISME J.*, 12(10):2544–2558, October 2018.

[4] Rhiannon Mondav, Ben J Woodcroft, Eun-Hae Kim, Carmody K McCalley, Suzanne B Hodgkins, Patrick M Crill, Jeffrey Chanton, Gregory B Hurst, Nathan C VerBerkmoes, Scott R Saleska, Philip Hugenholtz, Virginia I Rich, and Gene W Tyson. Discovery of a novel methanogen prevalent in thawing permafrost, 2014.

[5] E A G Schuur, A D McGuire, C Schädel, G Grosse, J W Harden, D J Hayes, G Hugelius, C D Koven, P Kuhry, D M Lawrence, S M Natali, D Olefeldt, V E Romanovsky, K Schaefer, M R Turetsky, C C Treat, and J E Vonk. Climate change and the permafrost carbon feedback. *Nature*, 520(7546):171–179, April 2015.

[6] Kurt E Williamson, Jeffry J Fuhrmann, K Eric Wommack, and Mark Radosevich. Viruses in soil ecosystems: An unknown quantity within an unexplored territory. *Annu Rev Virol*, 4(1):201–219, September 2017.

[7] Kurt E Williamson, K Eric Wommack, and Mark Radosevich. Sampling natural viral communities from soil for culture-independent analyses. *Appl. Environ. Microbiol.*, 69(11):6628–6633, November 2003.

[8] Gareth Trubl, Natalie Solonenko, Lauren Chittick, Sergei A Solonenko, Virginia I Rich, and Matthew B Sullivan. Optimization of viral resuspension methods for carbon-rich soils along a permafrost thaw gradient. *PeerJ*, 4:e1999, May 2016.

[9] Anja Narr, Ali Nawaz, Lukas Y Wick, Hauke Harms, and Antonis Chatzinotas. Soil viral communities vary temporally and along a land use transect as revealed by virus-like particle counting and a modified community fingerprinting approach (fRAPD). *Front. Microbiol.*, 8:1975, 2017.

[10] Paula Dalcin Martins, Robert E Danczak, Simon Roux, Jeroen Frank, Mikayla A Borton, Richard A Wolfe, Marie N Burris, and Michael J Wilkins. Viral and metabolic controls on high rates of microbial sulfur and carbon cycling in wetland ecosystems. *Microbiome*, 6(1):138, August 2018.

[11] Bonnie L Hurwitz and Jana M U'Ren. Viral metabolic reprogramming in marine ecosystems. *Curr. Opin. Microbiol.*, 31:161–168, June 2016.

[12] Simon Roux, Jennifer R Brum, Bas E Dutilh, Shinichi Sunagawa, Melissa B Duhaime, Alexander Loy, Bonnie T Poulos, Natalie Solonenko, Elena Lara, Julie Poulain, Stéphane Pesant, Stefanie Kandels-Lewis, Céline Dimier, Marc Picheral, Sarah Searson, Corinne Cruaud, Adriana Alberti, Carlos M Duarte, Josep M Gasol, Dolors Vaqué, Tara Oceans Coordinators, Peer Bork, Silvia G Acinas, Patrick Wincker, and Matthew B Sullivan. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, 537(7622):689–693, September 2016.

[13] Gareth Trubl, Ho Bin Jang, Simon Roux, Joanne B Emerson, Natalie Solonenko, Dean R Vik, Lindsey Solden, Jared Ellenbogen, Alexander T Runyon, Benjamin Bolduc, Ben J Woodcroft, Scott R Saleska, Gene W Tyson, Kelly C Wrighton, Matthew B Sullivan, and Virginia I Rich. Soil viruses are underexplored players in ecosystem carbon processing, 2018.

[14] Joanne B Emerson, Simon Roux, Jennifer R Brum, Benjamin Bolduc, Ben J Woodcroft, Ho Bin Jang, Caitlin M Singleton, Lindsey M Solden, Adrian E Naas, Joel A Boyd, Suzanne B Hodgkins, Rachel M Wilson, Gareth Trubl, Changsheng Li, Steve Frolking, Phillip B Pope, Kelly C Wrighton, Patrick M Crill, Jeffrey P Chanton, Scott R Saleska, Gene W Tyson, Virginia I Rich, and Matthew B Sullivan. Host-linked soil viral ecology along a permafrost thaw gradient. *Nature Microbiology*, 3(8):870–880, July 2018.

[15] Joanne B Emerson. Soil viruses: A new hope. *mSystems*, 4(3), May 2019.

[16] Ella T Sieradzki, J Cesar Ignacio-Espinoza, David M Needham, Erin B Fichot, and Jed A Fuhrman. Dynamic marine viral infections and major contribution to photosynthetic processes shown by spatiotemporal picoplankton metatranscriptomes. *Nat. Commun.*, 10(1):1169, March 2019.

[17] Mya Breitbart, Chelsea Bonnain, Kema Malki, and Natalie A Sawaya. Phage puppet masters of the marine microbial realm. *Nat Microbiol*, 3(7):754–766, July 2018.

[18] Jennifer R Brum and Matthew B Sullivan. Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat. Rev. Microbiol.*, 13(3):147–159, March 2015.

[19] Noah Fierer. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat. Rev. Microbiol.*, 15(10):579–590, October 2017.

[20] Akbar Adjie Pratama and Jan Dirk van Elsas. The 'neglected' soil virome - potential role and impact. *Trends Microbiol.*, 26(8):649–662, August 2018.

[21] Yakov Kuzyakov and Kyle Mason-Jones. Viruses in soil: Nano-scale undead drivers of microbial life, biogeochemical turnover and ecosystem functions. *Soil Biol. Biochem.*, 127:305–317, December 2018.

[22] Kurt E Williamson, Mark Radosevich, and K Eric Wommack. Abundance and diversity of viruses in six delaware soils. *Appl. Environ. Microbiol.*, 71(6):3119–3125, June 2005.

[23] Kurt E Williamson, Mark Radosevich, David W Smith, and K Eric Wommack. Incidence of lysogeny within temperate and extreme soil environments. *Environ. Microbiol.*, 9(10):2563–2574, October 2007.

[24] M M Swanson, G Fraser, T J Daniell, L Torrance, P J Gregory, and M Taliansky. Viruses in soils: morphological diversity and abundance in the rhizosphere. *Ann. Appl. Biol.*, 155(1):51–60, August 2009.

[25] Dhritiman Ghosh, Krishnakali Roy, Kurt E Williamson, Sharath Srinivasiah, K Eric Wommack, and Mark Radosevich. Acyl-homoserine lactones can induce virus production in lysogenic bacteria: an alternative paradigm for prophage induction. *Appl. Environ. Microbiol.*, 75(22):7142–7152, November 2009.

[26] Junjie Liu, Zhenhua Yu, Xinzhen Wang, Jian Jin, Xiaobing Liu, and Guanghua Wang. The distribution characteristics of the major capsid gene (g23) of t4-type phages in paddy floodwater in northeast china. *Soil Sci. Plant Nutr.*, 62(2):133–139, March 2016.

[27] Olivier Zablocki, Lonnie van Zyl, Evelien M Adriaenssens, Enrico Rubagotti, Marla Tuffin, Stephen Craig Cary, and Don Cowan. High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of antarctic soils. *Appl. Environ. Microbiol.*, 80(22):6888–6897, November 2014.

[28] Kurt E Williamson, Jennifer B Schnitker, Mark Radosevich, David W Smith, and K Eric Wommack. Cultivation-based assessment of lysogeny among soil bacteria. *Microb. Ecol.*, 56(3):437–447, October 2008.

[29] Dhritiman Ghosh, Krishnakali Roy, Kurt E Williamson, David C White, K Eric Wommack, Kerry L Sublette, and Mark Radosevich. Prevalence of lysogeny among soil bacteria and presence of 16S rRNA and trzn genes in viral-community DNA. *Appl. Environ. Microbiol.*, 74(2):495–502, January 2008.

[30] David Paez-Espino, Emiley A Eloe-Fadrosh, Georgios A Pavlopoulos, Alex D Thomas, Marcel Huntemann, Natalia Mikhailova, Edward Rubin, Natalia N Ivanova, and Nikos C Kyrpides. Uncovering earth's virome. *Nature*, 536(7617):425–430, August 2016.

[31] Evan P Starr, Erin E Nuccio, Jennifer Pett-Ridge, Jillian F Banfield, and Mary K Firestone. Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. *Proc. Natl. Acad. Sci. U. S. A.*, 116(51):25900–25908, December 2019.

[32] Joshua M A Stough, Max Kolton, Joel E Kostka, David J Weston, Dale A Pelletier, and Steven W Wilhelm. Diversity of active viral infections within the sphagnum microbiome. *Appl. Environ. Microbiol.*, 84(23), December 2018.

[33] Gareth Trubl, Simon Roux, Natalie Solonenko, Yueh-Fen Li, Benjamin Bolduc, Josué Rodríguez-Ramos, Emiley A Eloe-Fadrosh, Virginia I Rich, and Matthew B Sullivan. Towards optimized viral metagenomes for double-stranded and single-stranded DNA viruses from challenging soils. *PeerJ*, 7:e7265, July 2019.

[34] Christian Santos-Medellin, Laura A Zinke, Anneliek M ter Horst, Danielle L Gelardi, Sanjai J Parikh, and Joanne B Emerson. Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *The ISME Journal*, 15(7):1956–1970, 2021.

[35] Pauline C Göller, Jose M Haro-Moreno, Francisco Rodriguez-Valera, Martin J Loessner, and Elena Gómez-Sanz. Uncovering a hidden diversity: optimized protocols for the extraction of dsDNA bacteriophages from soil, 2020.

[36] Gareth Trubl, Paul Hyman, Simon Roux, and Stephen T Abedon. Coming-of-Age characterization of soil viruses: A user's guide to virus isolation, detection within metagenomes, and viromics, 2020.

[37] Simon Roux, Francois Enault, Bonnie L Hurwitz, and Matthew B Sullivan. VirSorter: mining viral signal from microbial genomic data. *PeerJ*, 3:e985, May 2015.

[38] Jie Ren, Kai Song, Chao Deng, Nathan A Ahlgren, Jed A Fuhrman, Yi Li, Xiaohui Xie, Ryan Poplin, and Fengzhu Sun. Identifying viruses from metagenomic data using deep learning. *Quantitative Biology*, 8(1):64–77, March 2020.

[39] Kristopher Kieft, Zhichao Zhou, and Karthik Anantharaman. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, 8(1):90, June 2020.

[40] Michael Shaffer, Mikayla A Borton, Bridget B McGivern, Ahmed A Zayed, Sabina Leanti La Rosa, Lindsey M Solden, Pengfei Liu, Adrienne B Narrowe, Josué Rodríguez-Ramos, Benjamin Bolduc, M Consuelo Gazitúa, Rebecca A Daly, Garrett J Smith, Dean R Vik, Phil B Pope, Matthew B Sullivan, Simon Roux, and Kelly C Wrighton. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.*, August 2020.

[41] Rachel Mackelprang, Scott R Saleska, Carsten Suhr Jacobsen, Janet K Jansson, and Neslihan Taş. Permafrost Meta-Omics and climate change. *Annu. Rev. Earth Planet. Sci.*, 44(1):439–462, June 2016.

[42] Janet K Jansson and Neslihan Taş. The microbial ecology of permafrost. *Nat. Rev. Microbiol.*, 12(6):414–425, June 2014.

[43] Ben J Woodcroft, Caitlin M Singleton, Joel A Boyd, Paul N Evans, Joanne B Emerson, Ahmed A F Zayed, Robert D Hoelzle, Timothy O Lamberton, Carmody K McCalley, Suzanne B Hodgkins, Rachel M Wilson, Samuel O Purvine, Carrie D Nicora, Changsheng Li, Steve Frolking, Jeffrey P Chanton, Patrick M Crill, Scott R Saleska, Virginia I Rich, and Gene W Tyson. Genome-centric view of carbon processing in thawing permafrost. *Nature*, 560(7716):49–54, August 2018.

[44] Xueju Lin, Malak M Tfaily, J Megan Steinweg, Patrick Chanton, Kaitlin Esson, Zamin K Yang, Jeffrey P Chanton, William Cooper, Christopher W Schadt, and Joel E Kostka. Microbial community stratification linked to utilization of carbohydrates and phosphorus limitation in a boreal peatland at marcell experimental forest, minnesota, USA. *Appl. Environ. Microbiol.*, 80(11):3518–3530, June 2014.

[45] Richard J Norby, Joanne Childs, Paul J Hanson, and Jeffrey M Warren. Rapid loss of an ecosystem engineer: Sphagnum decline in an experimentally warmed bog. *Ecol. Evol.*, 9(22):12571–12585, November 2019.

[46] Paul J Hanson, Jeffery S Riggs, W Robert Nettles, Jana R Phillips, Misha B Krassovski, Leslie A Hook, Lianhong Gu, Andrew D Richardson, Donald M Aubrecht, Daniel M Ricciuto, Jeffrey M Warren, and Charlotte Barbier. Attaining whole-ecosystem warming using air and deep-soil heating methods with an elevated CO_2 atmosphere. *Biogeosciences*, 14(4):861–883, February 2017.

[47] Nancy B Dise, Eville Gorham, and Elon S Verry. Environmental factors controlling methane emissions from peatlands in northern minnesota. *J. Geophys. Res.*, 98(D6):10583, 1993.

[48] Randall Kolka, Stephen Sebestyen, Elon S Verry, and Kenneth Brooks. *Peatland Biogeochemistry and Watershed Hydrology at the Marcell Experimental Forest.* CRC Press, February 2011.

[49] D F Grigal. Elemental dynamics in forested bogs in northern minnesota. *Can. J. Bot.*, 69(3):539–546, March 1991.

[50] Dale S Nichols and James M Brown. Evaporation from a sphagnum moss surface. *J. Hydrol.*, 48(3):289–302, November 1980.

[51] Elon S Verry and D R Timmons. Waterborne nutrient flow through an Upland-Peatland watershed in minnesota, 1982.

[52] Don H Boelter and Elon S Verry. *Peatland and Water in the Northern Lake States.* Department of Agriculture, Forest Service, North Central Forest Experiment Station, 1977.

[53] Andrew D Richardson, Koen Hufkens, Thomas Milliman, Donald M Aubrecht, Morgan E Furze, Bijan Seyednasrollah, Misha B Krassovski, John M Latimer, W Robert Nettles, Ryan R Heiderman, Jeffrey M Warren, and Paul J Hanson. Ecosystem warming extends vegetation activity but heightens vulnerability to cold temperatures. *Nature*, 560(7718):368–371, August 2018.

[54] Christopher W Fernandez, Katherine Heckman, Randall Kolka, and Peter G Kennedy. Melanin mitigates the accelerated decay of mycorrhizal necromass with peatland warming. *Ecol. Lett.*, 22(3):498–505, March 2019.

[55] Mara Y McPartland, Evan S Kane, Michael J Falkowski, Randy Kolka, Merritt R Turetsky, Brian Palik, and Rebecca A Montgomery. The response of boreal peatland community composition and NDVI to hydrologic change, warming, and elevated carbon dioxide. *Glob. Chang. Biol.*, 25(1):93–107, January 2019.

[56] A M Hopple, R M Wilson, M Kolton, C A Zalman, J P Chanton, J Kostka, P J Hanson, J K Keller, and S D Bridgham. Massive peatland carbon banks vulnerable to rising temperatures. *Nat. Commun.*, 11(1):2373, May 2020.

[57] Alyssa A Carrell, Max Kolton, Jennifer B Glass, Dale A Pelletier, Melissa J Warren, Joel E Kostka, Colleen M Iversen, Paul J Hanson, and David J Weston. Experimental warming alters the community composition, diversity, and N2 fixation activity of peat moss (sphagnum fallax) microbiomes. *Glob. Chang. Biol.*, 25(9):2993–3004, September 2019.

[58] Melissa J Warren, Xueju Lin, John C Gaby, Cecilia B Kretz, Max Kolton, Peter L Morton, Jennifer Pett-Ridge, David J Weston, Christopher W Schadt, Joel E Kostka, and Jennifer B Glass. Molybdenum-Based diazotrophy in a sphagnum peatland in northern minnesota, 2017.

[59] Laurel A Kluber, Eric R Johnston, Samantha A Allen, J Nicholas Hendershot, Paul J Hanson, and Christopher W Schadt. Constraints on microbial communities, decomposition and methane production in deep peat deposits. *PLoS One*, 15(2):e0223744, February 2020.

[60] Simon Roux, Evelien M Adriaenssens, Bas E Dutilh, Eugene V Koonin, Andrew M Kropinski, Mart Krupovic, Jens H Kuhn, Rob Lavigne, J Rodney Brister, Arvind Varsani, Clara Amid, Ramy K Aziz, Seth R Bordenstein, Peer Bork, Mya Breitbart, Guy R Cochrane, Rebecca A Daly, Christelle Desnues, Melissa B Duhaime, Joanne B Emerson, François Enault, Jed A Fuhrman, Pascal Hingamp, Philip Hugenholtz, Bonnie L Hurwitz, Natalia N Ivanova, Jessica M Labonté, Kyung-Bum Lee, Rex R Malmstrom, Manuel Martinez-Garcia, Ilene Karsch Mizrachi, Hiroyuki Ogata, David Páez-Espino, Marie-Agnès Petit, Catherine Putonti, Thomas Rattei, Alejandro Reyes, Francisco Rodriguez-Valera, Karyna Rosario, Lynn Schriml, Frederik Schulz, Grieg F Steward, Matthew B Sullivan, Shinichi Sunagawa, Curtis A Suttle, Ben Temperton, Susannah G Tringe, Rebecca Vega Thurber, Nicole S Webster, Katrine L Whiteson, Steven W Wilhelm, K Eric Wommack, Tanja Woyke, Kelly C Wrighton, Pelin Yilmaz, Takashi Yoshida, Mark J Young, Natalya Yutin, Lisa Zeigler Allen, Nikos C Kyrpides, and Emiley A Eloe-Fadrosh. Minimum information about an uncultivated virus genome (MIUViG). *Nat. Biotechnol.*, 37(1):29–37, January 2019.

[61] David Paez-Espino, I-Min A Chen, Krishna Palaniappan, Anna Ratner, Ken Chu, Ernest Szeto, Manoj Pillay, Jinghua Huang, Victor M Markowitz, Torben Nielsen, Marcel Huntemann, T B K Reddy, Georgios A Pavlopoulos, Matthew B Sullivan, Barbara J Campbell, Feng Chen, Katherine McMahon, Steve J Hallam, Vincent Denef, Ricardo Cavicchioli, Sean M Caffrey, Wolfgang R Streit, John Webster, Kim M Handley, Ghasem H Salekdeh, Nicolas Tsesmetzis, Joao C

Setubal, Phillip B Pope, Wen-Tso Liu, Adam R Rivers, Natalia N Ivanova, and Nikos C Kyrpides. IMG/VR: a database of cultured and uncultured DNA viruses and retroviruses. *Nucleic Acids Res.*, 45(D1):D457–D465, January 2017.

[62] Simon Roux, Steven J Hallam, Tanja Woyke, and Matthew B Sullivan. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *Elife*, 4:e08490, July 2015.

[63] Ann C Gregory, Ahmed A Zayed, Nádia Conceição-Neto, Ben Temperton, Ben Bolduc, Adriana Alberti, Mathieu Ardyna, Ksenia Arkhipova, Margaux Carmichael, Corinne Cruaud, Céline Dimier, Guillermo Domínguez-Huerta, Joannie Ferland, Stefanie Kandels, Yunxiao Liu, Claudie Marec, Stéphane Pesant, Marc Picheral, Sergey Pisarev, Julie Poulain, Jean-Éric Tremblay, Dean Vik, Tara Oceans Coordinators, Marcel Babin, Chris Bowler, Alexander I Culley, Colomban de Vargas, Bas E Dutilh, Daniele Iudicone, Lee Karp-Boss, Simon Roux, Shinichi Sunagawa, Patrick Wincker, and Matthew B Sullivan. Marine DNA viral macro- and microdiversity from pole to pole. *Cell*, 177(5):1109–1123.e14, May 2019.

[64] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, 35(Database issue):D61–5, January 2007.

[65] Simon Roux, Gareth Trubl, Danielle Goudeau, Nandita Nath, Estelle Couradeau, Nathan A Ahlgren, Yuanchao Zhan, David Marsan, Feng Chen, Jed A Fuhrman, Trent R Northen, Matthew B Sullivan, Virginia I Rich, Rex R Malmstrom, and Emiley A Eloe-Fadrosh. Optimizing de novo genome assembly from PCR-amplified metagenomes. *PeerJ*, 7:e6902, May 2019.

[66] Xiaolong Liang, Regan E Wagner, Jie Zhuang, Jennifer M DeBruyn, Steven W Wilhelm, Fang Liu, Lu Yang, Margaret E Staton, Andrew C Sherfy, and Mark Radosevich. Viral abundance and diversity vary with depth in a southeastern united states agricultural ultisol. *Soil Biol. Biochem.*, 137:107546, October 2019.

[67] Carmody K McCalley, Ben J Woodcroft, Suzanne B Hodgkins, Richard A Wehr, Eun-Hae Kim, Rhiannon Mondav, Patrick M Crill, Jeffrey P Chanton, Virginia I Rich, Gene W Tyson, and Scott R Saleska. Methane dynamics regulated by microbial community response to permafrost thaw. *Nature*, 514(7523):478–481, October 2014.

[68] Suzanne B Hodgkins, Jeffrey P Chanton, Lauren C Langford, Carmody K McCalley, Scott R Saleska, Virginia I Rich, Patrick M Crill, and William T Cooper. Soil incubations reproduce field methane dynamics in a subarctic wetland. *Biogeochemistry*, 126(1):241–249, November 2015.

[69] Erik A Hobbie, Janet Chen, Paul J Hanson, Colleen M Iversen, Karis J McFarlane, Nathan R Thorp, and Kirsten S Hofmockel. Long-term carbon and nitrogen dynamics at SPRUCE revealed through stable isotopes in peat profiles, 2017.

[70] Connor T Skennerton, Michael Imelfort, and Gene W Tyson. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Res.*, 41(10):e105, May 2013.

[71] Paul N Evans, Joel A Boyd, Andy O Leu, Ben J Woodcroft, Donovan H Parks, Philip Hugenholtz, and Gene W Tyson. An evolving view of methane metabolism in the archaea. *Nat. Rev. Microbiol.*, 17(4):219–232, April 2019.

[72] L A Zinke, P N Evans, A Schroeder, D H Parks, and others. Evidence for non-methanogenic metabolisms in globally distributed archaeal clades basal to the methanomassiliicoccales. *bioRxiv*, 2020.

[73] Min Jin, Xun Guo, Rui Zhang, Wu Qu, Boliang Gao, and Runying Zeng. Diversities and potential biogeochemical impacts of mangrove soil viruses. *Microbiome*, 7(1):58, April 2019.

[74] Andrea Du Toit. Permafrost thawing and carbon metabolism. *Nat. Rev. Microbiol.*, 16(9):519, September 2018.

[75] Kristen M DeAngelis, Grace Pold, Begüm D Topçuoğlu, Linda T A van Diepen, Rebecca M Varney, Jeffrey L Blanchard, Jerry Melillo, and Serita D Frey. Long-term forest soil warming alters microbial communities in temperate forest soils. *Front. Microbiol.*, 6:104, February 2015.

[76] Hui Wang, Shirong Liu, Andreas Schindlbacher, Jingxin Wang, Yujing Yang, Zhanchao Song, Yeming You, Zuomin Shi, Zhaoying Li, Lin Chen, Angang Ming, Lihua Lu, and Daoxiong Cai. Experimental warming reduced topsoil carbon content and increased soil bacterial diversity in a subtropical planted forest. *Soil Biol. Biochem.*, 133:155–164, June 2019.

[77] Max Kolton, Ansley Marks, Rachel M Wilson, Jeffrey P Chanton, and Joel E Kostka. Impact of warming on greenhouse gas production and microbial diversity in anoxic peat from a Sphagnum-Dominated bog (grand rapids, minnesota, united states). *Front. Microbiol.*, 10:870, April 2019.

[78] Adam C Martiny, Kathleen Treseder, and Gordon Pusch. Phylogenetic conservatism of functional traits in microorganisms. *ISME J.*, 7(4):830–838, April 2013.

[79] Susanne A Fritz and Andy Purvis. Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits, 2010.

[80] Benjamin Bolduc, Ho Bin Jang, Guilhem Doulcier, Zhi-Qiang You, Simon Roux, and Matthew B Sullivan. vConTACT: an ivirus tool to classify double-stranded DNA viruses that infect archaea and bacteria. *PeerJ*, 5:e3243, May 2017.

[81] Ho Bin Jang, Benjamin Bolduc, Olivier Zablocki, Jens H Kuhn, Simon Roux, Evelien M Adriaenssens, J Rodney Brister, Andrew M Kropinski, Mart Krupovic, Rob Lavigne, Dann Turner, and Matthew B Sullivan. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.*, 37(6):632–639, June 2019.

[82] Evelien M Adriaenssens, Matthew B Sullivan, Petar Knezevic, Leonardo J van Zyl, B L Sarkar, Bas E Dutilh, Poliane Alfenas-Zerbini, Małgorzata Łobocka, Yigang Tong, James Rodney Brister, Andrea I Moreno Switt, Jochen Klumpp, Ramy Karam Aziz, Jakub Barylski, Jumpei Uchiyama, Rob A Edwards, Andrew M Kropinski, Nicola K Petty, Martha R J Clokie, Alla I Kushkina, Vera V Morozova, Siobain Duffy, Annika Gillis, Janis Rumnieks, İpek Kurtböke, Nina Chanishvili, Lawrence Goodridge, Johannes Wittmann, Rob Lavigne, Ho Bin Jang, David Prangishvili, Francois Enault,

Dann Turner, Minna M Poranen, Hanna M Oksanen, and Mart Krupovic. Taxonomy of prokaryotic viruses: 2018-2019 update from the ICTV bacterial and archaeal viruses subcommittee. *Arch. Virol.*, 165(5):1253–1260, May 2020.

[83] Ann C Gregory, Olivier Zablocki, Ahmed A Zayed, Allison Howell, Benjamin Bolduc, and Matthew B Sullivan. The gut virome database reveals Age-Dependent patterns of virome diversity in the human gut. *Cell Host Microbe*, August 2020.

[84] Konstantinos T Konstantinidis, Ramon Rosselló-Móra, and Rudolf Amann. Uncultivated microbes in need of their own taxonomy. *ISME J.*, 11(11):2399–2406, November 2017.

[85] Catherine A Lozupone and Rob Knight. Global patterns in bacterial diversity. *Proc. Natl. Acad. Sci. U. S. A.*, 104(27):11436–11440, July 2007.

[86] Alexa M Nicolas, Alexander L Jaffe, Erin E Nuccio, Michiko E Taga, Mary K Firestone, and Jillian F Banfield. Unexpected diversity of CPR bacteria and nanoarchaea in the rare biosphere of rhizosphere-associated grassland soil. *bioRxiv*, page 2020.07.13.194282, July 2020.

[87] Shannon J Williamson, Lisa Zeigler Allen, Hernan A Lorenzi, Douglas W Fadrosh, Daniel Brami, Mathangi Thiagarajan, John P McCrow, Andrey Tovchigrechko, Shibu Yooseph, and J Craig Venter. Metagenomic exploration of viruses throughout the indian ocean. *PLoS One*, 7(10):e42047, October 2012.

[88] Joanne B Emerson, Karen Andrade, Brian C Thomas, Anders Norman, Eric E Allen, Karla B Heidelberg, and Jillian F Banfield. Virus-host and CRISPR dynamics in archaea-dominated hypersaline lake tyrrell, victoria, australia. *Archaea*, 2013:370871, June 2013.

[89] Joanne B Emerson, Brian C Thomas, Karen Andrade, Eric E Allen, Karla B Heidelberg, and Jillian F Banfield. Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. *Appl. Environ. Microbiol.*, 78(17):6309–6320, September 2012.

[90] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, August 2014.

[91] B Bushnell. BBTools software package. *URL http://sourceforge. net/projects/bbmap*, 2014.

[92] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31(10):1674–1676, May 2015.

[93] Brian Bushnell, Jonathan Rood, and Esther Singer. BBMerge – accurate paired shotgun read merging via overlap, 2017.

[94] Yu Peng, Henry C M Leung, S M Yiu, and Francis Y L Chin. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428, June 2012.

[95] Ying Huang, Beifang Niu, Ying Gao, Limin Fu, and Weizhong Li. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5):680–682, March 2010.

[96] Brian Bushnell. BBMap: a fast, accurate, splice-aware aligner. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2014.

[97] Simon Roux, Joanne B Emerson, Emiley A Eloe-Fadrosh, and Matthew B Sullivan. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ*, 5:e3817, September 2017.

[98] BamM. BamM - working with the BAM. `http://ecogenomics.github.io/BamM/`. Accessed: 2020-10-13.

[99] Aaron R Quinlan. BEDTools: the swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics*, 47(1):11–12, 2014.

[100] Jie Ren, Nathan A Ahlgren, Yang Young Lu, Jed A Fuhrman, and Fengzhu Sun. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, 5(1):69, July 2017.

[101] Donovan Parks. CompareM.

[102] Dongwan D Kang, Jeff Froula, Rob Egan, and Zhong Wang. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities, 2015.

[103] Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, 25(7):1043–1055, July 2015.

[104] Jackson W Sorensen, Taylor K Dunivin, Tammy C Tobin, and Ashley Shade. Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. *Nat Microbiol*, 4(1):55–61, January 2019.

[105] Eric R Johnston, Janet K Hatt, Zhili He, Liyou Wu, Xue Guo, Yiqi Luo, Edward A G Schuur, James M Tiedje, Jizhong Zhou, and Konstantinos T Konstantinidis. Responses of tundra soil microbial communities to half a decade of experimental warming at two critical depths. *Proc. Natl. Acad. Sci. U. S. A.*, 116(30):15096–15105, July 2019.

[106] Yu-Wei Wu, Blake A Simmons, and Steven W Singer. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4):605–607, February 2016.

[107] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9(4):357–359, March 2012.

[108] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[109] Donovan H Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J Woodcroft, Paul N Evans, Philip Hugenholtz, and Gene W Tyson. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol*, 2(11):1533–1542, November 2017.

[110] Matthew R Olm, Christopher T Brown, Brandon Brooks, and Jillian F Banfield. drep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.*, 11(12):2864–2868, December 2017.

[111] Pierre-Alain Chaumeil, Aaron J Mussig, Philip Hugenholtz, and Donovan H Parks. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics*, November 2019.

[112] Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.*, 36(10):996–1004, November 2018.

[113] Laura A Zinke, Paul N Evans, Christian Santos-Medellín, Alena L Schroeder, Donovan H Parks, Ruth K Varner, Virginia I Rich, Gene W Tyson, and Joanne B Emerson. Evidence for non-methanogenic metabolisms in globally distributed archaeal clades basal to the methanomassiliicoccales. *Environ. Microbiol.*, 23(1):340–357, January 2021.

[114] Doug Hyatt, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11:119, March 2010.

[115] Sean R Eddy. Accelerated profile HMM searches. *PLoS Comput. Biol.*, 7(10):e1002195, October 2011.

[116] David Burstein, Lucas B Harrington, Steven C Strutt, Alexander J Probst, Karthik Anantharaman, Brian C Thomas, Jennifer A Doudna, and Jillian F Banfield. New CRISPR-Cas systems from uncultivated microbes. *Nature*, 542(7640):237–241, February 2017.

[117] Sudhir Kumar, Glen Stecher, Michael Li, Christina Knyaz, and Koichiro Tamura. MEGA x: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.*, 35(6):1547–1549, June 2018.

[118] Laura A Hug, Brett J Baker, Karthik Anantharaman, Christopher T Brown, Alexander J Probst, Cindy J Castelle, Cristina N Butterfield, Alex W Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A Relman, Kari M Finstad, Ronald Amundson, Brian C Thomas, and Jillian F Banfield. A new view of the tree of life. *Nat Microbiol*, 1:16048, April 2016.

[119] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, July 2014.

[120] Robert C Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, October 2010.

[121] Mang Shi, Xian-Dan Lin, Jun-Hua Tian, Liang-Jun Chen, Xiao Chen, Ci-Xiu Li, Xin-Cheng Qin, Jun Li, Jian-Ping Cao, John-Sebastian Eden, Jan Buchmann, Wen Wang, Jianguo Xu, Edward C Holmes, and Yong-Zhen Zhang. Redefining the invertebrate RNA virosphere. *Nature*, 540(7634):539–543, December 2016.

[122] Kazutaka Katoh and Daron M Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, 30(4):772–780, April 2013.

[123] Salvador Capella-Gutiérrez, José M Silla-Martínez, and Toni Gabaldón. trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973, August 2009.

[124] Federico Abascal, Rafael Zardoya, and David Posada. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, 21(9):2104–2105, May 2005.

[125] Lam-Tung Nguyen, Heiko A Schmidt, Arndt von Haeseler, and Bui Quang Minh. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, 32(1):268–274, January 2015.

[126] Ivica Letunic and Peer Bork. Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.*, 47(W1):W256–W259, July 2019.

[127] Jari Oksanen, F Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R Minchin, R B O'hara, Gavin L Simpson, Peter Solymos, and Others. vegan: Community ecology package. R package version 2.4-3. *Vienna: R Foundation for Statistical Computing. [Google Scholar]*, 2016.

[128] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12):550, 2014.

Figures and tables

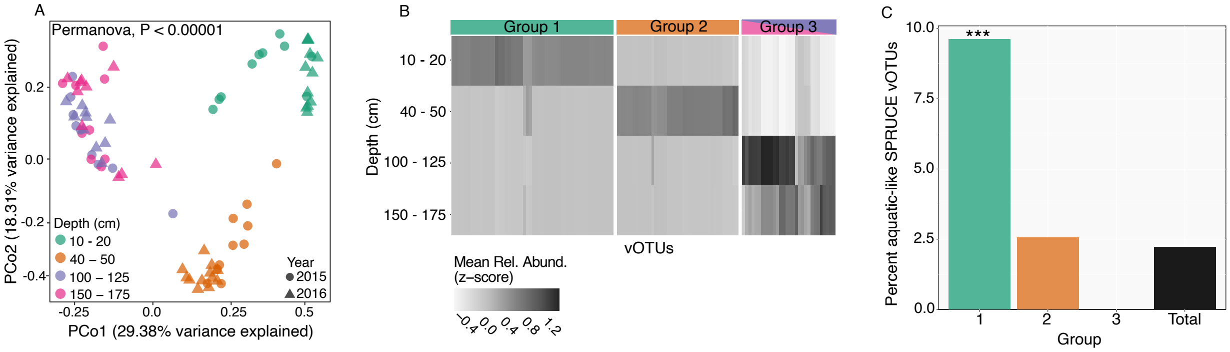


Figure 2.1: Peat viral community and population (vOTU) abundance patterns with depth in the SPRUCE experimental plots. A: Principal coordinates analysis (PCoA) of viral community composition in 82 samples (total soil metagenomes) from peat bog soil from the Marcell Experimental Forest in northern Minnesota (USA) collected from the SPRUCE experimental plots and chambers (temperature treatmentsranging from ambient to +9 °C above ambient), based on Bray-Curtis dissimilarities derived from the table of vOTU abundances (read mapping to vOTUs, n=2,699). Each point is one sample (n=82). **B:** Mean relative abundances (Z- transformed) of vOTUs significantly differentially abundant by depth (adjusted-p<0.05, Likelihood Ratio Test). Groups were identified through hierarchical clustering and are colored according to the depths in panel A. **C:** Percentage of vOTUs classified as "aquatic-like" in each of the groups identified in panel B (Groups 1-3) and in the whole dataset of 2,699 vOTUs (Total). SPRUCE vOTUs were considered "aquatic-like" if they shared a genus-level viral cluster (VC) with at least one vOTU from a marine or freshwater habitat in the PIGEON database. Note that the y-axis maximum is 10%. *** denotes a significantly larger proportion of aquatic-like vOTUs in that group, relative to the proportion of aquatic-like vOTUs in the full SPRUCE dataset (Total) (P < 0.05, Hypergeometric test)

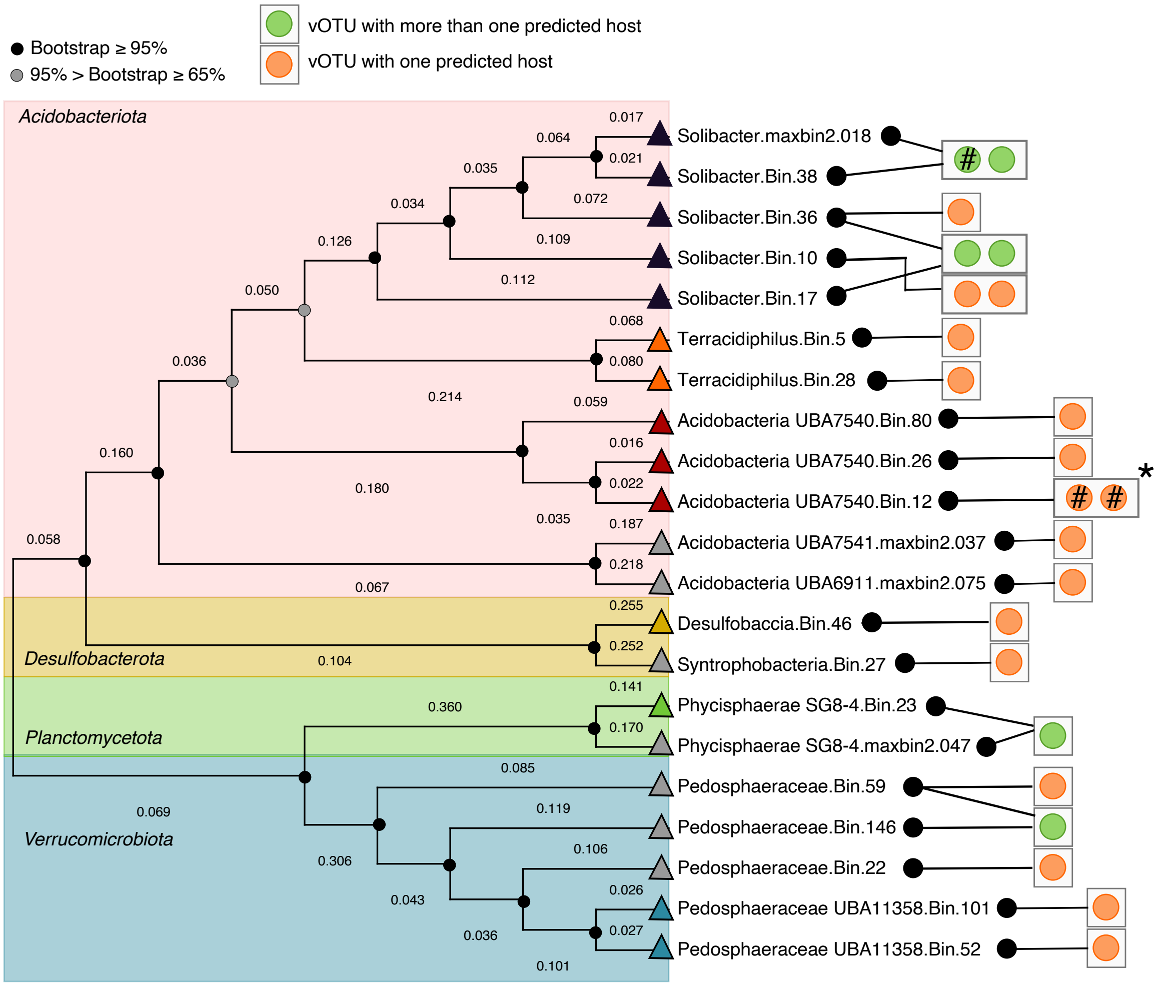


Figure 2.2: SPRUCE virus-host linkages according to host phylogeny. Unrooted phylogenetic tree (concatenated predicted protein alignment of 43 marker genes defined by CheckM [109]) of microbial host metagenome-assembled genomes (MAGs) with at least one vOTU (green and orange circles) linked via CRISPR sequence homology. Branch represent the expected number of substitutions per site. Lines between black circles and squares with orange or green circles link vOTUs to predicted host MAGs. Colored triangles indicate the MAG genus (the same color is the same genus, except for grey triangles, for which the corresponding MAG could only be classified to the family level). Asterisk indicates vOTUs in the same genus-level viral cluster (VC); remaining vOTUs were all in distinct VCs. Bootstrap support values are shown as circles on nodes, black circles indicate support ¿= 95%, grey indicates support between 65 and 95%. A pound sign inside an orange or green circle indicates a one-nucleotide CRISPR spacer-protospacer mismatch.

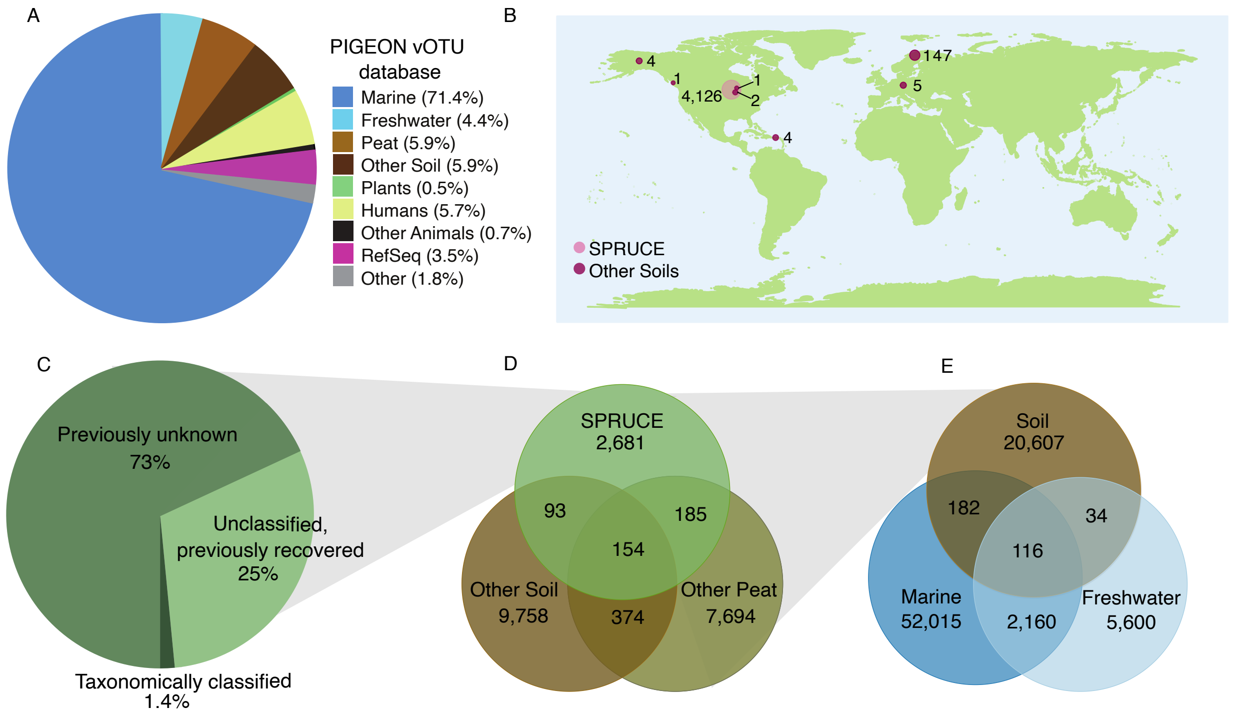


Figure 2.3: Habitat and global distribution of SPRUCE vOTUs and viral clusters (VCs), using the PIGEON database for context. **A:** Composition of the PIGEON database of vOTUs (n=266,805) by source environment. RefSeq includes isolate viral genomes from a variety of source environments (prokaryotic viruses in RefSeq v95). Plants = plant-associated, Humans = humanassociated, Other Animals = non-human animal-associated. **B:** vOTUs (n=4,326) recovered from SPRUCE peat by read mapping, according to the location from which they were first recovered. Numbers indicate SPRUCE vOTUs from a given location. Circle sizes are proportional to the number of vOTUs. **C:** Percentages of vOTUs recovered from SPRUCE that: had predicted taxonomy based on clustering with RefSeq viral genomes (Taxonomically classified), had unknown taxonomy but shared a genus-level viral cluster (VC) with one or more previously recovered vOTUs in the PIGEON database (Unclassified, previously recovered), or were previously unknown at the VC (genus) level (Previously unknown). **D:** Habitat(s) for each soil VC (n=20,939) in the PIGEON database, based on source habitat(s) for the vOTU(s) contained in each VC. For a given soil VC, either all vOTUs were exclusively derived from a single habitat (non-overlapping regions), or two or more vOTUs were derived from different soil habitats (overlapping regions). **E:** Similar to D, but for VCs with vOTUs from soil, marine, and/or freshwater habitats (n=80,714 VCs).

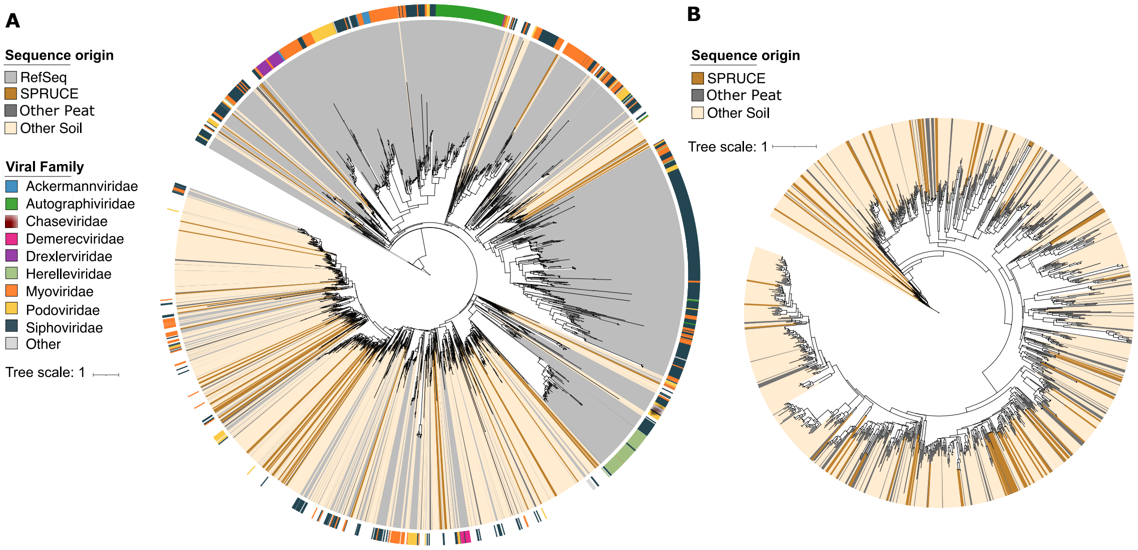


Figure 2.4: Unrooted phylogenetic trees of terminase large subunit (TerL) protein sequences from RefSeq prokaryotic viral genomes and soil vOTUs in the PIGEON database. Trees are color-coded by sequence source (RefSeq or soil category within PIGEON). Trees were constructed using IQ-tree and the LG+I+G4+F model of sequence evolution, using ultrafast bootstrapping and an SH-aLRT test. Bootstrap values are not displayed but can be found for each of the branches in Supplemental File 1. A: Phylogenetic tree of TerL protein sequences from RefSeq prokaryotic viral genomes (n=1,613) and PIGEON soil vOTUs (n=1,011). Outer ring color represents viral family of RefSeq genomes. Phylogenetic dispersion was estimated by using Fritz and Purvis D (D). D=-0.25 when comparing TerL sequences from RefSeq viral genomes and TerL sequences from soil vOTUs, with D $<$ 0 indicating phylogenetic clustering. B: Phylogenetic tree of TerL protein sequences from PIGEON soil vOTUs. D=0.58 for other soil (n=634) compared to peat, including SPRUCE (n=377), and D=0.41 when comparing SPRUCE (n=51) to all other soil sequences (n=960).

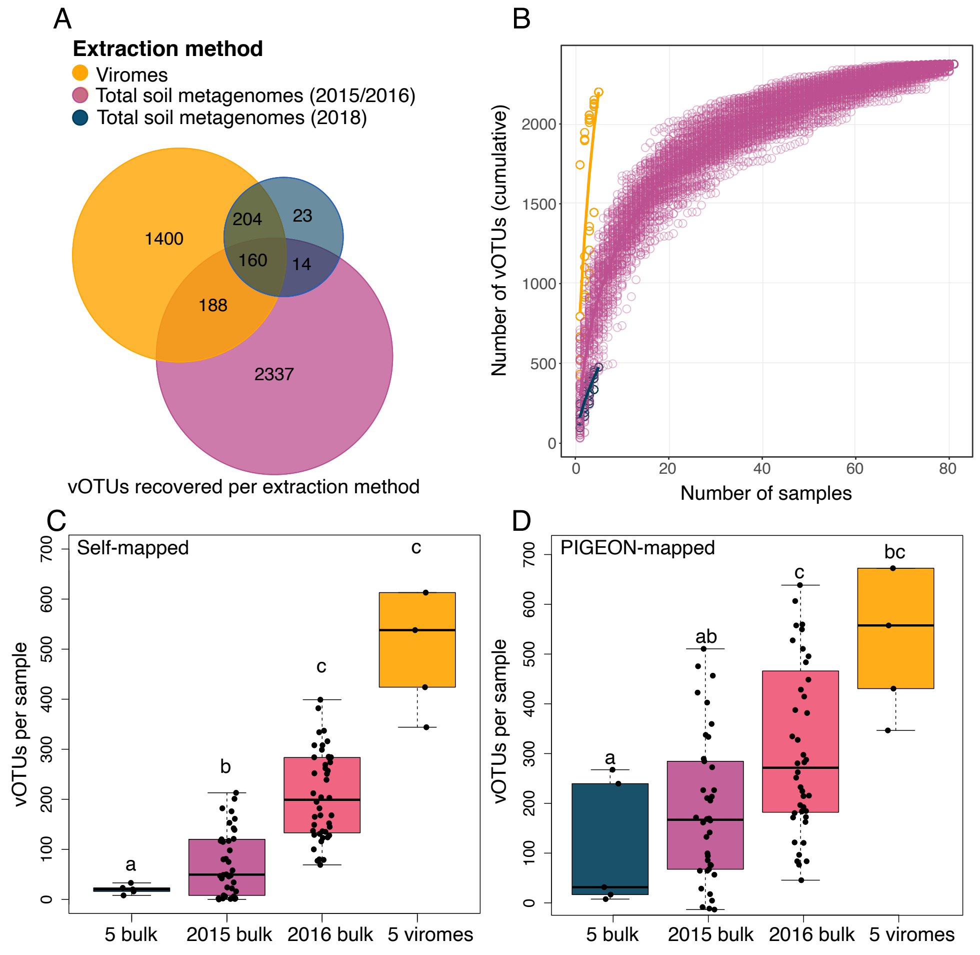


Figure 2.5: Comparison of vOTU recovery from SPRUCE viromes and total soil metagenomes. **A:** Distribution of vOTUs recovered in each of three extraction groups (grouped by extraction method and collection date), based on read mapping to the PIGEON database (n=5 viromes from 2018, 82 total soil metagenomes from 2015 and 2016, and 5 total soil metagenomes from 2018). **B:** Accumulation curves of distinct vOTUs recovered as sampling increases for each extraction method; 100 permutations of sample order are depicted as open circles, line shows the average of the permutations for each method. **C:** Number of vOTUs recovered per metagenome when reads were only allowed to map to vOTUs that assembled from metagenomes in the same category (self-mapped), considering four categories: 2018 bulk (n=5), 2015 bulk (n=38), 2016 bulk (n=44), 2018 viromes (n=5); bulk = total soil metagenomes. One outlier was excluded from the plot for ease of visualization; the y-axis value of the outlier in the 2018 viromes was 1,328. Letters above boxes correspond to significant differences between groups (Student's T-test, significant when $p < 0.05$). **D:** Similar to C, but reads were allowed to map to all vOTUs in the PIGEON database (PIGEON-mapped), including all vOTUs assembled from any of the SPRUCE metagenomes. Three outliers were removed from the plot for ease of visualization; the y-axis values of the two outliers from 2016 bulk were 1,415 and 1,818, and the value of the outlier from the 2018 viromes was 1,558.

CHAPTER 3

RNA viral communities are structured by host plant phylogeny in oak and conifer leaves

Anneliek M. ter Horst[1], Jane D. Fudyma[1], Aurélie Bak[1], Min Sook Hwang[1,2], Christian Santos-Medellín[1], Kristian A. Stevens[1,2], David M. Rizzo[1], Maher Al Rwahnih[1,2], and Joanne B. Emerson* [1]

[1] Department of Plant Pathology, University of California, Davis, Davis, CA, USA

[2] Foundation Plant Services, University of California, Davis, Davis, CA, USA

Abstract

Wild plants can suffer devastating diseases, experience asymptomatic, persistent infections, and serve as reservoirs for viruses of agricultural crops, yet we have a limited understanding of the natural plant virosphere. To access representatives of locally and globally distinct wild plants and investigate their viral diversity, we extracted and sequenced dsRNA from leaves from 16 healthy oak and conifer trees in the UC Davis Arboretum (Davis, California). From de novo assemblies, we recovered 389 RNA-dependent RNA polymerase (RdRp) gene sequences from 384 putative viral species, and identified 580 putative viral contigs via a virus prediction software followed by manual confirmation of virus annotation. Based on similarity to known viruses, most recovered viruses were predicted to infect plants or fungi, with the highest diversity and abundance observed in the *Totiviridae* and *Mitoviridae* families. Phyllosphere viral community composition differed significantly by host plant phylogeny, suggesting the potential for host-specific viromes. The phyllosphere viral community of one oak tree differed substantially from other oak viral communities and contained a greater proportion of putative mycoviral sequences, potentially due to the tree's more

advanced senescence at the time of sampling. These results suggest that oaks and conifers harbor a vast diversity of viruses with as-yet unknown roles in plant health and phyllosphere microbial ecology.

3.1. Introduction

Trees and other wild plants can act as reservoirs of viruses that can cause disease in economically important crops [**1**, **2**, **3**], but most research on plant viruses has been focused on viruses that cause disease in crops and ornamental plants [**4**, **5**]. Very little is known about the prevalence and effects of viral infection in wild plants [**5**, **6**], although these viruses can play important ecological roles in the phytobiome, even in asymptomatic plants [**1**, **2**, **3**]. In particular, forest trees, such as oaks and conifers, have a broad ecological distribution and substantial economic importance, yet there is a relative lack of information about their associated viral diversity [**7**, **8**, **9**, **10**].

Forests are among the world's most important ecosystems. They cover 30% of the Earth's land surface, preserve most of Earth's terrestrial biodiversity, are an important carbon sink, and play a role in climate regulation [**11**, **12**, **13**]. Moreover, the forestry industry in the United States provides four percent of the total manufacturing Gross Domestic Product (GDP), which equals an estimated contribution of over $200 billion per year [**14**]. The increase in global trade has accelerated the spread of invasive pathogens to forests [**15**, **16**]. These pathogens are either introduced by accident, and/or adapt to new host trees [**13**] and are responsible for major economic and ecological damages [**17**, **18**]. Unravelling the natural tree virome may help predict forest responses to these perturbations and increasing pathogen emergence.

With the emergence of next generation sequencing (NGS), great advances have been made in the discovery of plant viruses, and these studies have shown diverse viruses in wild plants [**1**, **2**, **19**]. Currently little is known about the functions of these viruses, but it is clear that viruses can be symbiotic members of their plant host microbial community [**20**], and recent research has uncovered many novel RNA viruses [**21**, **22**]. Viruses associated with plants, including viruses that infect the plant and viruses that infect members of the plant microbiome, are known to play various roles with respect to plant health. Even though most plant viruses have been studied in the context of disease [**20**], some plant-associated viruses have been shown to have positive, mutualistic, or neutral interactions with the plant host [**5**, **23**, **24**, **25**], either directly [**26**, **27**] or indirectly [**28**, **29**, **30**, **31**, **32**]. For example, the mycovirus *Cryphonectria hypovirus 1* causes hypovirulence (reduced virulence) of the fungus, *Cryphonectria parasitica*, that causes chestnut blight. [**28**, **29**, **30**, **31**, **32**]. As part of the plant microbiome, bacteria and fungi can have diverse ecological interactions with the plant host, but the role of viruses that infect these plant-associated microbiota is relatively poorly understood.

To better expand the natural plant virome, we used a double-stranded RNA enrichment protocol to obtain RNA viral nucleic acids from plant leaves. This dsRNA enrichment protocol has fostered in-depth analyses of virus-specific sequences from plants in the past [**5**, **19**], and we chose it over other common techniques for viral community analyses with the expectation that it could facilitate better access to the phyllosphere viral community. Further, the majority of RNA viruses have a dsRNA life stage [**33**], including most known plant-infecting viruses and mycoviruses [**34**, **35**]. In previous studies, enrichment of virus-like particles followed by nucleic acid extraction and sequencing has had variable results in plants [**25**, **36**], and extraction of total DNA or RNA followed by bioinformatic mining of viral sequences has resulted in less viral 'signal' in the sequencing data, as the majority of the sequencing reads tend to be derived from cellular organisms [**4**]. Thus, this approach allowed us to directly target the phyllosphere viral community in line with previous viral phyllosphere studies.

In this study, we sequenced dsRNA derived from leaves of 16 oak and conifer species in order to reconstruct RNA viral contigs and assess the RNA viral diversity within and among these host tree species. The assembled contigs were examined for RNA-dependent RNA polymerase (RdRp) genes, a conserved gene found in RNA viruses that lack a DNA stage [**21**], along with other viral signatures, and viral communities were compared across tree species. This study is one of the first to look at natural tree virosphere in conifer and oak trees, and provided a robust picture of putative novel viruses.

3.2. Results and Discussion

3.2.1. Dataset overview and viral contig recovery.

To investigate the diversity and abundance of RNA viruses in conifer and oak trees, we extracted dsRNA from the leaves of 16 tree species (5 *Cupressaceae*, 5 *Pinaceae*, and 6 *Fagaceae*, commonly called cypresses, pines, and oaks, respectively, Supplementary table 1) from the UC Davis Arboretum. Samples were sequenced to an approximate depth of 6.6 Gbp each. Reads were assembled into contigs $\geq$ 200 bp, which were searched for viral signatures using 1) established Hidden Markov Models (HMMs) to identify viral RdRp genes [**21**], 2) VIBRANT [**37**] virus prediction software, and 3) BLASTp and BLASTn [**38**] searches against the NCBI nr (BLASTp) and nt (BLASTn) databases to manually investigate the putative viral contigs found via the first two approaches (Supplementary Figure 1). Of 186,591 total contigs $\geq$ 200 bp in the dataset, 1,166 were tentatively predicted as viral using approaches 1 and 2, and 202 of those were removed after manual investigation revealed an ambiguous or likely non-viral origin using approach 3, most often due to evidence that the contig was derived from a retrotransposon, not a virus. We note that, despite the dsRNA extraction that should have yielded predominantly viral RNA, many contigs were not predicted to be viral. Although we do not know the reason for certain, our virus detection approach was meant to be conservative to limit false positives, so we have likely missed some viruses, and a BLASTp search of all 192,336 predicted protein sequences revealed 69.5% of the sequences to be of likely plant origin (Supplementary Figure 3.1).

We detected 389 RdRp gene sequences on 384 contigs and a further 614 putative viral proteins on 580 contigs via VIBRANT, for a total of 964 putative viral contigs (average length 955 bases, Supplementary Table 2) that passed the manual curation step. Of those, eight had three or more predicted proteins and a genome length $>$ 1 kb, deemed sufficient for genome analysis, which revealed their divergence from known viruses on account of the presence of >30% hypothetical proteins and/or best BLAST hits to viruses of diverse taxa (*e.g.*, fungi and plants) within the same genome (Supplementary Figure 3.2). In order to maximize recovery of viral sequences that may not have assembled into contigs and to calculate the relative abundance of each putative virus in each sample, we mapped reads from each of the 16 samples to both the 964 viral contigs recovered de novo and 4,495 RefSeq viral genomes [**39**]. We detected 963 viral contigs through read mapping (889 from the Arboretum contigs and 74 from RefSeq). The 389 RdRp sequences from the Arboretum were translated to protein as described in the following section, and used for phylogenetic analyses, and the remaining analyses of viral populations and communities within and among trees considered the 963 viral contigs recovered through read mapping.

3.2.2. Viral RdRp diversity in leaves from 16 oak and conifer species.

To investigate viral diversity within the 16 tree species (Supplementary Table 1), we explored the phylogenies of the recovered RdRps in the context of known RdRps in the RefSeq database. We first translated the 389 predicted RdRp contigs into protein sequences and then dereplicated them at 99% amino acid identity (AAI) into 337 putative RdRp protein sequences, which were phylogenetically grouped via BLASTp searches against RefSeq RdRp sequences into 14 viral families. A phylogenetic tree of our discovered RdRps and 635 RefSeq RdRps from these 14 viral families was then constructed (Figure 3.1A, Supplementary table 3). Most of our RdRps grouped phylogenetically with unclassified viruses (n=92), followed by *Totiviridae* (n=38), *Bunyaviridae* (n=33), and *Secoviridae* (n=30) (Figure 3.1A). *Totiviridae* are known to infect fungi and protozoa [**40**], *Bunyaviridae* are known to infect plants, insects, and vertebrates [**41**], and *Secoviridae* are known to infect plants [**42**]. Our results suggest that both plant-infecting viruses and viruses that infect members of the phytobiome were recovered.

In a comparison of viral taxonomic diversity according to tree family (the *Pinaceae* and *Cupressaceae* families of conifers and the Fagaceae family of oaks), viral taxonomic composition was similar overall, but substantially more viruses were identified from oaks (n=247) than from the two conifer tree families (n=68 in *Cupressaceae* and n=22 in *Pinaceae*) (Figure 3.1B). For both the *Cupressaceae* and *Fagaceae* families, the 'taxonomic' category that represented the most RdRps was the 'unclassified viruses' group, at 29% and 28% of total RdRps, respectively. The largest taxonomic group associated with the *Pinaceae* was the *Bunyaviridae* family (18%). Besides unclassified viruses, most RdRps in all three tree families were associated with *Totiviridae*, *Bunyaviridae*, and *Secoviridae*, consistent with the dominance of these groups in the dataset overall.

To better understand the potential ecological implications of the recovered viruses we next sought to consider the potential hosts of the recovered viruses, based on phylogenetic affiliation of their RdRps with those of viruses with known hosts in RefSeq (Figure 3.1C). Of the 337 RdRps in the dataset, 32% were phylogenetically affiliated with viruses currently only known to infect plants (*Bromoviridae, Closteroviridae, Secoviridae, Solemoviridae or Tombusviridae* [**43**]), 20% were associated with viruses known to infect fungi (*Mitoviridae* or *Totiviridae* [**43**]), and 19% were aligned with viral families known to infect both plants and fungi (*Alphaflexiviridae*, *Betaflexiviridae*, *Partitiviridae* [**43**]). A further 18% grouped with unclassified viruses and thus could not be assigned to a putative host, while 10% were associated with viral families known to infect both plants and animals such as insects (*Bunyaviridae* [**41**]). Overall, the dataset was dominated by putative plant and fungal viruses.

However, 2% of the RdRps were associated with viruses from the *Caliciviridae* or *Flaviviridae* [**43**] that are thus far only known to infect *Animalia* (Figure 3.1C). Four of these RdRps aligned with the *Flaviviridae*, which are known to infect vertebrates and are transmitted by arthropods [**44**], and two aligned with *Caliciviridae*, which are known to only infect vertebrates [**41**]. Since both the *Caliciviridae* and *Flaviviridae* are primarily recovered in research central to human and animal pathogens [**44, 45**], these RdRps are well represented in the RefSeq database [**39**]. Thus, we suspected that the *Caliciviridae* and *Flaviviridae* RdRps in our dataset could have been erroneously assigned to these groups, partly on account of this database bias.

To further investigate whether these RdRps could represent true *Caliciviridae* or *Flaviviridae*, we performed a manual, web-based BLASTp search against the NCBI non-redundant protein database. Two of our RdRp sequences had significant alignments with RdRps of totiviruses, one with partitiviruses, one with mitoviruses, one with tombus-like viruses, and one with tymoviruses. All of these virus groups include known plant and/or fungal viruses [**43**, **46**], thus we believe all of the RdRps originally assigned to *Caliciviridae* or *Flaviviridae* were more likely of plant or fungal virus origin.

Overall, results are consistent with plant and/or fungal viruses dominating the RNA viral communities in these oak and conifer phytobiomes. Interestingly, RdRps from bacteriophages were not detected in our dataset, despite potential host bacteria presumably representing a large component of the phyllosphere and endosphere microbiome [**21**, **47**]. We infer that either bacteriophages with RNA genomes were not abundant in these samples or that they were not amenable to the laboratory and/or bioinformatic approaches used for viral recovery. Most known bacteriophages have dsDNA genomes [**48**], which would not have been recovered through our dsRNA library preparation. Therefore, the lack of detectable bacteriophage RdRps does not suggest that bacteriophages were absent from these plant phytobiomes.

3.2.3. Viral population composition detected within and across tree families.

We next sought to investigate the extent to which specific viruses were shared within and among the three tree families. We used the 389 RdRp-containing contigs, 580 putative viral contigs identified by VIBRANT [**37**], and 4,495 viral genomes from RefSeq as a reference database for read mapping from the 16 dsRNA metagenomes to assess the presence of each virus in each tree. Only viruses that were detected in two or more tree species were taken into account in this analysis, in order to compare within and between families. Viruses were most commonly shared among tree species within the same family, with the *Pinaceae* having the most shared viruses in the dataset (164), followed by the *Fagaceae* (78), and then the *Cupressaceae* (70) (Figure 3.2). After similarities within families, the *Cupressaceae* and *Pinaceae* together shared the most viral contigs (63). These results are unsurprising, since both of these tree families belong to the order *Pinales* (conifers) and are more closely related to each other than they are to the *Fagaceae.* More viral contigs were shared across all three tree families (*i.e.*, detected in at least three trees, with at least one species from each tree family) than were shared between the *Fagaceae* and either the *Cupressaceae* or *Pinaceae* alone. These results could indicate that there are both specific, host-associated viromes that align with tree phylogeny (for example, due to the presence of specific plant metabolites and/or host-associated microbiomes that could in turn select for specific viruses), as well as suggest a core virome common across the three tree families, perhaps due to their shared location within the UC Davis Arboretum.

3.2.4. Comparing viral community composition to host tree phylogeny.

Given that more viruses were shared within than among tree families, we wanted to see whether leaf viral communities would separate according to host tree phylogeny. Based on read mapping from each sample to all viral contigs and RefSeq viral genomes, we we generated a presence/absence matrix and computed pairwise correlations between viral communities using the Pearson method and compared the resulting hierarchical cluster of viral community composition with a phylogenetic tree of the tree species, derived from the chloroplast *rbcL* gene (commonly used to define tree phylogeny [**49**, **50**]). As in the phylogenetic tree of the trees, the hierarchical cluster of viral communities showed clear separation between the oaks and the conifers, along with separation according to the two families within the conifers (*Cupressaceae* and *Pinaceae*), yielding three primary clusters of viral community composition separated by tree family (Figure 3.3A). In fact, for all 10 conifer species, the dendrogram of viral community composition and the phylogenetic tree of trees aligned exactly, suggesting strong ties between host plant phylogeny and viral community composition, presumably due to host specificity for the viruses to the plants themselves and/or to their specific microbiomes. In contrast, within the oak (*Fagaceae*) family, the dendrogram of viral communities and the host phylogenetic tree exhibited slight differences. We partially attribute these differences to the phylogenetic dispersion captured in the oaks compared to the conifers: all of the oaks were from the same genus (*Quercus*), whereas the conifers spanned seven genera. Consistent with this, the three conifer species in the *Pinus* genus had viral communities that were more similar to each other than to the other two members of the *Pinaceae* family from different genera. We speculate that, at the shorter phylogenetic distances captured within a tree genus, the influence of host phylogeny on properties that would influence viral community composition was much smaller than at longer phylogenetic distances, where, in all cases, viral community composition aligned with tree phylogeny.

To further investigate the viral communities across the 16 tree species, we performed a principal coordinates analysis (PCoA, Figure 3.3B). This analysis now considers the relative abundances of all of the viruses recovered in the dataset through read mapping, whereas the analysis above was based solely on detection (presence/absence). Consistent with the presence/absence analysis, viral communities differed significantly by tree family (PERMANOVA $p = 0.001$) (Figure 3.3B). Viral community similarity in the PCoA plot followed approximately the same patterns as in the dendrogram derived from viral presence/absence data. However, the 'outlier' viral community from the *Q. douglasii* oak was particularly pronounced as separate from the other oaks in the PCoA plot, and this 'outlier' sample warranted further consideration, as described in the next section.

3.2.5. Evidence for differences in leaf viral community composition linked to senescence. The viral community of the oak tree *Q. douglasii* differed substantially from the viral communities of all other oak trees (Figure 3.3A, 3.3B), likely in part because the largest number of viral contigs (333) was recovered from that tree. This recovery was starkly different from other samples, where the second-largest number of viral contigs recovered from a tree was nearly 100 less than *Q. douglasii* (n=251 in *Pinus sylvestris*), and the average number for all other tree species was 130 (Supplementary Figure 3). *Q. douglasii* was the only species in this study known to be drought deciduous, meaning that leaves are shed in response to drought [**51, 52**]. Consistent with the drought deciduous lifestyle resulting in earlier senescence and loss of leaves, we found that the leaves of *Q. douglasii* were in a further state of senescence compared to other samples, as shown by their brown color (Supplementary Figure 4). As plants senescence and turn into dead organic matter, saprobic fungi have been shown to increase in activity as they decompose this fresh organic matter [**53**]. We hypothesized that viruses of these fungi might have become more active as a response to increased host activity and abundance. This was of particular interest, since viral communities of symbiotic and saprotrophic fungi are poorly understood [**54, 55**]. To investigate whether the senescing *Q. douglasii* sample contained more mycoviral sequences than other samples, we focused on the subset of RdRp HMM searches with matches to *Mitoviridae*, which was the most abundant viral family in the dataset exclusively known to infect fungi. A total of 83 matches to the *Mitoviridae* RdRp HMM was found in the dataset overall, and 87% of these (72 out of 83) were recovered through read mapping in the *Q. douglasii* sample, compared to a maximum of 15% in all other samples. Of all the *Q. douglasii* RdRp HMM matches that were recovered through read mapping (n=231), 31% matched the *Mitoviridae* HMM (72 out of 231), compared to a maximum of 16% (27 out of 170) in all other samples. Thus, both the diversity and proportion of *Mitoviridae* sequences was higher within *Q. douglasii* than in any of the other trees. We infer that *Q. douglasii* had more mycoviral sequences than any of the other trees, and we suspect that this may have been due to increased fungal and subsequent mycovirus activity within the leaves as the tree went into senescence.

3.3. Conclusions

By extracting dsRNA from tree leaves and mining the assembled contigs for viral signatures, we recovered 964 putative viral contigs from asymptomatic oak and conifer plants. The phylogenetic affiliation of many of these viruses with known plant and fungal viruses suggests that their primary hosts are plants or fungi, potentially indicating persistent infection of the host plant and/or infection of members of the plant microbiome. Although some viruses were detected in all three tree families examined, suggesting the potential for a core virome (*e.g.*, due to close proximity within the UC Davis Arboretum), most viruses that were detected in more than one tree were limited to tree species within the same tree family, suggesting host- and/or host microbiome-specific factors that could be structuring these viral communities. Interestingly, more putative mycoviruses were recovered from senescing oak leaves than from any other sample in the dataset, suggesting the potential for increased mycoviral activity coinciding with increased activity of saprobic fungi during senescence. Much viral diversity remains to be discovered, and here we have provided a framework to further investigate viral diversity in wild tree species for a more complete understanding of the plant holobiont and for identifying potential reservoirs of emergent plant diseases.

Bibliography

[1] Ian Cooper and Roger A C Jones. Wild plants and viruses: Under-Investigated ecosystems. In *Advances in Virus Research*, volume 67, pages 1–47. Academic Press, January 2006.

[2] Yuxin Ma, Armelle Marais, Marie Lefebvre, Sébastien Theil, Laurence Svanella-Dumas, Chantal Faure, and Thierry Candresse. Phytovirome analysis of wild plant populations: Comparison of Double-Stranded RNA and Virion-Associated nucleic acid metagenomic approaches. *J. Virol.*, 94(1), December 2019.

[3] James E Schoelz and Lucy R Stewart. The role of viruses in the phytobiome. *Annual Review of Virology*, 5:93:111, July 2018.

[4] Marilyn J Roossinck, Darren P Martin, and Philippe Roumagnac. Plant virus metagenomics: Advances in virus discovery. *Phytopathology*, 105(6):716–727, June 2015.

[5] Marilyn J Roossinck. Metagenomics of plant and fungal viruses reveals an abundance of persistent lifestyles. *Front. Microbiol.*, 5:767, 2014.

[6] Holly R Prendeville, Xiaohong Ye, T Jack Morris, and Diana Pilson. Virus infections in wild plant populations are both frequent and often unapparent. *Am. J. Bot.*, 99(6):1033–1042, June 2012.

[7] C Büttner, S von Bargen, M Bandte, and H P Mühlbach. Forest diseases caused by viruses. In *Infectious forest diseases*, pages 50–75. CABI, Wallingford, 2013.

[8] Preston R Aldrich and Jeannine Cavender-Bares. Quercus. In Chittaranjan Kole, editor, *Wild Crop Relatives: Genomic and Breeding Resources: Forest Trees*, pages 89–129. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[9] Aljos Farjon. The kew review: Conifers of the world. *Kew Bull.*, 73(1):8, March 2018.

[10] David M Richardson, Philip W Rundel, Stephen T Jackson, Robert O Teskey, James Aronson, Andrzej Bytnerowicz, Michael J Wingfield, and Şerban Procheş. Human impacts in pine forests: Past, present, and future. *Annu. Rev. Ecol. Evol. Syst.*, 38(1):275–297, December 2007.

[11] Xavier Morin, Lorenz Fahse, Hervé Jactel, Michael Scherer-Lorenzen, Raúl García-Valdés, and Harald Bugmann. Long-term response of forest productivity to climate change is mostly driven by change in tree species composition. *Sci. Rep.*, 8(1):5627, April 2018.

[12] S Trumbore, P Brando, and H Hartmann. Forest health and global change. *Science*, 349(6250):814–818, August 2015.

[13] M J Wingfield, E G Brockerhoff, B D Wingfield, and B Slippers. Planted forest health: The need for a global strategy. *Science*, 349(6250):832–836, August 2015.

[14] The state of the forest. `https://www.usda.gov/media/blog/2019/04/22/state-forest`. Accessed: 2021-9-10.

[15] S Gauthier, P Bernier, T Kuuluvainen, A Z Shvidenko, and D G Schepaschenko. Boreal forest health and global change. *Science*, 349(6250):819–822, August 2015.

[16] David W Langor, Erin K Cameron, Chris J K MacQuarrie, Alec McBeath, Alec McClay, Brian Peter, Margo Pybus, Tod Ramsfield, Krista Ryall, Taylor Scarr, Denys Yemshanov, Ian DeMerchant, Robert Foottit, and Greg R Pohl. Non-native species in canada's boreal zone: diversity, impacts, and risk. *Environ. Rev.*, 22(4):372–420, December 2014.

[17] Paolo Gonthier and Giovanni Nicolotti. *Infectious Forest Diseases*. CABI, 2013.

[18] Flávia Milene Barros Nery. Viral diversity in tree species. December 2020.

[19] Marilyn J Roossinck. Plant virus metagenomics: biodiversity and ecology. *Annu. Rev. Genet.*, 46:359–369, August 2012.

[20] Marilyn J Roossinck. The good viruses: viral mutualistic symbioses. *Nat. Rev. Microbiol.*, 9(2):99–108, February 2011.

[21] Evan P Starr, Erin E Nuccio, Jennifer Pett-Ridge, Jillian F Banfield, and Mary K Firestone. Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. *Proc. Natl. Acad. Sci. U. S. A.*, 116(51):25900–25908, December 2019.

[22] M Chiapello, J Rodríguez-Romero, M A Ayllón, and M Turina. Analysis of the virome associated to grapevine downy mildew lesions reveals new mycovirus lineages. *Virus Evolution*, 6(2), nov 2020. veaa058.

[23] Pierre Lefeuvre, Darren P Martin, Santiago F Elena, Dionne N Shepherd, Philippe Roumagnac, and Arvind Varsani. Evolution and ecology of plant viruses. *Nat. Rev. Microbiol.*, 17(10):632–644, October 2019.

[24] Marilyn J Roossinck. Move over, bacteria! viruses make their mark as mutualistic microbial symbionts. *J. Virol.*, 89(13):6532–6535, July 2015.

[25] Marilyn J Roossinck. Deep sequencing for discovery and evolutionary analysis of plant viruses. *Virus Res.*, 239:82–86, July 2017.

[26] Ping Xu, Fang Chen, Jonathan P Mannas, Tracy Feldman, Lloyd W Sumner, and Marilyn J Roossinck. Virus infection improves drought tolerance. *New Phytol.*, 180(4):911–921, September 2008.

[27] Jack H Westwood, Lucy McCann, Matthew Naish, Heather Dixon, Alex M Murphy, Matthew A Stancombe, Mark H Bennett, Glen Powell, Alex A R Webb, and John P Carr. A viral RNA silencing suppressor interferes with abscisic acid-mediated signalling and induces drought tolerance in arabidopsis thaliana. *Mol. Plant Pathol.*, 14(2):158–170, February 2013.

[28] Donald L Nuss. Hypovirulence and chestnut blight: From the field to the laboratory and back. In J W Kronstad, editor, *Fungal Pathology*, pages 149–170. Springer Netherlands, Dordrecht, 2000.

[29] Donald L Nuss. Hypovirulence: mycoviruses at the fungal-plant interface. *Nat. Rev. Microbiol.*, 3(8):632–642, August 2005.

[30] Xiaofang Wang, Zhong Wei, Keming Yang, Jianing Wang, Alexandre Jousset, Yangchun Xu, Qirong Shen, and Ville-Petri Friman. Phage combination therapies for bacterial wilt disease in tomato. *Nat. Biotechnol.*, 37(12):1513–1520, December 2019.

[31] B Balogh, Jeffrey B Jones, F B Iriarte, and M T Momol. Phage therapy for plant disease control. *Curr. Pharm. Biotechnol.*, 11(1):48–57, January 2010.

[32] Colin Buttimer, Olivia McAuliffe, R P Ross, Colin Hill, Jim O'Mahony, and Aidan Coffey. Bacteriophages and bacterial plant diseases. *Front. Microbiol.*, 8:34, January 2017.

[33] Marilyn J Roossinck. Plants, viruses and the environment: Ecology and mutualism. *Virology*, 479-480:271–277, May 2015.

[34] R C Gergerich and V V Dolja. Introduction to plant viruses, the invisible foe. *Plant Health Instr.*, 2006.

[35] Luis Rubio, Luis Galipienso, and Inmaculada Ferriol. Detection of plant viruses and disease management: Relevance of genetic diversity and evolution. *Front. Plant Sci.*, 11:1092, July 2020.

[36] Ulrich Melcher, Vijay Muthukumar, Graham B Wiley, Byoung Eun Min, Michael W Palmer, Jeanmarie Verchot-Lubicz, Akhtar Ali, Richard S Nelson, Bruce A Roe, Vaskar Thapa, and Margaret L Pierce. Evidence for novel viruses by analysis of nucleic acids in virus-like particle fractions from ambrosia psilostachya, 2008.

[37] Kristopher Kieft, Zhichao Zhou, and Karthik Anantharaman. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, 8(1):90, June 2020.

[38] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, October 1990.

[39] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, 35(Database issue):D61–5, January 2007.

[40] Kenta Okamoto, Naoyuki Miyazaki, Daniel S D Larsson, Daisuke Kobayashi, Martin Svenda, Kerstin Mühlig, Filipe R N C Maia, Laura H Gunn, Haruhiko Isawa, Mutsuo Kobayashi, Kyoko Sawabe, Kazuyoshi Murata, and Janos Hajdu. The infectious particle of insect-borne totivirus-like omono river virus has raised ridges and lacks fibre complexes. *Sci. Rep.*, 6:33170, September 2016.

[41] Susan Payne. *Viruses: From Understanding to Investigation*. Academic Press, August 2017.

[42] Jeremy R Thompson, Nitin Kamath, and Keith L Perry. An evolutionary analysis of the secoviridae family of viruses. *PLoS One*, 9(9):e106305, September 2014.

[43] Gustavo Fermin. Host range, host–virus interactions, and virus transmission. *Viruses*, page 101, 2018.

[44] Flaviviridae. In *Fenner's Veterinary Virology*, pages 525–545. Elsevier, 2017.

[45] Ulrich Desselberger. Caliciviridae other than noroviruses. *Viruses*, 11(3), March 2019.

[46] G P Martelli, A A Agranovsky, M Bar-Joseph, D Boscia, T Candresse, R H A Coutts, V V Dolja, B W Falk, D Gonsalves, W Jelkmann, A V Karasev, A Minafra, S Namba, H J Vetten, G C Wisler, and N Yoshikawa. The family Closteroviridae revised. *Archives of Virology*, 147(10):2039–2044, 2002.

[47] Vasvi Chaudhry, Paul Runge, Priyamedha Sengupta, Gunther Doehlemann, Jane E Parker, and Eric Kemen. Shaping the leaf microbiota: plant–microbe–microbe interactions. *J. Exp. Bot.*, 72(1):36–56, September 2020.

[48] Moïra B Dion, Frank Oechslin, and Sylvain Moineau. Phage diversity, genomics and phylogeny. *Nat. Rev. Microbiol.*, 18(3):125–138, March 2020.

[49] J R Manhart. Phylogenetic analysis of green plant rbcl sequences. *Mol. Phylogenet. Evol.*, 3(2):114–127, June 1994.

[50] Yong Kang, Zhiyan Deng, Runguo Zang, and Wenxing Long. DNA barcoding analysis and phylogenetic relationships of tree species in tropical cloud forests. *Sci. Rep.*, 7(1):12564, October 2017.

[51] D. D. McCreary. Native california oaks losing leaves early. university of california. Available at https://ucanr.edu/blogs/blogcore/postdetail.cfm?postnum=8276, 2021.

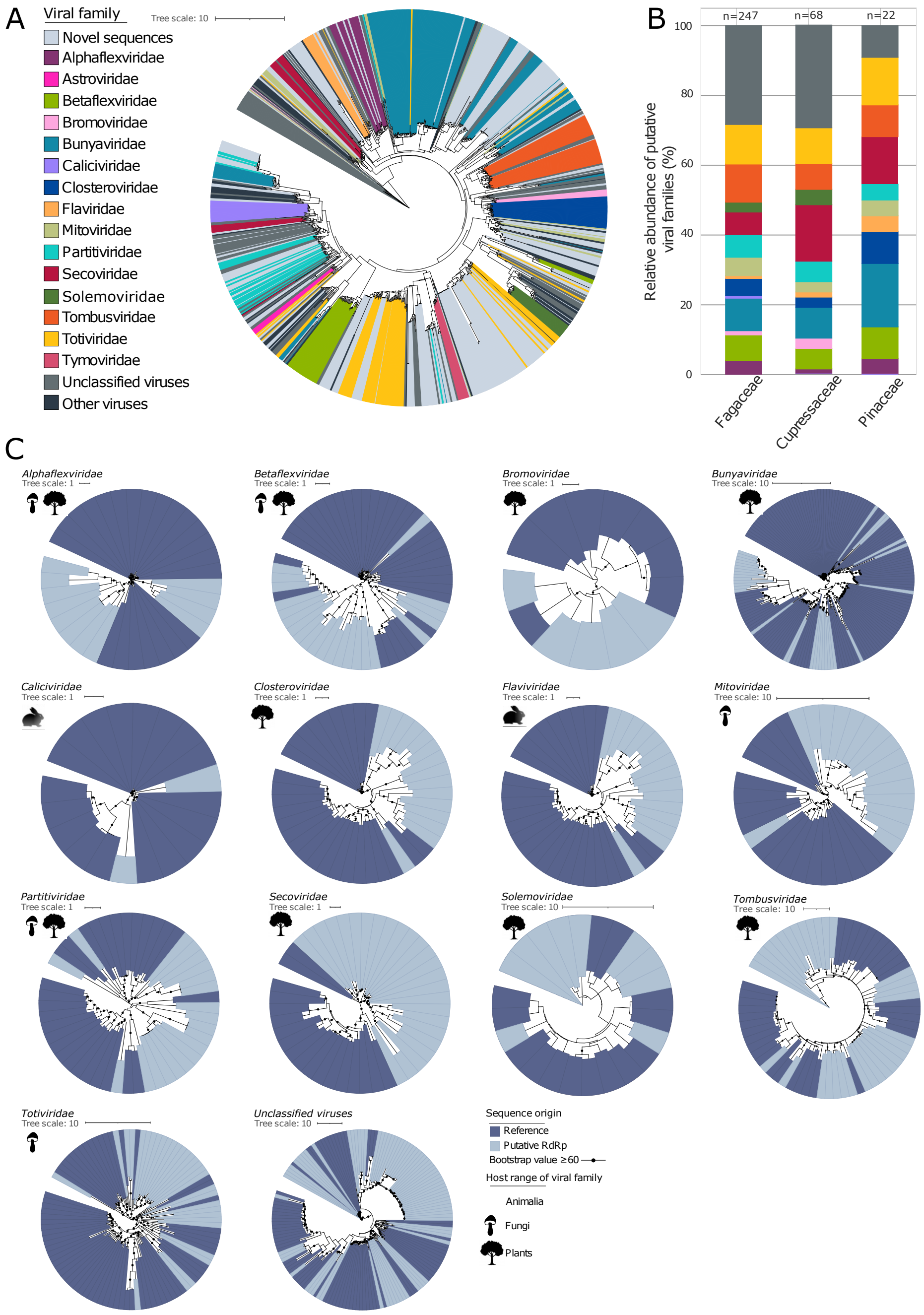
[52] Marc D. Abrams. Adaptations and responses to drought in Quercus species of North America. *Tree Physiology*, 7(1-2-3-4):227–238, 12 1990.

[53] Björn D Lindahl and Anders Tunlid. Ectomycorrhizal fungi - potential organic matter decomposers, yet not saprotrophs. *New Phytol.*, 205(4):1443–1447, March 2015.

[54] Suvi Sutela, Anna Poimala, and Eeva J Vainio. Viruses of fungi and oomycetes in the soil environment. *FEMS Microbiology Ecology*, 95(9), 07 2019. fiz119.

[55] Suvi Sutela, Marco Forgia, Eeva J Vainio, Marco Chiapello, Stefania Daghino, Marta Vallino, Elena Martino, Mariangela Girlanda, Silvia Perotto, and Massimo Turina. The virome from a collection of endomycorrhizal fungi reveals new viral taxa with unprecedented genome organization. *Virus Evolution*, 6(2), 10 2020. veaa076.

Figures and tables



(Continued on the following page)

Figure 3.1: Phylogenetic classification of viral contigs, based on alignment of RdRp gene sequences from RefSeq and this dataset. A) Unrooted phylogenetic tree (concatenated predicted protein alignment) of RdRp sequences from all newly identified contigs ('Novel sequences', this dataset, n=337) and RefSeq (n=635). The tree is colored by viral family phylogeny from RefSeq. 'Unclassified viruses' and 'Other viruses' are also from RefSeq but did not have a taxonomic assignment or were assigned to other viral groups, respectively. **B)** Average relative abundances of viral taxa within tree families, based on putative taxonomic assignments for RdRp contig sequences (derived from significant BLAST hits for the RdRp gene to known RdRp sequences in RefSeq). Colors correspond to the legend in panel A. 'Unclassified viruses' indicate RdRps from our dataset that had significant best BLAST hits to unclassified viruses in RefSeq. Relative abundances of each RdRp-containing contig in each sample were derived from read mapping to the 384 RdRp-containing contigs, and relative abundances were summed for each taxon and averaged across tree species within each tree family to generate the stacked bar charts. Numbers at the tops of bars indicate the total number of RdRp-containing contigs detected for each tree family. **C)** Unrooted phylogenetic trees of RdRp protein sequences. The most significant RdRp BLAST bit score was used to assign each contig to a viral family, and phylogenetic trees were constructed separately for each family. Bootstrap support values $\geq$ 60% are shown as circles on nodes, and were calculated using an approximate likelihood ratio test (aLRT) with the Shimodaira–Hasegawa-like procedure (SH-aLRT), using 1000 bootstrap replicates.

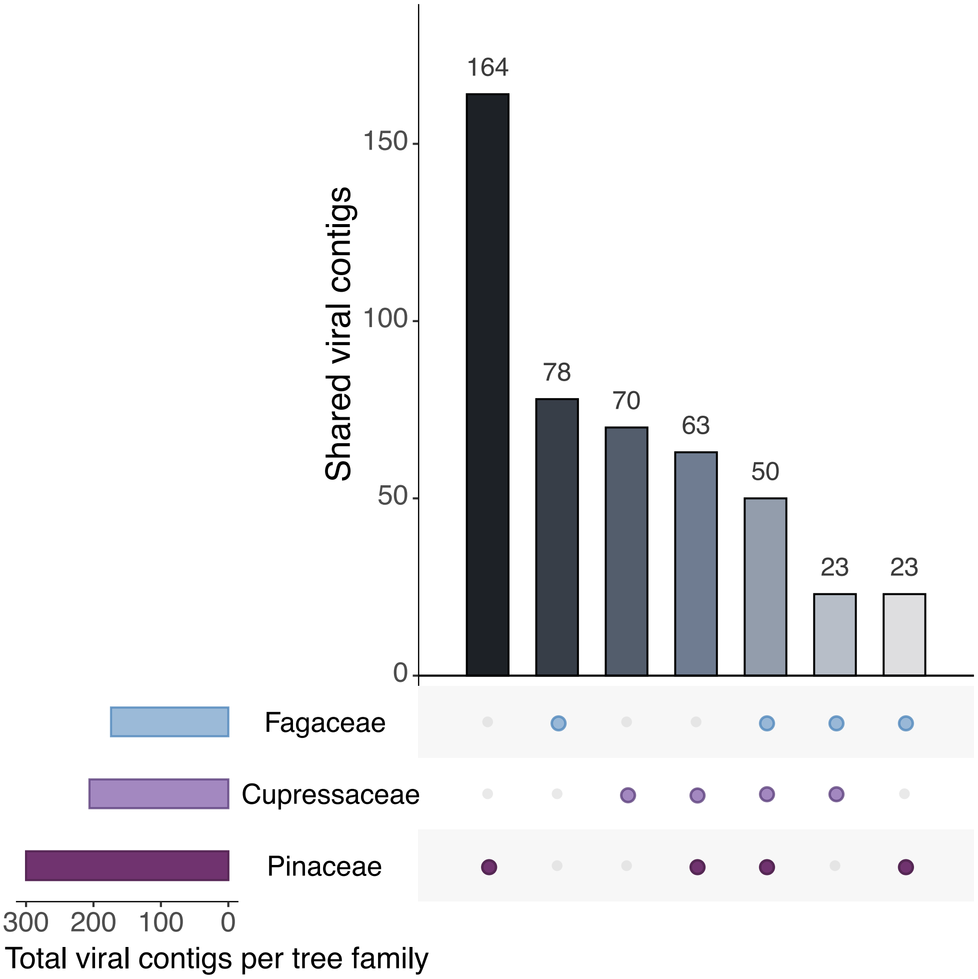


Figure 3.2: Upset plot of shared viral contig sequences between tree families based on the presence/absence version of the table of viral contig sequences in each sample (Supplementary Table 5). Colored dots below the bars indicate the tree family or families included in each bar. Only viral contigs detected in two or more tree species were included in this analysis (n=471).

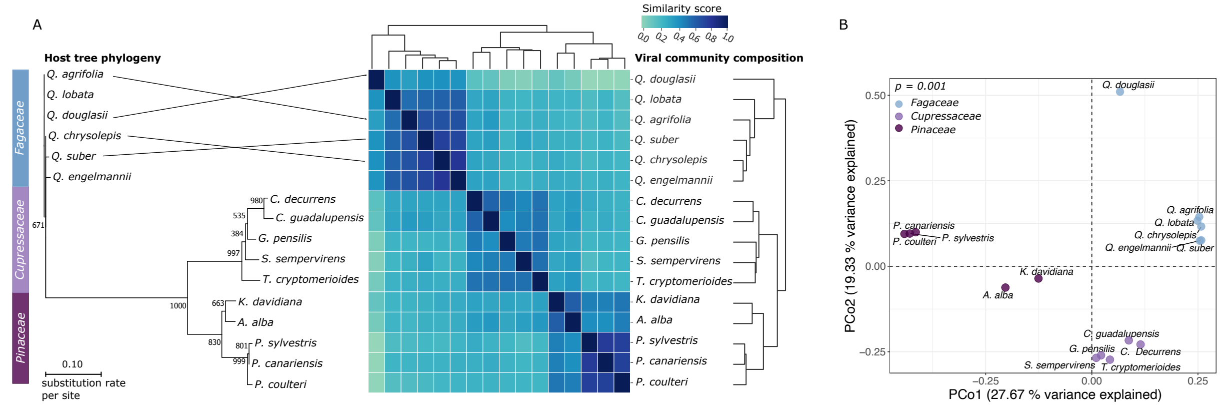


Figure 3.3: Viral community composition by tree taxonomy. A) Unrooted phylogenetic tree of tree species, based on sequence alignment of the *rbcL* gene (left), connected via a tanglegram (lines after tree species names; no line indicates the same row, equivalent to a horizontal straight line) to a heatmap and associated dendrograms (right and top, same dendrogram repeated) of pairwise viral community similarity between tree leaf samples. Viral community similarity was measured as Pearson similarity between pairs of samples, starting from a presence-absence matrix of 964 viral contigs in each sample, with detection patterns based on read mapping to the viral contigs. Crossed lines in the tanglegram indicate discrepancies between tree phylogeny and the viral community composition dendrogram. B) Principal Coordinates Analysis (PCoA) of viral community composition, based on pairwise Bray-Curtis dissimilarities derived from per-sample read mapping to 964 viral contigs. Each point represents a sample from one of the 16 tree species, labeled by species and colored by tree family. The p-value is from a PERMANOVA test for differences in viral community composition according to the three tree families.

CHAPTER 4

Dispersal, habitat filtering, and eco-evolutionary dynamics as drivers of local and global wetland viral biogeography

Anneliek M. ter Horst[1], Jane D. Fudyma[1], Jacqueline L. Sones[2], and Joanne B. Emerson* [1]

[1] Department of Plant Pathology, University of California, Davis, Davis, CA, USA

[2] Bodega Marine Reserve, University of California, Davis, Bodega Bay, CA, USA

Abstract

Wetlands store 20-30% of the world's soil carbon, and identifying the microbial controls on these carbon reserves is essential to predicting feedbacks to climate change. Although viral infections likely play important roles in wetland ecosystem dynamics, we lack a basic understanding of wetland viral ecology. Here 63 viral size-fraction metagenomes (viromes) and paired total metagenomes were generated from three time points in 2021 at seven fresh- and saltwater wetlands in the California Bodega Bay Marine Reserve. We recovered 12,826 viral population genomic sequences (vOTUs), 4.4% of which were also detected at the same field site two years prior, indicating a small degree of population stability or recurrence. Viral communities differed most significantly across the seven wetland sites and were also structured by habitat (plant community composition and salinity). Read mapping to a new version of our reference database, PIGEONv2.0 (now with 515,763 vOTUs), revealed 196 vOTUs present over large geographic distances, often reflecting shared habitat characteristics. Wetland vOTU microdiversity was significantly lower locally than globally and lower within than between time points, indicating greater divergence with increasing spatiotemporal distance. Viruses tended to have broad predicted host ranges via CRISPR spacer linkages to metagenome-assembled genomes (whether this reflects true biology remains to be seen), and increased SNP frequencies in CRISPR-targeted major tail protein genes suggest viral eco-evolutionary dynamics,

potentially in response to both immune targeting and to changes in host cell receptors involved in viral attachment. Together, these results highlight the importance of dispersal, environmental selection, and eco-evolutionary dynamics as drivers of local and global wetland viral biogeography.

4.1. Introduction

Wetlands are an important carbon sink, estimated to store between 20-30% of the global soil carbon [**1**]. They also provide ecosystem services, such as flood control, drought prevention, and water quality protection, and they support a rich biodiversity [**1**, **2**, **3**, **4**]. However, these ecosystems are currently being lost at an estimated annual rate of 1.5% globally, releasing stored carbon into the atmosphere [**2**, **5**]. Moreover, due to climate change, soil salinity is increasing in formerly freshwater wetlands, causing changes to microbial and plant communities [**6**, **7**, **8**] and potentially leading to biodiversity loss [**9**]. Microorganisms play central roles in carbon turnover and the emission of greenhouse gasses from wetland ecosystems [**10**], and, by infecting, controlling the metabolism of, and lysing microorganisms, viruses also likely impact these biogeochemical cycles [**10**, **11**]. It is therefore important to characterize fresh- and saltwater wetland microbial and viral communities, in order to understand the ecological and biogeochemical responses of these fragile ecosystems under a changing climate [**12**, **13**, **14**].

While viruses are highly abundant in peat wetlands and other soils [**15**, **16**, **17**, **18**], we still know relatively little about wetland viral ecology, as methodological improvements have only recently made it possible to study soil viral communities in detail. While some prior efforts have focused on bioinformatic mining of viral sequences from total soil metagenomes [**19**, **20**], by purifying the viral size fraction through 0.22 μm filtration prior to metagenomic sequencing (viromics), a much higher viral diversity can be recovered [**11**, **15**, **21**]. Application of these methods to peatlands and a variety of other soils is beginning to reveal ecological factors important to soil viral biogeography.

Recent studies have shown substantial differences in soil viral community composition among habitats at both regional and global scales [**15**, **22**]. For example, soil viral 'species' (vOTUs) were rarely shared among four different habitats (grasslands, shrublands, woodlands, and wetlands) in northern California [**22**], and similarly, few RNA viral sequences were shared between grasslands and peatlands in the United Kingdom [**23**]. Despite repeated observations of soil viral community heterogeneity at regional or continental scales [**18**, **24**], the same viral 'species' (vOTUs) can be found on different continents, usually in the same habitat (e.g. peat viruses tend to be restricted to other peatlands) [**15**]. While habitat seems to be an important contributor to soil viral biogeography, given the sparseness of the data, further studies are needed to assess the generalizability of these patterns.

At more local scales, soil viral community spatial structuring and temporal turnover have been observed, with viral communities showing seasonal dynamics [**25**] and exhibiting stronger spatial and temporal distance-decay relationships than bacterial communities [**18**, **21**]. However, those studies were conducted within the same habitat or soil type; differences in viral community composition across habitats have rarely been considered at local scales. In two studies that did compare viral communities by habitat in the same Swedish ecosystem, viral communities were found to be distinct among three habitats along a peatland permafrost thaw gradient [**16**, **19**]. However, those three habitats were also spatially separated, making the relative influences of habitat and spatial location difficult to disentangle. Similarly, viral community compositional differences along a grassland pH gradient also reflected spatial separation, but pH was seemingly the predominant factor driving viral community composition, which was corroborated in a meta-analysis of other soil and peat viral datasets [**26**]. Disentangling the relative impacts of habitat characteristics and spatial location on soil viral community composition is thus an important near-term goal for advancing the field, but appropriate spatiotemporal scales for sampling soil viral communities are still unknown.

Building on our prior regional study of 30 viromes from four habitat types with very little overlap in viral ‘species’ (vOTUs) across samples [**22**], here we hypothesized that reducing complexity from the regional to local scale and restricting the diversity of habitats considered (only wetlands) would yield sufficient vOTU co-occurrence to link viral ecological patterns to their potential underlying drivers. We sampled seven different wetland sites across a 0.6 km^2 area at three time points in 2021 at the Bodega Marine Reserve on the California Pacific Coast (USA). We generated 63 viral size-fraction metagenomes (viromes) and 63 total soil metagenomes to profile the dsDNA viral communities and bacterial and archaeal (prokaryotic) communities, respectively, in these wetlands. We also compared results to our viromic dataset from Bodega Bay collected two years prior (in 2019) [**22**]. Here we explore local and global wetland viral biogeography, investigate which factors among spatial distance, plant and microbial community composition, soil physicochemical properties, and time have the strongest influence on viral community composition, and evaluate the influences of spatial and temporal distance on viral population microheterogeneity and virus-host eco-evolutionary dynamics.

4.2. Results and discussion

4.2.1. Dataset overview.

To investigate wetland dsDNA viral biogeography on a local scale, we sampled seven nearby wetland sites within a 0.6 km^2 area in the Bodega Marine Reserve, California, USA (Figure 4.1A, map). Sampling sites were initially selected to represent freshwater, brackish, and saltwater wetlands, based on institutional knowledge of plant community composition, and we subsequently measured both plant communities and salinity to empirically define the sampled habitats. Near-surface (top 15 cm) wetland soils were collected at three time points (March, May, and July of 2021), with three replicate samples per time point per wetland site. Replicates were collected on average 17 m apart, with the closest samples within a site (regardless of the time point) 1.7 m apart and the farthest 89 m apart (Supplementary Table 1). All 63 samples (7 sites x 3 replicates x 3 time points) underwent viral size-fraction metagenomics (viromics) and total metagenomics to measure viral and prokaryotic community composition, respectively. A suite of soil physicochemical properties was also measured for each sample (Supplementary Table 2). In total, 12,826 viral operational taxonomic units (vOTUs, $\geq$ 10 kbp, $\geq$ 95% average nucleotide identity, approximately species-level taxonomy [**27**]) and 219 metagenome-assembled genomes (MAGs,$\geq$ 50% complete, $\leq$ 10% contaminated [**28**]), Supplementary Table 3) were detected in our samples. From the viromes, we assembled 17,703 viral contigs de novo, which clustered into 12,261 vOTUs, and we recovered an additional 565 vOTUs by read mapping to our Phages and Integrated Genomes Encapsidated Or Not (PIGEONv2.0) database of 515,763 vOTUs from diverse ecosystems, including 369 vOTUs recovered from Bodega Bay viromes collected in 2019 [**22**]. Read mapping to these sets of vOTUs and MAGs yielded the estimated relative abundances of each vOTU and MAG in each sample, used for downstream community compositional analyses (Supplementary Tables 4,5).

4.2.2. Habitat features (plant community composition and salinity).

We identified 32 plant species across the seven sites (Supplementary Table 6), and plant communities separated the sites into two vegetation groups, based on the presence or absence of halophytes (salt-tolerant plants). There were no overlapping plant species between the two groups, and while most sites in the same vegetation group shared at least one dominant plant species, plant community composition differed at each of the seven sites (Supplementary Table 6). We also used salinity measurements to define habitats, with electrical conductivity measurements ranging from 0 to 82 mmhos/cm in our wetlands, and those between 0 to 2 mmhos/cm considered non-saline, 2 to 4 slightly saline, 4 to 8 moderately saline, 8 to 16 strongly saline, and 16 or greater extremely saline wetlands [**29**]. Although vegetation tended to be indicative of soil salinity, our salinity measurements varied both within and among sites, and two sites had consistently mismatched salinity and vegetation measurements (site M1 had a 'no halophyte' plant community with non-to-strongly saline soils, and site M2 had a 'halophyte' plant community with non-saline soils). These seemingly contradictory vegetation and salinity results left us initially concerned that our salinity measurements might have been faulty, but evaporation during dry periods, seasonal waterlogging, precipitation, and leaching of water can all influence soil salinity in short time spans [**30**]. Halophytic plants outcompete non-halophyte plants in saline environments [**31**], and coastal salt marshes such as site M2 experience tidal flooding with seawater, leading us to believe that M2 likely sometimes experiences higher soil salinity than we measured, promoting halophyte growth. Halophytes are not competitive in non-saline habitats [**32**], and since there were no halophytes at site M1 despite the moderate salinity measurements, we speculate that M1 soil is often non-saline. Regardless of the underlying mechanism(s) for the differences, we separated the seven sites into four habitat groups: "Halophyte (H)" for the two sites with halophyte plants and overall medium to extreme soil salinity (H1 and H2), "No Halophyte (NH)" for the three sites with no halophytes and low to slight soil salinity (NH1, NH2, and NH3), and two "Mismatched (M)" groups (M1 and M2) for the two sites for which the vegetation did not correspond with soil salinity. Importantly, the mismatched (M) sites did not share the same vegetation and salinity mismatch, so they do not represent the same habitat type, leading to four habitat groups (H, NH, M1, M2).

4.2.3. Viral and prokaryotic communities were distinct at each of the seven wetland sites but were more similar within than between habitat types.

Most (90%) of the viral 'species' (vOTUs) were restricted to only one of the seven wetland sites. While 38% of the vOTUs were detected in only one of the 63 viromes (Supplementary Figure 1), the proportion of these 'singleton' vOTUs was substantially reduced, compared to our prior regional-scale comparison of 30 viromes across grassland, shrubland, woodland, and wetland habitats, in which 81% of the vOTUs were detected in only one virome [**22**]. Thus, the localized focus in one area and restriction to wetland habitats here, as well as increased spatiotemporal resolution, improved our ability to identify vOTUs shared across samples, as is necessary for recognizing biogeographical patterns. Of the 62% of vOTUs detected in more than one virome, 6,680 (52%) were recovered only within one wetland site, and viral community composition was significantly different at each site (PERMANOVA $p < 0.001$, Figure 4.1B, 4.1C). Viral community beta-diversity was significantly negatively correlated with spatial distance (Supplementary

Figure 4.2A), implicating dispersal limitation as one potential driver of these patterns (as also suggested in previous work [**18**, **21**, **22**]).

Despite overarching differences among wetland sites, viral communities grouped secondarily according to habitat type (Figure 4.1B, 4.1C), with significant differences among the four habitat groups (H, NH, M1, and M2, PERMANOVA, $p < 0.001$). Consistent with edaphic factors as potential drivers of these differences, viral community composition correlated significantly with soil chemical measurements (Supplementary Figure 4.2B, Supplementary Table 2), such as pH, moisture content, and sulfate concentrations (Supplementary Figure 4.2C). Considering only between-habitat beta-diversity, the viral communities from sites NH1 and NH2 were the most similar (Figure 4.1C), despite being physically far apart (Figure 4.1A), perhaps related to their similar salinity and plant communities (Supplementary Table 6). This is consistent with prior work that has suggested that plant cover type could play an important role in shaping soil viral communities [**33**]. Similar salinity and plant communities were likely also drivers of viral community compositional similarity at sites H1 and H2. Those sites are also connected by a culvert (a human-made water tunnel beneath a road) (Figure 4.1A), presumably facilitating dispersal between the sites. Finally, the two mismatched sites each had distinct viral communities, potentially due to their unique combinations of plant composition and salinity.

A co-occurrence analysis revealed that vOTUs were most often shared across samples from the same habitat type. Specifically, vOTUs from all three non-halophyte soils (NH1, NH2, and NH3) tended to co-occur, as did vOTUs from wetlands with halophyte plants (H1 and H2). Perhaps reflecting the lack of other samples from the same habitat types in this dataset, vOTUs from each of the mismatched sites tended to co-occur only with other samples from the same site. Interestingly, a small subnetwork of vOTUs from the halophyte site H2 co-occurred with vOTUs from the non-halophyte wetlands. All of those co-occurring vOTUs were either the most abundant in or only detected in one particular H2 sample, H2-1-T2, which had low salinity (1.69 mmho/cm) (Supplementary Table 2). This suggests that environmental selection (presumably by way of microbial hosts) can act on wetland viral communities on very short time scales, and/or that an influx of new vOTUs was brought to site H2, being already adapted to conditions in the less saline water that brought them there.

The two sites with the most within-site vOTU co-occurrences also had the highest moisture content, consistent with hydrological mixing facilitating greater viral community homogeneity. Specifically, site NH3 harbored viral communities distinct from all other sites (Figure 4.1B), despite its similar salinity and plant community composition to the other two non-halophyte sites and its close proximity to NH2 (Figure 4.1A, Supplementary Table 2). A comparatively large percentage of vOTUs was shared across samples within the NH3 site (35% of vOTUs were shared among five or more NH3 samples, relative to only 8% on average for the other two non-halophyte sites, Supplementary Figure 1). Similarly, communities from site M1 were also distinct, with 30% of their vOTUs detected in five or more samples from the same site, whereas the five other sites (not NH3 or M1) shared only 9% of their vOTUs across five or more samples from the same site. Soil moisture content was highest at sites NH3 (83% on average) and M1 (52% on average), compared to 34% on average at the other five sites, likely facilitating more mixing and greater viral community homogeneity due to greater hydrologic connectivity. Overall, the viral community compositional and vOTU co-occurrence patterns revealed both dispersal (and dispersal limitation) and environmental selection (biotic and abiotic habitat characteristics) as likely drivers of local wetland viral biogeographic patterns.

To determine whether prokaryotic communities exhibited similar patterns to the viral communities, prokaryotic community composition and co-occurrence were also investigated. Briefly, the relative abundances and co-occurrences of MAGs and, separately, of 16S rRNA gene fragments recovered from total metagenomes were used for these analyses. While most of the prokaryotic communities were significantly different at each of the wetland sites (Figure 4.1E, Supplementary Figure 3A), the communities from the Halophyte sites (H1 and H2) were not significantly different from each other (PERMANOVA, p=0.055), grouping more by habitat type than did the viral communities. Co-occurrence networks for MAGs showed similar patterns to those of the viral communities, largely reflecting shared MAGs within the same habitat type, though relatively few MAGs were recovered from the non-halophyte wetlands (Supplementary Figure 3B,C). Although OTUs also showed the most co-occurrence within habitat types, OTUs were detected in multiple habitats far more often than were MAGs, suggesting that increased resolution (i.e., not requiring assembly into MAGs) revealed more co-occurrence, presumably due to increased access to rare community members. Overall, patterns for prokaryotic communities were similar to those of their viruses, and viral community composition was significantly correlated with prokaryotic community composition (Mantel test, $p < 0.001$), suggesting that at least some of the observed viral biogeographical patterns were due to habitat filtering (environmental selection) by way of their hosts.

4.2.4. Global distribution patterns for Bodega Bay vOTUs suggest that wetland viral biogeography reflects habitat and salinity.

To compare vOTUs recovered at Bodega Bay to the global viral metacommunity, we leveraged a new version of our Phages and Integrated Genomes Encapsidated Or Not (PIGEONv2.0) database, which we introduce here. Since the first iteration of PIGEON (PIGEONv1.0), which contained 266,125 vOTUs [**15**], PIGEONv2.0 has almost doubled in size, now including 515,763 vOTU sequences. Most notably, we increased the number of soil vOTUs from 15,892 to 61,757, predominantly from our in-house soil viromics data, including previously unpublished datasets that we are now making publicly available in PIGEONv2.0. The number of freshwater vOTUs also substantially increased, largely due to the addition of viruses from aquatic viromes from Lake Baikal in Russia [**34**]. Here, these PIGEON improvements have facilitated global comparisons of Bodega Bay vOTU occurrence patterns.

Of the 12,826 vOTUs recovered at Bodega Bay, 196 (1.5%) were previously detected at other sites throughout the world (Figure 4.2A), recovered here through read mapping to PIGEONv2.0 (Figure 4.2B, Supplementary Table 7). Bodega Bay vOTUs were previously recovered from non-wetland soils (83), freshwater lakes (57), marine ecosystems (33), non-peat freshwater wetlands (14), and peat wetlands (8), indicating globally present viruses in relatively similar ecosystems throughout the world (Figure 4.2A, Supplementary Figure 4A). Notably, zero vOTUs from human-associated habitats were detected in these wetlands, perhaps indicating species boundaries between these very different habitat types. Most vOTUs that were detected in non-saline or slightly saline wetlands at Bodega Bay were originally recovered from non-wetland soils (62) or freshwater ecosystems (46), whereas most vOTUs from saline wetlands were previously recovered from marine (34) or non-wetland soil (28) ecosystems (Supplementary Figure 4.4B), again suggesting that habitat characteristics underlie global viral biogeographic patterns. Similarly, we also considered the relationship between vegetation group at Bodega Bay and the habitat in which a given vOTU was originally recovered (Figure 4.2D) and found that vOTUs from the non-halophyte sites were most often previously detected in non-wetland soils (79) or freshwater ecosystems (57), while vOTUs from the halophyte sites were most often previously detected in marine ecosystems (24) (Figure 4.2C). The detection of marine vOTUs in these wetlands is counter to our previous study of freshwater peatlands in Minnesota, USA, in which zero marine vOTUs from PIGEONv1.0 were detected [**15**], consistent with salinity as a habitat filter for vOTUs in both oceans and wetlands. Together, these results indicate that habitat characteristics – in this case, salinity and salinity indicators (halophyte or non-halophyte plant community composition) – can drive wetland viral community biogeography on a global scale.

4.2.5. Wetland viral microdiversity was lower locally than globally and lower within than between time points.

To investigate the contributions of viral genotypic heterogeneity to local and global viral ecology, we used inStrain [**35**]to calculate vOTU microdiversity profiles and compared dominant allelic variants over time and space. Specifically, we compared vOTU reference sequences initially recovered from PIGEONv2.0 (not assembled from Bodega Bay, 196), assembled from Bodega Bay in 2019 (2,377) [**22**], and assembled from Bodega Bay in 2021 (this study, 12,261) to their variants recovered in different samples at Bodega Bay. For each vOTU, we calculated pairwise average nucleotide identities (ANIs) between each sample-specific consensus variant sequence from Bodega Bay and the reference vOTU sequence. Genomic similarity between Bodega Bay variants and PIGEON references was significantly lower on average (average ANI 97.48%) than that for variants that were both assembled and recovered from Bodega Bay (average ANI 99.55%, Figure 4.3A, $p < 0.001$, Student's T-test). Given the global scale of PIGEON and local scale of Bodega Bay, this indicates greater viral population allelic variance (genomic heterogeneity) with increasing distance and/or time between samplings, a pattern known as 'isolation by distance' that has been studied for geographic distance, whereby populations in closer proximity are more genetically similar than populations that are farther away [**36**].

A relatively small number of the Bodega Bay vOTUs detected in 2021 were also recovered from Bodega Bay in 2019 (568 vOTUs, 4.4% of the 2021 dataset). This suggests that a small part of the wetland soil virosphere was stable or consistently recurrent over time. However, for vOTUs that were assembled and recovered through read mapping in the same year, the genomic similarity of dominant allelic variants was higher (99.86%) than for vOTUs that were assembled and recovered in different years (99.25%, Student's T-test, $p < 0.001$, Figure 4.3B). Thus, although these viral 'species' persisted over time, their strain-level heterogeneity increased over the two years, consistent with temporal 'isolation by distance' [**36**], with populations farther apart in time exhibiting more genomic divergence.

Sub-population dynamics for vOTUs that were recovered multiple times within the same Bodega Bay wetland site in 2021 were also compared to assess short-term eco-evolutionary dynamics. Genomic similarity of dominant allelic variants was highest for vOTUs recovered through read mapping at the time point from which they were assembled (Figure 4.3C) and was significantly lower at both of the other time points. This indicates that, even over short time scales of one to two months, variants significantly fluctuated in abundance and/or diverged. Given that there was no linear trajectory in variant ANI divergence with time (variants were just as different between adjacent time points as between the first and third time points), abundance fluctuations seem more likely to explain these patterns than divergence.

We also used inStrain to compare MAG allelic variants in the 2021 Bodega Bay dataset. MAG variants recovered and assembled at the same time point were most genomically similar (had the highest ANI), whereas MAGs from different time points had significantly lower ANI (Supplementary Figure 4.3D). Interestingly, in contrast to the vOTU variants, MAG sub-population dynamics exhibited temporal progression, with sub-population pairs from the same time point most similar, those from the first and last time points most distinct, and those from adjacent time points (i.e., from time points 1 and 2 or 2 and 3) exhibiting intermediate similarity in their ANIs. Additional time points would be required to determine whether this is likely due to divergence over time, but results show sub-population dynamics for both viral and prokaryotic populations over months.

4.2.6. Viral 'species' (vOTUs) tended to have broad predicted host ranges, and on average, MAGs had evidence for interactions with more than 10 vOTUs past.

To investigate putative host ranges, we bioinformatically linked vOTUs to MAGs, using CRISPR arrays [**37**]. All 12,826 vOTUs and 219 MAGs (210 Bacteria and 9 Archaea) were used for this analysis. A total of 29,709 CRISPR arrays was recovered from the metagenomes, and 683 virus-host linkages were predicted between 378 vOTUs and 53 MAGs. All identified host MAGs were bacteria and could be classified to at least the phylum level, with Proteobacteria and Actinobacteriota among the most commonly reconstructed MAGs (Figure 4.4A). Samples from medium to extremely saline wetlands had significantly more CRISPR arrays and spacers than others, perhaps suggesting increased viral predation, but there was no significant relationship between the number of CRISPR arrays or spacers and the number of vOTUs in a given sample (Supplementary Figure 5). The average MAG was linked to 13 vOTUs, indicating that wetland prokaryotic populations can be infected by (or otherwise interact with [**38**]) multiple, diverse viral species. On average, each vOTU was linked to four MAGs, and 164 vOTUs (45% of those with predicted hosts) had putative linkages to MAGs in different phyla (Figure 4.4A, 4.4B). The average vOTU was linked to MAGs in two phyla, and when only considering vOTUs linked to more than one MAG, vOTUs were linked to MAGs across three or more phyla on average.

These results suggest either that CRISPR spacer matches to viral proto-spacers are imperfect for predicting virus-host linkages associated with infections in these systems, or that wetland viruses have much broader host ranges than previously appreciated. Recent studies have suggested that viral interactions with hosts may be far less specific than previously understood, with viruses infecting (or otherwise interacting with) prokaryotes across different phyla [**38, 39**]. The mechanisms that could routinely enable viruses to infect different phyla are unknown, but recent evidence for diverse plasmid-dependent phages [**40**] (which target conjugation proteins encoded by horizontally transferrable plasmids) offers one interesting possibility for cross-infection that bears further exploration. Cross-phylum CRISPR linkages could also reflect non-specific interactions (e.g., uptake of viral particles or DNA by non-primary hosts, or horizontal transfer of CRISPR regions), as opposed to infections, and these interactions have been suggested to be more common than previously appreciated [**38**].

To investigate viral evolution in response to host immunity, we calculated the allelic variance within and outside of the viral genomic regions linked to CRISPR spacers, using the originally assembled vOTU sequence as the reference for SNP identification for each vOTU. Viral genomic regions with a CRISPR-spacer match had on average 5.6 SNPs/Kbp, whereas the genome outside of the match had on average 3.3 SNPs/Kbp, indicating more allelic variance in CRISPR-targeted regions, compared to the rest of the viral genome. This has been seen previously, for example in Streptococcus thermophilus phage-host coevolution experiments and in an acid mine drainage system [**41**, **42**], and it suggests increased phage genome diversification in CRISPR targeted regions to promote immune evasion. Of the predicted proteins in the CRISPR-targeted viral genomic regions with SNPs, 87% were annotated as hypothetical proteins, and 9% were putative major tail proteins. A significantly larger proportion of putative tail proteins were found in these regions than were annotated as putative major tail proteins in the whole dataset (0.93%, $p < 0.00001$, Z-test). This suggests that there is selection for accelerated evolution in viral genomic regions targeted by CRISPRs, particularly in tail proteins likely involved in attachment to host cell receptors [**43**]. Evidence for higher mutation rates in phage tail protein genes is presumably due to viral adaptation to changes in host cell receptors to facilitate attachment, as previously suggested [**44**, **45**].

4.3. Conclusions

Here, we analyzed dsDNA viral communities from the Bodega Bay, California wetland ecosystem and showed significant differences in viral community composition across seven wetland sites, with evidence for dispersal, dispersal limitation, and habitat filtering as underlying drivers of the observed patterns. Although wetland viral communities differed predominantly by location within Bodega Bay, perhaps reflecting local dispersal limitation, the two wetland sites with the most homogeneous communities had the highest soil moisture content, suggesting hydrologic mixing and more opportunities for within-site dispersal with increasing moisture content. Local wetland viral communities were secondarily structured by habitat characteristics, such as plant community composition and soil salinity, indicating environmental filtering, perhaps by way of host adaptation. A small fraction (1.5%) of the vOTUs were previously recovered elsewhere, with global biogeographical patterns largely linked to habitat characteristics; marine vOTUs tended to be recovered in saline wetlands, freshwater vOTUs in non-saline wetlands, and soil vOTUs across wetland habitats.

In addition to dispersal and environmental filtering, eco-evolutionary dynamics (e.g., diversification and/or compositional shifts among dominant allelic variants) contributed to local and global viral biogeographical patterns. Pairwise ANI % between dominant allelic variants (sub-populations) differed significantly between years and over the four-month timescale of this study. In addition, Bodega Bay vOTU variants tended to be more divergent from reference vOTUs recovered elsewhere globally than from reference sequences assembled from Bodega Bay. The observed greater divergence over larger spatiotemporal scales is consistent with patterns of 'isolation by distance', whereby variants closer together in time and/or space likely had greater opportunities for gene flow. On a global scale, this may reflect local diversification and global dispersal limitation of most variants. Our limited ability to link viruses to their hosts (a limitation of the current state of the field) makes the contributions of virus-host co-evolutionary dynamics to biogeographic patterns difficult to evaluate, but we did see evidence for virus-host interactions spanning multiple phyla. Taken together, these results highlight dispersal, environmental filtering, and eco-evolutionary dynamics as likely drivers of both local and global wetland viral biogeographical patterns, expanding our understanding of the highly diverse and dynamic global soil virosphere.

4.4. Materials and Methods

4.4.1. Field site and sample collection.

Samples were collected three times over six months at the University of California, Davis Bodega Marine Reserve, in seven wetland soil ecosystems within the reserve (Supplementary Table 2). Sample collections were performed on March 17th, May 13th,and July 15th, 2021 (T1, T2 and T3, respectively) from each of seven distinct wetland sites.The plant community at each site was used as an indicator for soil salinity (Supplementary Table 6), such that the seven wetland sites were initially selected to represent three low-salinity and four high-salinity habitats, but subsequent analyses revealed more nuance in these habitat types (see main text). At each time point, three replicate surface soil samples (0-15 cm deep, 2.5 x 2.5 cm square area) were collected per wetland site, using a soil knife. The soil within each sample was homogenized and stored at -80 °C until further processing.

4.4.2. Virome DNA extraction, library construction, and sequencing.

Soil virions were enriched using a modified version of a previously published protocol [**46**]. For each sample, 10 grams of soil were suspended in 30 mL of protein-supplemented phosphate-buffered saline solution (PPBS: 2% bovine serum albumin, 10% phosphate-buffered saline, 1% potassium citrate, and 150 mM $MgSO_4$ in ultrapure water), briefly vortexed, placed on an orbital shaker (30 min, 400 rpm, 4 °C), and then centrifuged (10 min, 3,095 x g, 4 °C). Supernatant was then centrifuged twice (8 min, 10,000 x g, 4 °C) to remove residual soil particles. The purified supernatants were then filtered through a 0.22 μm polyethersulfone membrane to remove most cells. The resulting filtrate was ultracentrifuged (2 hrs 25 min, 32,000 rpm, 4 °C) to pellet the virions, using an Optima LE-80K ultracentrifuge with a 50.2 Ti rotor (Beckman-Coulter Life Sciences). Supernatants were decanted, and pellets were resuspended in 100 μl of ultrapure water. DNase treatment was not performed, as soil samples were stored frozen prior to processing, due to COVID-19 lockdown restrictions, and avoiding DNase treatment on such samples has been shown to improve viromic DNA yields without substantially compromising the viral 'signal' in the data [**47**]. DNA was extracted from the viral-enriched fraction, using the DNeasy PowerSoil Pro kit (Qiagen, Hilden, Germany), following the manufacturer's instructions, with an added step of a 10-min incubation at 65 °C before the bead-beating step. Libraries were constructed by the UC Davis DNA Technologies Core, using the DNA Hyper Prep library kit (Kapa Biosystems-Roche, Basel, Switzerland), and paired-end 150 bp sequencing was done using the NovaSeq S4 platform (Illumina) to an approximate depth of 10 Gbp per virome.

4.4.3. Total DNA extraction, library construction, and sequencing.

Total DNA was extracted from 0.25 g of soil per sample with the DNeasy PowerSoil Pro kit (Qiagen, Hilden, Germany), following the manufacturer's instructions, with an added step of a 10-min incubation at 65 °C before the bead-beating step. Libraries were constructed by the UC Davis DNA Technologies Core, using the DNA Hyper Prep library kit (Kapa Biosystems-Roche, Basel, Switzerland), and paired-end 150 bp sequencing was done using the NovaSeq S4 platform (Illumina) to approximate depth of 20 Gbp per total metagenome.

4.4.4. Soil chemistry and moisture.

Soil moisture was defined by calculating the gravimetric water content of the soil. Soil chemistry measurements were performed by Ward Laboratories (Kearney, NE, USA). Briefly, soil pH and soluble salts were measured using a 1:1 soil:water suspension. Soil organic matter was calculated as percent mass loss on ignition. Nitrate was measured using a KCl extraction. Potassium, calcium, magnesium and sodium were measured using an ammonium acetate extraction. Zinc, iron, manganese and copper were measured using a DTPA extraction. Phosphorus was measured using the Olsen method and sulfate was measured using a Mehlich-3 extraction.

4.4.5. Virome bioinformatic processing.

Reads were trimmed using Trimmomatic v0.39 [**48**] to remove Illumina adapters and for quality trimming, using paired-end trimming, a sliding window size of 3:40, and a minimum read length of 50 bp. PhiX sequences were removed using BBDuk, from the BBMap v38-72 package [**49**], using k=31 and hdist=1. De novo assemblies were generated separately for each virome from the quality-trimmed, phiX-free reads, using MEGAHIT v1.0.6 [**50**], with k-min of 27, presets meta-large, and a minimum contig length of 1000 bp. Contigs were renamed, using the rename command from the BBMap package [**49**], using standard settings, and only contigs $\geq$10kbp were retained, using reformat from BBmap with the setting minlength=10000. Viral contigs were predicted using VIBRANT v1.2.0 [**51**], in virome mode and retained for downstream analyses if VIBRANT classified the contig as viral. Viral contigs were dereplicated into vOTUs using dRep v3.2.0 [**52**] at 95% ANI with a minimum coverage threshold of 85%, using the ANImf algorithm. Reads were mapped to the vOTUs, using Bowtie2 v2.4.2 [**53**] in sensitive mode, and the resulting samfiles were converted to bamfiles via SAMtools v1.15.1 [**54**]. A coverage Table was produced using CoverM v0.6.1 [**55**], using CoverM contig with the mean coverage and a minimum covered fraction (breadth) of 75% (Supplementary Table 4). Reads were subsequently mapped back to our PIGEONv2.0 database, using CoverM with the same settings.

4.4.6. Total metagenome bioinformatic processing.

Read trimming, PhiX removal, and assembly were done the same way as for the viromes. Contigs were renamed using using the rename command from the BBMap package [**49**], using standard settings, and only contigs $\geq$2 kbp were retained, using reformat from BBmap with the settings minlength=2000. CD-hit v2007-013 [**56**] was used to deduplicate the contigs at approximately 99% ANI, using the -c 0.99 -aS 0.99 settings. Bowtie2 v2.4.2 [**53**] was used to map reads to the contigs, using sensitive mode, and the samfiles were converted to bamfiles using SAMtools v1.15.1 [**54**]. A depth file for binning was created using MetaBAT2 v2.12.1 [**57**], using jgi_summarize_bam_contig_depths. Bins were then created using MetaBAT2 v2.12.1, using standard settings. dRep v3.2.0 [**52**] was used to dereplicate the bins, using primary clustering at 90%, secondary clustering at 95%, coverage method larger, a contamination threshold of 5%, and a coverage threshold of 30% [**58**]. CheckM v1.0.13 [**59**] was used to estimate completeness and contamination of genome bins, and bins $\geq$ 50% complete and $\leq$ 10% contaminated were retained [**28**]. RefineM v0.1.2 [**60**] was used to refine the recovered bins, using standard settings. Reads were mapped to these bins (metagenome-assembled genomes, MAGs) using Bowtie2 v2.4.2 [**53**] with default settings except setting –min-covered-fraction to 0.5 [**61**] (Supplementary Table 5). Phylogenetic trees were constructed using gtdbtk v2.1.0 [**62**], using the classify-wf command for phylogenetic inference and for aligning the identified marker genes. After this, gtdbtk infer was used to create the phylogenetic tree, using standard settings.

4.4.7. Recovery and analysis of 16S rRNA gene sequences from metagenomes.

SortMeRNA v4.2.0 [**63**] was used against the bacterial and archaeal SILVA databases [**64**] to recover reads containing 16S rRNA gene sequences from the total soil metagenomes. RDP tools v11 [**65**] was used to taxonomically classify the sequences, using the RDP database v18 as a reference [**66**]. A count Table of the 16S rRNA gene OTUs was generated using the hier2phyloseq() function from the RDPutils package [**67**].

4.4.8. PIGEONv2.0.

To build further upon PIGEON 1.0 [**15**], we added more viral sequences mostly from in-house soil data, both published [**18**, **22**, **47**] and currently unpublished, and from recent publications of viral ecology in soil [**26**], lakes [**34**, **68**] and oceans [**69**, **70**]. We also mined a total soil metagenome dataset for viral sequences [**71**]. A prefix was added to all sequence headers, in order to quickly identify what dataset the original sequence came from (Supplementary table 9). All sequences were dereplicated using cd-hit 2007-0131 [**56**], because the dataset was too large to use other programs for dereplication.

4.4.9. Microdiversity profiles.

Within-population genetic diversity was calculated using inStrain v1.4.0 [**35**]. The bam files created by bowtie2 from the viromes were used as input for the inStrain profile option to identify divergent sites for each of the vOTUs. Variants were only called if they had a minimum coverage of 5 reads. MAG population genetic diversity was calculated the same way, using the bam files created by bowtie2 from the total metagenomes as input for InStrain.

4.4.10. CRISPR-spacer analyses for virus-host linkages.

Crass v1.0.1 [**37**] was used to assemble spacer and repeat sequences in the total metagenomes, using -l 4. All spacer sequences were then compared to the vOTUs, using blastN v2.7.1 [**72**], retaining hits with fewer than two mismatches and >95% nucleotide identity. All repeat sequences were compared to the MAGs using blastN, retaining hits that had no mismatches and 100% nucleotide identity (Supplementary Table 8).

4.4.11. Data analysis and visualization.

All statistical analyses were done using R v 4.1.0 [**73**]. Analysis for viral community composition were done on the mean coverage vOTU abundance Table, unless otherwise noted. Bray-Curtis dissimilarities were calculated on log-transformed relative abundances, using the vegdist function from the vegan package v2.6-2 [**74**]. PERMANOVA analyses were done using the adonis2 function from vegan. Principal coordinates analyses were performed with the pcoa() function from ape v5.4-2 [**75**]. The BIO-ENV analysis was done using the bioenv function from vegan. Co-occurrence analyses for vOTUs, MAGs and 16S rRNA OTUs were done using the coocur package in R [**76**], using a presence-absence version of the abundance Tables. Only significantly positive co-occurrences ($p<0.001$) were used for visualization. Co-occurrence networks were visualized using Cytoscape v3.7.1 [**77**], using the edge-weighted spring embedded model, placing vOTUs that co-occur more frequently in closer proximity to each other in the figure. Upset plots were created using the UpSetR package v1.4.0 [**78**], using a presence-absence version of the vOTU abundance Table. All maps were created using the R package ggmaps [**79**]. Pie charts and bar charts were created with Python v3.8, using matplotlib v3.4.2 [**80**] and seaborn v0.11.2. The phylogenetic tree was created using the iTOL website [**81**], and the CRISPR-repeat network linking viruses to hosts was created using Cytoscape. All other plots were created, using the R package ggplot2 v3.3.5 [**82**]. Correlation tests between community Jaccard Dissimilarity and spatial or environmental distance were done using the cor.test() function, using the pearson method with the alternative parameter set to two-sided. The linear regression slopes were calculated using the lm function, as has been done previously [**18**]. All scripts are available at https://github.com/AnneliektH/BodegaBay2021.

Bibliography

[1] A M Nahlik and M S Fennessy. Carbon storage in US wetlands. *Nat. Commun.*, 7:13835, December 2016.

[2] Charles S Hopkinson, Wei-Jun Cai, and Xinping Hu. Carbon sequestration in wetland dominated coastal systems—a global sink of rapidly diminishing magnitude. *Current Opinion in Environmental Sustainability*, 4(2):186–194, May 2012.

[3] William J Mitsch and James G Gosselink. The value of wetlands: importance of scale and landscape setting. *Ecol. Econ.*, 35(1):25–33, October 2000.

[4] William J Mitsch, Blanca Bernal, and Maria E Hernandez. Ecosystem services of wetlands. *Int. J. Biodivers. Sci. Eco. Srvcs. Mgmt.*, 11(1):1–4, January 2015.

[5] Stephanie A Yarwood. The role of wetland microorganisms in plant-litter decomposition and soil organic matter formation: a critical review. *FEMS Microbiol. Ecol.*, 94(11), November 2018.

[6] Minghua Zhou, Klaus Butterbach-Bahl, Harry Vereecken, and Nicolas Brüggemann. A meta-analysis of soil salinization effects on nitrogen pools, cycles and fluxes in coastal ecosystems. *Glob. Chang. Biol.*, 23(3):1338–1352, March 2017.

[7] Marcelo Ardón, Ashley M Helton, and Emily S Bernhardt. Salinity effects on greenhouse gas emissions from wetland soils are contingent upon hydrologic setting: a microcosm experiment. *Biogeochemistry*, 140(2):217–232, September 2018.

[8] Guangliang Zhang, Junhong Bai, Christoph C Tebbe, Qingqing Zhao, Jia Jia, Wei Wang, Xin Wang, and Lu Yu. Salinity controls soil microbial community structure and function in coastal estuarine wetlands. *Environ. Microbiol.*, 23(2):1020–1037, February 2021.

[9] Andy J Green, Paloma Alcorlo, Edwin Thm Peeters, Edward P Morris, José L Espinar, Miguel Angel Bravo-Utrera, Javier Bustamante, Ricardo Díaz-Delgado, Albert A Koelmans, Rafael Mateo, Wolf M Mooij, Miguel Rodríguez-Rodríguez, Egbert H van Nes, and Marten Scheffer. Creating a safe operating space for wetlands in a changing climate. *Front. Ecol. Environ.*, 15(2):99–107, March 2017.

[10] Paula Dalcin Martins, Robert E Danczak, Simon Roux, Jeroen Frank, Mikayla A Borton, Richard A Wolfe, Marie N Burris, and Michael J Wilkins. Viral and metabolic controls on high rates of microbial sulfur and carbon cycling in wetland ecosystems. *Microbiome*, 6(1):138, August 2018.

[11] Joanne B Emerson. Soil viruses: A new hope. *mSystems*, 4(3), May 2019.

[12] Carmody K McCalley, Ben J Woodcroft, Suzanne B Hodgkins, Richard A Wehr, Eun-Hae Kim, Rhiannon Mondav, Patrick M Crill, Jeffrey P Chanton, Virginia I Rich, Gene W Tyson, and Scott R Saleska. Methane dynamics regulated by microbial community response to permafrost thaw. *Nature*, 514(7523):478–481, October 2014.

[13] Ricardo Cavicchioli, William J Ripple, Kenneth N Timmis, Farooq Azam, Lars R Bakken, Matthew Baylis, Michael J Behrenfeld, Antje Boetius, Philip W Boyd, Aimée T Classen, Thomas W Crowther, Roberto Danovaro, Christine M Foreman, Jef Huisman, David A Hutchins, Janet K Jansson, David M Karl, Britt Koskella, David B Mark Welch, Jennifer B H Martiny, Mary Ann Moran, Victoria J Orphan, David S Reay, Justin V Remais, Virginia I Rich, Brajesh K Singh, Lisa Y Stein, Frank J Stewart, Matthew B Sullivan, Madeleine J H van Oppen, Scott C Weaver, Eric A Webb, and Nicole S Webster. Scientists' warning to humanity: microorganisms and climate change. *Nat. Rev. Microbiol.*, 17(9):569–586, September 2019.

[14] Ben J Woodcroft, Caitlin M Singleton, Joel A Boyd, Paul N Evans, Joanne B Emerson, Ahmed A F Zayed, Robert D Hoelzle, Timothy O Lamberton, Carmody K McCalley, Suzanne B Hodgkins, Rachel M Wilson, Samuel O Purvine, Carrie D Nicora, Changsheng Li, Steve Frolking, Jeffrey P Chanton, Patrick M Crill, Scott R Saleska, Virginia I Rich, and Gene W Tyson. Genome-centric view of carbon processing in thawing permafrost. *Nature*, 560(7716):49–54, August 2018.

[15] Anneliek M Ter Horst, Christian Santos-Medellín, Jackson W Sorensen, Laura A Zinke, Rachel M Wilson, Eric R Johnston, Gareth Trubl, Jennifer Pett-Ridge, Steven J Blazewicz, Paul J Hanson, Jeffrey P Chanton, Christopher W Schadt, Joel E Kostka, and Joanne B Emerson. Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. *Microbiome*, 9(1):233, November 2021.

[16] Joanne B Emerson, Simon Roux, Jennifer R Brum, Benjamin Bolduc, Ben J Woodcroft, Ho Bin Jang, Caitlin M Singleton, Lindsey M Solden, Adrian E Naas, Joel A Boyd, Suzanne B Hodgkins, Rachel M Wilson, Gareth Trubl, Changsheng Li, Steve Frolking, Phillip B Pope, Kelly C Wrighton, Patrick M Crill, Jeffrey P Chanton, Scott R Saleska, Gene W Tyson, Virginia I Rich, and Matthew B Sullivan. Host-linked soil viral ecology along a permafrost thaw gradient. *Nature Microbiology*, 3(8):870–880, July 2018.

[17] Gareth Trubl, Jeffrey A Kimbrel, Jose Liquet-Gonzalez, Erin E Nuccio, Peter K Weber, Jennifer Pett-Ridge, Janet K Jansson, Mark P Waldrop, and Steven J Blazewicz. Active virus-host interactions at sub-freezing temperatures in arctic peat soil. *Microbiome*, 9(1):208, October 2021.

[18] Christian Santos-Medellín, Katerina Estera-Molina, Mengting Yuan, Jennifer Pett-Ridge, Mary K Firestone, and Joanne B Emerson. Spatial turnover of soil viral populations and genotypes overlain by cohesive responses to moisture in grasslands. *Proc. Natl. Acad. Sci. U. S. A.*, 119(45):e2209132119, November 2022.

[19] Gareth Trubl, Ho Bin Jang, Simon Roux, Joanne B Emerson, Natalie Solonenko, Dean R Vik, Lindsey Solden, Jared Ellenbogen, Alexander T Runyon, Benjamin Bolduc, Ben J Woodcroft, Scott R Saleska, Gene W Tyson, Kelly C Wrighton, Matthew B Sullivan, and Virginia I Rich. Soil viruses are underexplored players in ecosystem carbon processing, 2018.

[20] David Paez-Espino, Emiley A Eloe-Fadrosh, Georgios A Pavlopoulos, Alex D Thomas, Marcel Huntemann, Natalia Mikhailova, Edward Rubin, Natalia N Ivanova, and Nikos C Kyrpides. Uncovering earth's virome. *Nature*, 536(7617):425–430, August 2016.

[21] Christian Santos-Medellin, Laura A Zinke, Anneliek M Ter Horst, Danielle L Gelardi, Sanjai J Parikh, and Joanne B Emerson. Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *ISME J.*, 15(7):1956–1970, July 2021.

[22] Devyn M Durham, Ella T Sieradzki, Anneliek M ter Horst, Christian Santos-Medellín, C Winston A Bess, Sara E Geonczy, and Joanne B Emerson. Substantial differences in soil viral community composition within and among four northern california habitats. *ISME Communications*, 2(1):1–5, October 2022.

[23] Luke S Hillary, Evelien M Adriaenssens, David L Jones, and James E McDonald. RNA-viromics reveals diverse communities of soil RNA viruses with the potential to affect grassland ecosystems across multiple trophic levels. *ISME Commun*, 2:34, April 2022.

[24] Ruonan Wu, Michelle R Davison, William C Nelson, Emily B Graham, Sarah J Fansler, Yuliya Farris, Sheryl L Bell, Iobani Godinez, Jason E Mcdermott, Kirsten S Hofmockel, and Janet K Jansson. DNA viral diversity, abundance, and functional potential vary across grassland soils with a range of historical moisture regimes. *MBio*, 12(6):e0259521, December 2021.

[25] Clement Coclet, Patrick O Sorensen, Ulas Karaoz, Shi Wang, Eoin L Brodie, Emiley A Eloe-Fadrosh, and Simon Roux. Virus diversity and activity is driven by snowmelt and host dynamics in a high-altitude watershed soil ecosystem. March 2023.

[26] Sungeun Lee, Jackson W Sorensen, Robin L Walker, Joanne B Emerson, Graeme W Nicol, and Christina Hazard. Soil ph influences the structure of virus communities at local and global scales. *Soil Biol. Biochem.*, 166:108569, March 2022.

[27] Simon Roux, Evelien M Adriaenssens, Bas E Dutilh, Eugene V Koonin, Andrew M Kropinski, Mart Krupovic, Jens H Kuhn, Rob Lavigne, J Rodney Brister, Arvind Varsani, Clara Amid, Ramy K Aziz, Seth R Bordenstein, Peer Bork, Mya Breitbart, Guy R Cochrane, Rebecca A Daly, Christelle Desnues, Melissa B Duhaime, Joanne B Emerson, François Enault, Jed A Fuhrman, Pascal Hingamp, Philip Hugenholtz, Bonnie L Hurwitz, Natalia N Ivanova, Jessica M Labonté, Kyung-Bum Lee, Rex R Malmstrom, Manuel Martinez-Garcia, Ilene Karsch Mizrachi, Hiroyuki Ogata, David Páez-Espino, Marie-Agnès Petit, Catherine Putonti, Thomas Rattei, Alejandro Reyes, Francisco Rodriguez-Valera, Karyna Rosario, Lynn Schriml, Frederik Schulz, Grieg F Steward, Matthew B Sullivan, Shinichi Sunagawa, Curtis A Suttle, Ben Temperton, Susannah G Tringe, Rebecca Vega Thurber, Nicole S Webster, Katrine L Whiteson, Steven W Wilhelm, K Eric Wommack, Tanja Woyke, Kelly C Wrighton, Pelin Yilmaz, Takashi Yoshida, Mark J Young, Natalya Yutin, Lisa Zeigler Allen, Nikos C Kyrpides, and Emiley A Eloe-Fadrosh. Minimum information about an uncultivated virus genome (MIUViG). *Nat. Biotechnol.*, 37(1):29–37, January 2019.

[28] Alexandre Almeida, Alex L Mitchell, Miguel Boland, Samuel C Forster, Gregory B Gloor, Aleksandra Tarkowska, Trevor D Lawley, and Robert D Finn. A new genomic blueprint of the human gut microbiota. *Nature*, 568(7753):499–504, April 2019.

[29] D L Corwin and E Scudiero. Chapter one - review of soil salinity assessment for agriculture across multiple scales using proximal and/or remote sensors. In Donald L Sparks, editor, *Advances in Agronomy*, volume 158, pages 1–130. Academic Press, January 2019.

[30] Elisabeth N Bui. Causes of soil salinization, sodification, and alkalinization. In *Oxford Research Encyclopedia of Environmental Science*. 2017.

[31] Sergey Shabala. Learning from halophytes: physiological basis and strategies to improve abiotic stress tolerance in crops. *Ann. Bot.*, 112(7):1209–1221, November 2013.

[32] Irwin A Ungar. Are biotic factors significant in influencing the distribution of halophytes in saline habitats? *Bot. Rev.*, 64(2):176–199, April 1998.

[33] Xiaolong Liang, Yusong Wang, Ying Zhang, Jie Zhuang, and Mark Radosevich. Viral abundance, community structure and correlation with bacterial community in soils of different cover plants. *Appl. Soil Ecol.*, 168:104138, December 2021.

[34] F H Coutinho, P J Cabello-Yeves, R Gonzalez-Serrano, R Rosselli, M López-Pérez, T I Zemskaya, A S Zakharenko, V G Ivanov, and F Rodriguez-Valera. New viral biogeochemical roles revealed through metagenomic analysis of lake baikal. *Microbiome*, 8(1):163, November 2020.

[35] Matthew R Olm, Alexander Crits-Christoph, Keith Bouma-Gregson, Brian A Firek, Michael J Morowitz, and Jillian F Banfield. instrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.*, 39(6):727–736, June 2021.

[36] China A Hanson, Jed A Fuhrman, M Claire Horner-Devine, and Jennifer B H Martiny. Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat. Rev. Microbiol.*, 10(7):497–506, May 2012.

[37] Connor T Skennerton, Michael Imelfort, and Gene W Tyson. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Res.*, 41(10):e105, May 2013.

[38] Yunha Hwang, Simon Roux, Clément Coclet, Sebastian J E Krause, and Peter R Girguis. Viruses interact with hosts that span distantly related microbial domains in dense hydrothermal mats. *Nat Microbiol*, April 2023.

[39] Nikhil A George and Laura A Hug. CRISPR-resolved virus-host interactions in a municipal landfill include non-specific viruses, hyper-targeted viral populations, and interviral conflicts. *Sci. Rep.*, 13(1):5611, April 2023.

[40] Natalia Quinones-Olvera, Siân V Owen, Lucy M McCully, Maximillian G Marin, Eleanor A Rand, Alice C Fan, Oluremi J Martins Dosumu, Kay Paul, Cleotilde E Sanchez Castaño, Rachel Petherbridge, Jillian S Paull, and Michael Baym. Diverse and abundant viruses exploit conjugative plasmids. *bioRxiv*, March 2023.

[41] David Paez-Espino, Itai Sharon, Wesley Morovic, Buffy Stahl, Brian C Thomas, Rodolphe Barrangou, and Jillian F Banfield. CRISPR immunity drives rapid phage genome evolution in streptococcus thermophilus. *MBio*, 6(2), April 2015.

[42] Christine L Sun, Rodolphe Barrangou, Brian C Thomas, Philippe Horvath, Christophe Fremaux, and Jillian F Banfield. Phage mutations in response to CRISPR diversification in a bacterial population. *Environ. Microbiol.*, 15(2):463–470, February 2013.

[43] Britt Koskella and Michael A Brockhurst. Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol. Rev.*, 38(5):916–931, September 2014.

[44] A Betts, C Gray, M Zelek, R C MacLean, and K C King. High parasite diversity accelerates host adaptation and diversification. *Science*, 360(6391):907–911, May 2018.

[45] Pauline D Scanlan, Alex R Hall, Laura D C Lopez-Pascua, and Angus Buckling. Genetic basis of infectivity evolution in a bacteriophage. *Mol. Ecol.*, 20(5):981–989, March 2011.

[46] Pauline C Göller, Jose M Haro-Moreno, Francisco Rodriguez-Valera, Martin J Loessner, and Elena Gómez-Sanz. Uncovering a hidden diversity: optimized protocols for the extraction of dsDNA bacteriophages from soil, 2020.

[47] Jackson W Sorensen, Laura A Zinke, Anneliek M Ter Horst, Christian Santos-Medellín, Alena Schroeder, and Joanne B Emerson. DNase treatment improves viral enrichment in agricultural soil viromes. *mSystems*, page e0061421, September 2021.

[48] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, August 2014.

[49] Brian Bushnell. BBMap: a fast, accurate, splice-aware aligner. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2014.

[50] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31(10):1674–1676, May 2015.

[51] Kristopher Kieft, Zhichao Zhou, and Karthik Anantharaman. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, 8(1):90, June 2020.

[52] Matthew R Olm, Christopher T Brown, Brandon Brooks, and Jillian F Banfield. drep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.*, 11(12):2864–2868, December 2017.

[53] B Longmead and S L Salzberg. Fast gapped-read alignment with bowtie2. *Nat. Methods*, 9(4):357–359, 2012.

[54] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[55] Ben J Woodcroft. CoverM: Read coverage calculator for metagenomics.

[56] Ying Huang, Beifang Niu, Ying Gao, Limin Fu, and Weizhong Li. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5):680–682, March 2010.

[57] Dongwan D Kang, Jeff Froula, Rob Egan, and Zhong Wang. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities, 2015.

[58] Edoardo Pasolli, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, Paolo Manghi, Adrian Tett, Paolo Ghensi, Maria Carmen Collado, Benjamin L Rice, Casey DuLong, Xochitl C Morgan, Christopher D Golden, Christopher Quince, Curtis Huttenhower, and Nicola Segata. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, 176(3):649–662.e20, January 2019.

[59] Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, 25(7):1043–1055, July 2015.

[60] Donovan H Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J Woodcroft, Paul N Evans, Philip Hugenholtz, and Gene W Tyson. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol*, 2(11):1533–1542, November 2017.

[61] Matthew R Olm, Dylan Dahan, Matthew M Carter, Bryan D Merrill, Feiqiao B Yu, Sunit Jain, Xiandong Meng, Surya Tripathi, Hannah Wastyk, Norma Neff, Susan Holmes, Erica D Sonnenburg, Aashish R Jha, and Justin L Sonnenburg. Robust variation in infant gut microbiome assembly across a spectrum of lifestyles. *Science*, 376(6598):1220–1223, June 2022.

[62] Pierre-Alain Chaumeil, Aaron J Mussig, Philip Hugenholtz, and Donovan H Parks. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics*, 38(23):5315–5316, November 2022.

[63] Evguenia Kopylova, Laurent Noé, and Hélène Touzet. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, 28(24):3211–3217, December 2012.

[64] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, 41(Database issue):D590–6, January 2013.

[65] Qiong Wang, George M Garrity, James M Tiedje, and James R Cole. Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, 73(16):5261–5267, August 2007.

[66] James R Cole, Qiong Wang, Jordan A Fish, Benli Chai, Donna M McGarrell, Yanni Sun, C Titus Brown, Andrea Porras-Alfaro, Cheryl R Kuske, and James M Tiedje. Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.*, 42(Database issue):D633–42, January 2014.

[67] John Quensen. Jfq3/RDPutils: R utilities for processing RDPTool output. `https://rdrr.io/github/jfq3/RDPutils/`, November 2019. Accessed: 2022-12-5.

[68] Myriam Labbé, Catherine Girard, Warwick F Vincent, and Alexander I Culley. Extreme viral partitioning in a Marine-Derived high arctic lake. *mSphere*, 5(3), May 2020.

[69] Dean Vik, Maria Consuelo Gazitúa, Christine L Sun, Ahmed A Zayed, Montserrat Aldunate, Margaret R Mulholland, Osvaldo Ulloa, and Matthew B Sullivan. Genome-resolved viral ecology in a marine oxygen minimum zone. *Environ. Microbiol.*, 23(6):2858–2874, June 2021.

[70] Zexin Li, Donald Pan, Guangshan Wei, Weiling Pi, Chuwen Zhang, Jiang-Hai Wang, Yongyi Peng, Lu Zhang, Yong Wang, Casey R J Hubert, and Xiyang Dong. Deep sea sediments associated with cold seeps are a subsurface reservoir of viral diversity. *ISME J.*, 15(8):2366–2378, August 2021.

[71] Jin Xu, Yunzeng Zhang, Pengfan Zhang, Pankaj Trivedi, Nadia Riera, Yayu Wang, Xin Liu, Guangyi Fan, Jiliang Tang, Helvécio D Coletta-Filho, Jaime Cubero, Xiaoling Deng, Veronica Ancona, Zhanjun Lu, Balian Zhong, M Caroline Roper, Nieves Capote, Vittoria Catara, Gerhard Pietersen, Christian Vernière, Abdullah M Al-Sadi, Lei Li, Fan Yang, Xun Xu, Jian Wang, Huanming Yang, Tao Jin, and Nian Wang. The structure and function of the global citrus rhizosphere microbiome, 2018.

[72] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, October 1990.

[73] Team RCore. R: A language and environment for statistical computing. R foundation for statistical computing, vienna, austria, 2016.

[74] J Oksanen, F G Blanchet, M Friendly, R Kindt, P Legendre, D McGlinn, P R Minchin, R B O'Hara, G L Simpson, P Solymos, and Others. vegan: Community ecology package. R package version 2.5-2. 2018, 2018.

[75] Emmanuel Paradis and Klaus Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3):526–528, February 2019.

[76] Daniel M Griffith, Joseph A Veech, and Charles J Marsh. cooccur: Probabilistic species Co-Occurrence analysis in R. *J. Stat. Softw.*, 69:1–17, February 2016.

[77] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, November 2003.

[78] Jake R Conway, Alexander Lex, and Nils Gehlenborg. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18):2938–2940, September 2017.

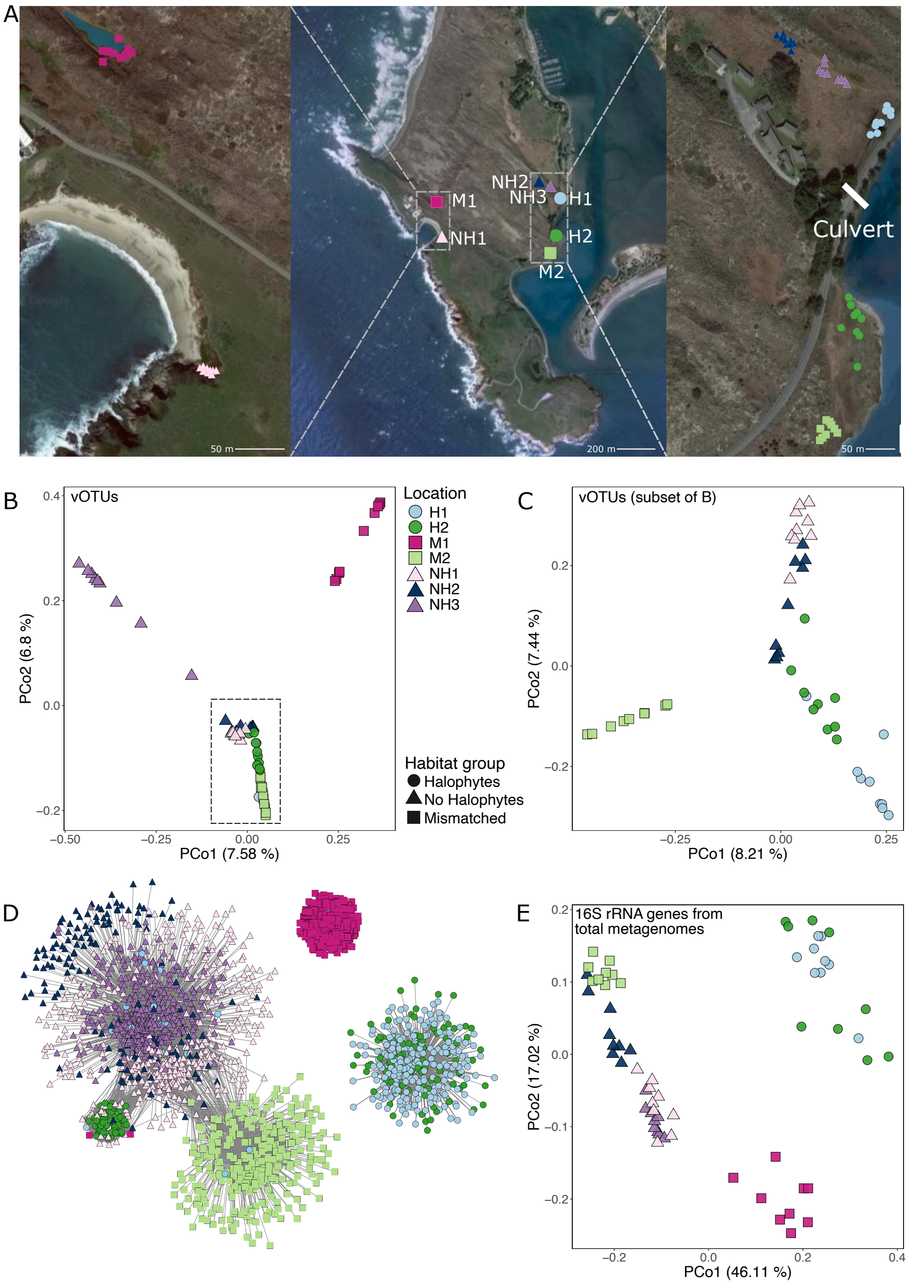
[79] D Kahle and H Wickham. ggmap: Spatial visualization with ggplot2. the R journal, 5 (1), 144-161. *URL https://journal. r-project. org/archive/2013-1/kahle.*

[80] Hunter. Matplotlib: A 2D graphics environment. 9:90–95, May 2007.

[81] Ivica Letunic and Peer Bork. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.*, 49(W1):W293–W296, July 2021.

[82] H Wickham. ggplot2: elegant graphics for data analysis. data. 2016.

Figures and tables



(Continued on the following page)

Figure 4.1: Sampling design and overarching compositional patterns for Bodega Bay viral and prokaryotic communities. A) Sampling locations for all Bodega Bay samples. Center: locations of the seven wetland sites within the Bodega Marine Reserve, Left and Right: locations of each of the nine samples per site (a zoomed in view of each site with individual sample labels is in Supplementary Figure 1). Per the legend below the images, circles correspond to locations with halophyte vegetation and saline soils, triangles correspond to locations without halophytes and non-saline soil, and squares correspond to mismatched locations. The 'culvert' label indicates the location of a human-made pipe below the road that allows for water movement. B-C) Principal coordinates analysis (PCoA), based on Bray-Curtis dissimilarities derived from the table of vOTU abundances (read mapping to vOTUs). Each point is one sample (one virome), with viral communities from B) all 63 viromes, and C) the 45 viromes indicated by the dashed rectangle in B. Panel C is a new PCoA to better show separation among overlapping samples in B. D) Co-occurrence network of vOTUs detected in more than one Bodega Bay virome, colored by the site in which they were most abundant (had the highest average per-bp coverage depth). Nodes represent vOTUs, and edges represent a significant co-occurrence between the vOTUs, calculated using a probabilistic co-occurrence model with the R package cooccur. E) PCoA based on Bray-Curtis dissimilarities of 16S rRNA gene OTU community composition from 63 total metagenomes. For all PCoA plots (B, C, E), the percent variance explained by each axis is indicated in parenthesis.

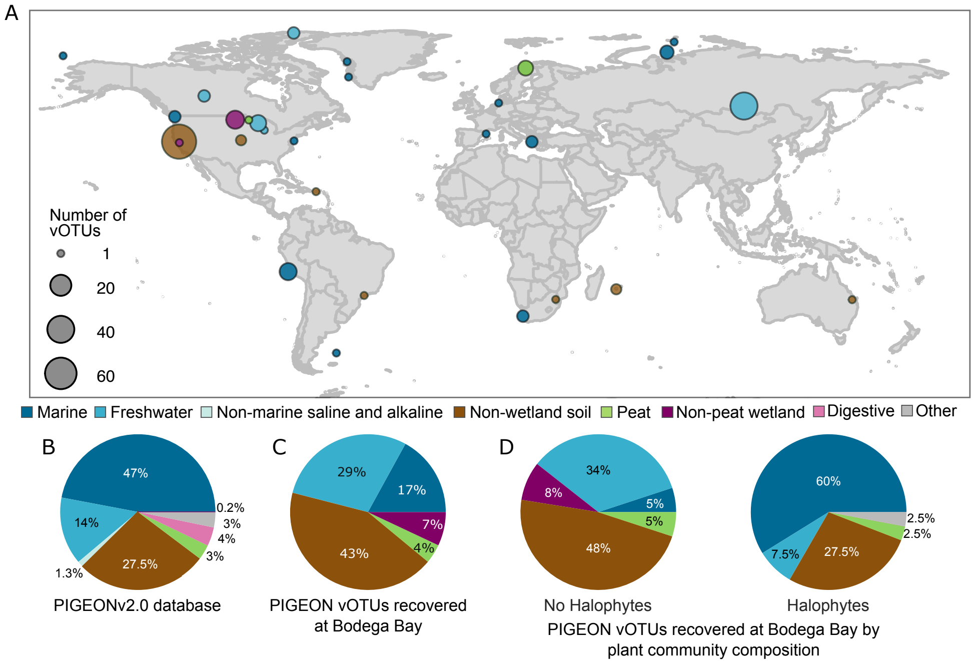


Figure 4.2: Global distribution and habitat context of Bodega Bay vOTUs, leveraging the PIGEONv2.0 database. A) vOTUs (n=196) from PIGEONv2.0 recovered at Bodega Bay by read mapping, according to the location where they were first recovered, colored by the environment in which they were originally recovered. Circle size indicates the number of vOTUs. B) Composition of the PIGEONv2.0 database of 515,763 vOTU sequences, colored by environment. C) Relative proportions of all vOTUs recovered from PIGEONv2.0 at Bodega Bay, colored by the original environment from which they were recovered. D) Relative proportions of vOTUs recovered from PIGEONv2.0 at Bodega Bay, as in panel C, but separated by the Bodega vegetation group in which they were recovered, colored by original source environment. If a vOTU was recovered in both vegetation groups, it appears twice in the chart.

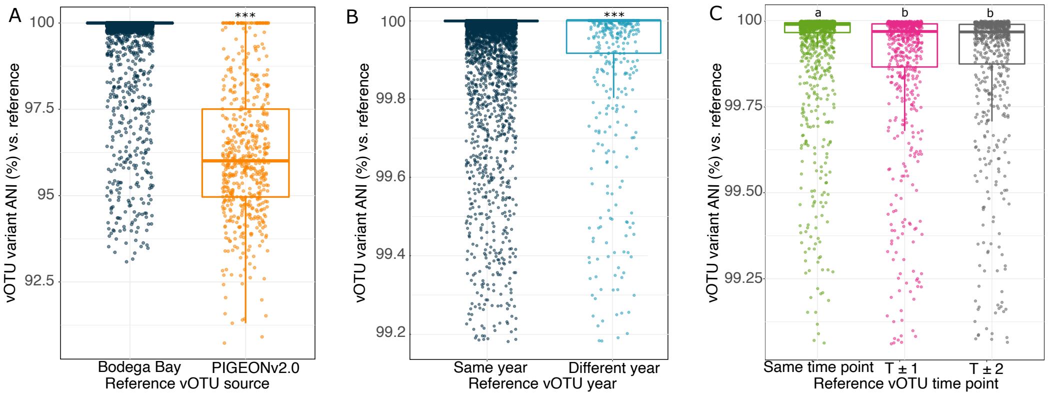


Figure 4.3: Comparisons of viral variant (sub-population) diversity in local and global contexts. Pairwise average nucleotide identities (ANIs) between vOTU variants, calculated between each sample-specific vOTU consensus sequence and the originally assembled (reference) vOTU sequence, using inStrain. Each point is the ANI for one vOTU variant in one Bodega Bay virome compared to the reference sequence for that vOTU. A) Variant ANIs for: (left) vOTUs both assembled and recovered through read mapping from the Bodega Bay dataset (Bodega Bay reference sequences), and (right) vOTUs recovered at Bodega Bay via read mapping but originally derived from PIGEONv2.0 (PIGEONv2.0 reference sequences). Stars above boxes correspond to significant differences between groups (Student's T test, significant when $p < 0.0001$). B) Variant ANIs for vOTUs both assembled and recovered via read mapping from Bodega Bay, either: (left) assembled and recovered in the same year (2019-2019 or 2021-2021), or (right) in different years (2019-2021 or 2021-2019). C) Variant ANIs for vOTUs assembled from Bodega Bay in 2021, either assembled and recovered through read mapping at the same sampling time point, or at different time points, where T±1 equals 2 months between samplings, and T±2 equals 4 months. Letters above boxes correspond to significant differences between groups (Student's T test, significant when $p < 0.0001$). In all three panels, boxes show the median and interquartile range (IQR), and whiskers extend to Q1-1.5*IQR and Q3+1.5*IQR.

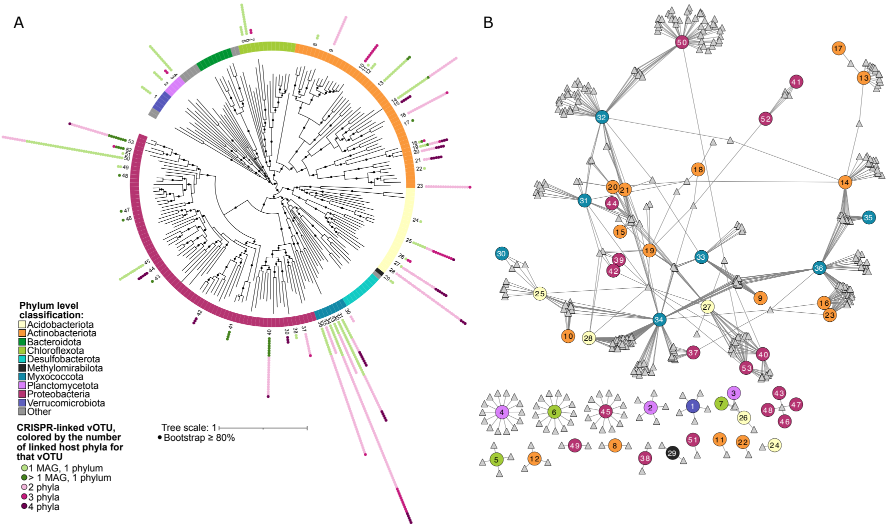


Figure 4.4: Bodega Bay virus-host linkages and putative interactions derived fromCRISPR spacer-protospacer matches. A) Unrooted phylogenetic tree (concatenated predicted protein alignment of 43 marker genes defined by CheckM) of prokaryotic metagenome-assembled genomes (MAGs) with at least one vOTU linked by CRISPR sequence homology. The numbers for MAGs correspond to numbers in the network in panel B. Tree was constructed using gtdbtk under the WAG model. B) Virus-host linkage network for MAGs with at least one vOTU linked through CRISPR homology. Circle nodes represent MAGs and are colored by phylum, while triangles represent vOTUs.

CHAPTER 5

Soil and rhizosphere viral communities are differently structured by plot location, treatment with mycorrhizal fungi, and time during the tomato growing season

Anneliek M. ter Horst[1], Katherine Simpson-Johnson[2], Christian Santos-Medellín[1], Jane D. Fudyma[1], Amélie C. M. Gaudin[2] and Joanne B. Emerson* [1]

[1] Department of Plant Pathology, University of California, Davis, CA, USA

[2] Department of Plant Sciences, University of California, Davis, CA, USA

Abstract

The rhizosphere microbiome plays an important role in plant health, growth, and nutrient acquisition, and by infecting rhizosphere microbes, viruses have the potential to impact these processes. To interrogate viral communities in rhizosphere soils, here we collected tomato rhizosphere and bulk soil samples just prior to and during the 2021 tomato growing season (four time points) in Davis, CA, USA, and we generated 78 viral size-fraction metagenomes (viromes) and associated 16S rRNA gene and ITS amplicon sequencing datasets, as well as metatranscriptomes from 33 rhizospheres from these samples. Half of the plants were treated with arbuscular mycorrhizal fungi (AMF). We recovered 63,924 viral 'species' sequences (vOTUs), and rhizospheres had significantly higher viral richness than bulk soils, counter to the usual trend of higher diversity in bulk soils for other microbiota, although this trend is management dependent. Bulk soil viral communities differed most significantly by soil moisture content, and at high but not low moisture content, bulk soil viral communities were similar to those from the rhizosphere, suggesting viral compositional similarities in areas of high host activity. In both bulk and rhizosphere soils,

viral community composition differed significantly by plot location and over time, but only rhizosphere viromes exhibited significant differences between AMF-treated and untreated samples. Approximately 25% of the vOTUs had been previously recovered in other datasets, predominantly in agricultural systems, suggesting habitat filtering for these vOTUs. RNA viral communities recovered from rhizosphere metatranscriptomes differed significantly over time and plot location, but not by AMF treatment. Prokaryotic and fungal community composition differed most significantly by soil compartment, but when considering bulk and rhizosphere soils separately, they both differed most significantly over time, which was a substantially less important factor in structuring viral communities. These results indicate that viruses are dynamic members of the tomato rhizosphere microbiome that, presumably by way of their hosts, respond to changing environmental conditions, plant growth stages, and soil microbiota.

5.1. Introduction

Plants exude a significant amount of their fixed carbon into the rhizosphere, thereby feeding plant-associated soil microbial communities and influencing their composition and activity [1,2,3]. In return, beneficial soil microbes aid in plant pathogen resistance, nutrient uptake, and synthesis of growth promoting hormones [1,3,4,5,6]. As such, there has been an increasing interest in using microbial amendments in agricultural management to improve crop production [7,8]. One such group of organisms is represented by the arbuscular mycorrhizal fungi (AMF), which can aid the plant in phosphorus uptake [9]. While AMF can benefit plant productivity in natural systems, it is up for debate whether these benefits extend to production-oriented agricultural systems [10]. The effects of AMF inoculation on rhizosphere microbiomes and, particularly, viromes are unknown.

Whereas some previous research has relied on bioinformatically mining viral sequences from total soil metagenomes [11,12], improvements in laboratory methods have recently made it possible to explore soil viral communities in more detail [13,14,15,16,17,18,19]. By purifying the viral fraction through 0.22 μm filtration, it is now possible to obtain a much higher viral diversity within each sample [13,14,16], revealing previously unrecognized viral ecological patterns. For example, recent studies have found that soil viral communities are extremely heterogeneous over meters-scale distances and are significantly different across habitats (e.g. grasslands, peatlands, woodlands and wetlands) at both regional and global scales [16,20,21,22]. However, the same viral species can be observed across the globe in similar habitats (e.g. viruses from wetlands can be found in other wetland ecosystems) [16,23]. Locally, soil viral communities display stronger spatial and temporal distance-decay relationships than prokaryotic communities [**14**, **21**]. Differences between natural and agricultural soils are also beginning to be revealed, but how these biogeographical patterns translate to rhizosphere viral communities is virtually unknown [**2**, **24**, **25**, **26**, **27**].

Recent advances in sequencing techniques have revealed important information about the structure and function of the microbial and fungal communities in rhizosphere microbiomes, providing some clues about the ecological patterns that we might expect in rhizosphere viromes. Many studies have shown that soil type an soil management practices have an effect on rhizosphere prokaryotic and fungal communities [**28**, **29**, **30**], but plant species, temporal (plant developmental stage) and spatial scales are important factors in shaping rhizosphere prokaryotic and fungal communities as well [**28**, **31**, **32**, **33**, **34**]. However, even though different plant species assemble relatively different rhizosphere microbiomes [**35**], these communities can still be relatively similar, even over spatial distance and in different environments [**36**, **37**], because plants select for specific microbes from the bulk soil pool, in order to improve plant fitness, thereby reducing microbial diversity in the rhizosphere [**38**].

It still remains to be seen whether rhizosphere viral communities follow similar patterns to those known for their likely host bacteria and fungi, for example, whether the significant differences between bulk soil and rhizospheres are generalizable across plants and soil conditions.

Although we know very little about rhizosphere viral communities, some early studies of viruses in the rhizosphere offer some hints of expected patterns [**2**, **24**, **25**, **26**, **27**]. In 2009, a transmission electron microscopy study found that, while viruses were present in the rhizosphere, no difference in virion abundances could be found between rhizosphere and bulk soils [**27**]. A more recent study of RNA viral communities mined from metatranscriptomes revealed significant differences among bulk, rhizosphere, and detritosphere compartments of wild oat [**39**]. A comparison of four bulk and four rhizosphere soils showed significantly different DNA viral community composition between bulk and rhizosphere soils in a maize cropping system [**24**]. Here, we generated 78 viral size-fraction metagenomes (viromes) to characterize viral community composition in 36 tomato plant rhizospheres and their associated bulk soils (as well as bulk soils before planting) throughout one growing season in an agricultural field in Davis, CA, USA. We also generated 33 rhizosphere metatranscriptomes to investigate the RNA viral community and 78 16S rRNA gene and ITS amplicon sequencing datasets to investigate the microbial and fungal communities in bulk and rhizosphere soils. Half of the tomato plants were treated with arbuscular mycorrhizal fungi (AMF) as part of an ongoing study to test the impacts of AMF on tomato yield. We also compared the recovered vOTUs to our PIGEONv2.0 reference database of 466,057 vOTUs [**23**] from diverse ecosystems, in order to investigate global soil viral biogeographical patterns. We explored viruses as dynamic members of the tomato rhizosphere microbiome and investigated what factors, such as plant growth stage, plot location, AMF treatment condition, and composition of other microbiota influenced viral community composition.

5.2. Results and discussion

5.2.1. Dataset overview.

To investigate tomato rhizosphere and bulk soil viral community composition and its underlying drivers over five months, we sampled three conventionally managed plots, which received biocides, (64 × 64 m) within the UC Davis Russell Ranch Century Experiment [**40**, **41**], where tomato plants were either untreated controls or inoculated with EndoMaxx Prime, an arbuscular mycorrhizal fungi (AMF)-containing formula from Valent (Valent, CA, USA). Near-surface (top 15 cm) bulk soils were collected at four time points [March (pre-planting), June (vegetative state), July (flowering), and August (harvest) of 2021], and rhizosphere (root-adherent) soils were collected at the three time points with plants. For both bulk and rhizosphere soils, two replicate samples were collected per time point per treatment per plot, and bulk soils accompanying rhizosphere soils were collected approximately 30 cm from the plant base. All 78 samples [2 compartments (bulk/rhizosphere) x 2 replicates x 2 treatments x 3 plots x 3 time points + 6 bulk soils from March], went through total DNA extraction for 16S rRNA gene and ITS amplicon sequencing and viral-size-fraction metagenome (virome) generation for viral community analyses, and total RNA was extracted from all 36 rhizosphere soil samples (RNA extraction of three samples failed, for a total of 33 metatranscriptomes analyzed).

A total of 116,884 viral contigs was recovered via de novo assembly of the viromes, and a further 17,332 were recovered through read mapping to our PIGEONv2.0 database [**23**], of which 10,173 had been recovered at the same field site (Russell Ranch Century Experiment) in 2018 [**15**]. Together, these viral contigs clustered into 67,038 viral operational taxonomic units (vOTUs, $\geq$ 10 kbp, $\geq$ 95% average nucleotide identity, approximately species-level taxonomy [**42**]). Of these vOTUs, 2% could be taxonomically classified via vConTACT2 [**43**] clustering with viral genomes from RefSeq (version 85 [**44**]). A total of 16,146 16S rRNA sequences and 6,684 ITS1 sequences were obtained, and we recovered 380 viral RNA-dependent RNA polymerase (RdRp) genes from the metatranscriptomic data. Relative abundances of these sets of vOTUs, RdRp containing sequences, and OTUs, were derived from read mapping (vOTUs and RdRps) or clustering-based counts (OTUs) and used for downstream community compositional analyses.

5.2.2. Viral community composition was structured by a combination of soil compartment and moisture content, but within each compartment, different relative influences of plot location, AMF treatment, and time were revealed.

Considering the full 78-sample viromic dataset, most vOTUs were recovered in more than one virome (87%). This is counter to the typical recovery of only 38-81% of vOTUs in single viromes in natural soil systems [**20**, **23**] and similar to a prior study of viromes from the same field site in 2018 [**45**], perhaps indicating greater viral community homogeneity in agricultural relative to natural soil systems. We hypothesize that these differences could be due to tilling (mixing) of the soil and relatively uniform fertilizer and irrigation inputs, as well as crop homogeneity in agricultural systems, compared to less mixing and greater biotic and abiotic heterogeneity in natural soils.

To interrogate ecological patterns in our dataset, we first sought to explore what variables structured the viral communities in both bulk and rhizosphere soils. Although viral community composition differed significantly between bulk and rhizosphere soils (PERMANOVA $p < 0.00001$), rhizosphere soils from all three sampled time points had similar communities to bulk soils from March and June, whereas bulk soils from July and August had significantly different viral communities from all of the other samples (Figure 5.1A). At one of the time points for which bulk soil viromes appeared similar to those from rhizospheres (in March), tomatoes were not yet planted, eliminating similarity in paired bulk-rhizosphere samples as a predominant driver of this pattern and suggesting the potential for other habitat and/or microbial host community similarities between some bulk soil and all rhizosphere samples. Over the course of the growing season, the soil dried down from an average of 12% soil moisture in March and June to 6% on average in July and August (Figure 5.1B), with the higher moisture content in the bulk soil samples with viromes most similar to rhizospheres. Soil moisture has already been shown to be an important factor in shaping soil viral communities [**46**, **47**], and it is an important driver of microbial community activity [**48**, **49**, **50**]. As rhizosphere soils are also zones of high microbial activity [**31**, **51**, **52**] similar viral communities in different soil compartments were associated with more active microbial communities, suggesting that microbial activity could be a more important driver of viral community composition than soil compartment alone. Consistent with greater viral production and diversity under conditions known to increase the activity of other microbes, here vOTU richness was highest in rhizosphere soils, followed closely by wetter bulk soils from March and June, and was significantly lower in drier bulk soils from July and August (Figure 5.1C).

We next sought to explore the relative importance of soil moisture and other physicochemical properties (measured for bulk soils only), treatment with AMF, plot location, and time on viral community composition within the two soil compartments (bulk and rhizosphere soils). Bulk soils were primarily structured by soil moisture content (Figure 5.2A, $p<0.00001$) and secondarily by plot location in the field ($p=0.001$) (Figure 5.2B). Bulk soil viral communities were not significantly structured by AMF treatment condition or by time point (apart from the aforementioned temporal differences in moisture content). Rhizosphere viral communities separately most significantly by AMF treatment condition (Figure 5.2C) and secondarily but still significantly by plot location (Figure 5.2D).

As might be expected, viral community composition correlated significantly with host prokaryotic community composition (Mantel test, $p < 0.00001$), suggesting environmental filtering by way of hosts. However, prokaryotic community composition separated most significantly by soil compartment (Figure 5.3A), and, within each compartment, by time for both bulk soil (Supplementary figure 5.1A) and rhizosphere communities (Figure 5.3B). Rhizosphere prokaryotic communities were also significantly different in AMF treated versus untreated plants per plot (Figure 5.3C), but treatment with AMF alone did not have a significant effect (PERMANOVA, $p=0.03$). Like prokaryotic community structure, fungal community structure correlated with compartment, then by time and then, within the rhizosphere, by plot location, followed by treatment with AMF per plot. Here, treatment with AMF alone also did not have a significant effect on fungal community composition (PERMANOVA, $p=0.025$).

Spatial distance appeared to influence the viral community more significantly than the prokaryotic community, whereas prokaryotic communities were more affected by temporal shifts. Both of these patterns have been observed before [**14**, **21**]. As viruses rely on hosts for replication, explanations for these differences are unknown, but, as suggested previously, there could be differences in the scales of measurement, with viruses representing more active members of the microbiome, or, dispersal limitation may play a bigger role in viral community assembly compared to prokaryotic community assembly.

5.2.3. Approximately 25% of the vOTUs had been previously recovered in other datasets, with global distribution patterns suggestive of habitat filtering for agricultural vOTUs.

To investigate the global habitat distribution of vOTUs recovered at Russell Ranch, we leveraged our Phages and Integrated Genomes Encapsidated Or Not (PIGEONv2.0) database [**23**] (Figure 5.4A). Of the 67,038 vOTUs recovered at Russell Ranch in this study, 17,332 (25.8%) were previously detected, including

10,173 (15.1%) previously recovered at the same field site in 2018 [**15**], suggesting that part of the agricultural soil virome is stable and/or recurring. Interestingly, 63.3% of those vOTUs were from two deeply sequenced viromes with DNA that had been density gradient fractionated to capture different GC% in different 'mini-metagenome' viromes to facilitate better assembly, resulting in substantially greater vOTU recovery in those samples compared to our typical viromes. These results suggest both that deeper and more targeted sequencing can facilitate greater access to the rare virosphere and that the rare virosphere contains a substantial seed bank, perhaps similar to previous intensive sequencing of soil microbes, in which the diversity in Central Park, USA reflected most of the known global diversity [**53**] and deep sequencing of marine microbes, in which one deeply sequenced sample captured most of the diversity in the English Channel over a longer, more shallowly sequenced time series [**54**]. Perhaps everything is (or can get) mostly everywhere [**55**].

For the global habitat analysis, we leveraged PIGEONv2.0 but excluded all vOTUs previously recovered from Russell Ranch (n=21,839). The remaining 7,159 (10.7%) previously recovered vOTUs were previously detected at other locations, primarily (98.7%) at other agricultural sites in California (Figure 5.4B). This result almost certainly reflects geographic sampling bias, as most (90.1%) of the soil vOTUs in PIGEONv2.0 are from our group's locally sampled viromes in California. For example, of the previously detected vOTUs in this study that were not from earlier samplings of the same field site, 80% were found in rhizosphere or bulk soils from almond orchards in California [**56**], which, to our knowledge, is the only other large-scale investigation of rhizosphere viromes. More generally, vOTUs from our dataset were previously recovered in other agricultural bulk soils (3,086), other agricultural rhizosphere soils (3,978), and natural soils (67) (Figure 5.4C). The skew towards vOTUs from other agricultural systems, despite substantial representation of natural soil vOTUs in the PIGEONv2.0 database (5.9% of PIGEONv2.0 vOTUs) suggests that similar viruses may be adapted to similar ecosystems and/or host communities, as has been suggested previously for peat and other wetland viruses [**16**, **23**]. However, in contrast to wetland ecosystems, from which 0.7% of the vOTUs were previously recovered from marine or freshwater environments, no vOTUs from marine ecosystems were recovered here, and only one vOTU was recovered from freshwater, suggesting that there are strong habitat boundaries for these agricultural soil viruses.

To spatially compare the vOTUs recovered at Russell Ranch and elsewhere in California, we used our almond dataset, in an attempt to minimize other confounding factors, such as the crop grown. We only compared vOTUs recovered at Russell Ranch and at almond orchards throughout California [**56**]. A total of 75,375 vOTUs was used for this analysis, of which 37.9% came from Davis, 12.8% came from Woodland, 14.6% came from Escalon, and 34.7% came from Madera (all cities in California). Most vOTUs recovered at Russell Ranch and at an almond orchard came from an almond orchard in Davis (n=4,820, 83.9%), followed by vOTUs from Woodland (n=523, 9.1%), then Madera (n=289, n=5%, then Escalon (n=115, 2%), suggesting that space may also play a role in the global distribution of soil viruses, but that the size of the reference database also may influence the number of vOTUs recovered. Together, these results show that viruses can be conserved over large spatial distances, and that habitat characteristics play an important role in shaping the global soil virome.

5.2.4. Rhizosphere RNA viral communities were structured by plot location and time more than by AMF treatment.

To investigate the RNA viral community of the rhizosphere microbiome, we leveraged total soil metatranscriptomes to recover RdRp genes [**39**, **57**, **58**]. We recovered 380 RdRp sequences, including 174 putative viruses of fungi (mycoviruses), 166 putative viruses of prokaryotes, 20 putative viruses of animals (insects and vertebrates), 15 putative viruses of animals and/or plants, and 3 putative viruses of plants. Significant differences in RNA viral community composition were observed by plot location (PERMANOVA p = 0.001) and by time point (PERMANOVA p=0.001), but these were not immediately visible in the first three principal coordinates axes, so we opted instead to show them via a canonical analyses of principal coordinates (CAP, Figure 5.5 A, B). AMF treatment alone did not have a significant effect on RNA viral community composition (PERMANOVA p=0.7) , but location in the field and treatment together significantly impacted the RNA viral community composition (p=0.007).

Counter to our previous study of RNA viral communities in oak leaves, in which a greater proportion of putative mycoviral sequences was associated with senescing leaves (presumably dominated by saprobic fungi feeding on the decaying material, compared to healthy plant leaves with presumably less fungal activity), here we found no significant enrichment in putative mycoviral sequences in samples treated with AMF. There could be a number of potential reasons for this result, including unsuccessful AMF establishment (AMF growth was confirmed in our companion study, excluding this explanation, data not shown) and/or a lack of viruses in the soil or AMF inoculum that were capable of successfully infecting the ‘invading’ AMF. It is also possible that our laboratory methods failed to sufficiently recover or our bioinformatic methods failed to recognize these viruses. Still, given the observed shifts in RNA viral community composition, these results indicate that RNA viruses are active members of the tomato rhizosphere microbiome that turnover during the growing season.

5.3. Conclusions

Here, we investigated viral community assembly patterns and their underlying drivers in the rhizosphere microbiome of tomato plants over one growing season in Davis, CA. We analyzed dsDNA and RNA viral communities from tomato rhizosphere soils and their accompanying bulk soils and showed significant differences in viral community composition between bulk and rhizosphere soils, where bulk soil viral communities with higher soil moisture content were more similar to rhizosphere viral communities than to bulk soil communities with lower soil moisture. Rhizosphere viral communities were primarily structured by treatment with AMF and secondarily by plot location, whereas viral communities in bulk soils were primarily structured by soil moisture and secondarily by location in the field (though we note that soil moisture was not measured in the rhizosphere due to sample size limitations). Prokaryotic and fungal community composition were primarily structured by time instead of plot location for both bulk and rhizosphere soils, perhaps indicating dispersal limitations for viruses in soil compared to prokaryotes and fungi. 25% of the vOTUs were previously detected, 15.1% at the Russell Ranch field site, suggesting that part of the soil virome is stable or recurring, and 10.6% of the vOTUs were recovered at other locations, primarily other agricultural sites, suggesting habitat filtering for these viruses. Together, these results indicate that tomato rhizosphere viral communities are a dynamic part of the rhizosphere microbiome, that respond to changes in the environment, such as soil compartment, soil moisture level, and other members of the soil microbiota such as AMF, that form different communities on a relatively small spatial scale.

5.4. Materials and methods

5.4.1. Sample collection and processing.

Samples were collected at the Russell Ranch Sustainable Agriculture Facility (Davis, California, United States, 38.54'N, 121.87'W) at four time points during one tomato growing season in 2021: March (pre-planting), June (6 weeks post-planting), July (10 weeks post-planting), and August (pre-harvest, 16 weeks post-planting)). The Heinz 1662 tomato cultivar was used for all experiments, and three plots were sampled at each time point. All three plots were conventionally managed. Briefly, each plot received 156 kg per hectare of mineral fertilizer (urea ammonium nitrate solution), via drip line fertilization 3-4 times throughout the growing season and were left fallow during the winter months. At each time point, two plants and their accompanying bulk soil (30 cm from the base of the plant) were randomly selected for processing. This is with exception of the March time point, from which only bulk soils were sampled (in approximately the same location, i.e., 30 cm from where plants were to be planted), since this was prior to planting. Seedlings with the first two true leaves were transplanted into the field, and immediately prior to transplanting dip inoculated. Plants that underwent treatment with EndoMaxx Prime were inoculated using a root-dip method, following manufacturer instructions, and control plants were inoculated with autoclaved EndoMaxx Prime (Valent, San Ramon, CA, USA).

5.4.2. Virome DNA extraction, library construction, and shotgun sequencing.

Soils were processed immediately after collection for viromics. To enrich the samples for soil virions, a modified version of a previously published protocol was used [**59**]. Per sample, 10 grams of soil were suspended in 30 mL of protein-supplemented phosphate-buffered saline solution (PPBS: 2% bovine serum albumin, 10% phosphate-buffered saline, 1% potassium citrate, and 150 mM $MgSO_4$ and then briefly vortexed and placed on an orbital shaker (30 min, 400 rpm, 4 °C). For rhizosphere samples, the roots were vigorously shaken to remove loose soil particles, and only the adhered portion was analyzed. The roots were suspended in the above-mentioned buffer and shaken as defined above, and we made no difference between the rhizosphere and endosphere. The shaken soil buffer solution was then centrifuged (10 min, 3,095 x g, 4°C) and the resulting supernatant was then centrifuged two times (8 min, 10,000 x g, 4°C) to remove residual soil particles. The centrifuged supernatants were then filtered through a 0.22 μm polyethersulfone membrane to remove cells. The filtrate was then ultracentrifuged (2 hrs 25 min, 32,000 x g, 4 °C) to pellet the virions, using a Optima LE-80K 293 ultracentrifuge with a 50.2 Ti rotor (Beckman-Coulter Life Sciences). Supernatant was discarded, and pellets were resuspended in 100 μl of ultrapure water and treated with DNase to remove free DNA not encapsidated in a virion, using 10 U of RQ1 RNase-free DNase and 10 μl of 10× DNase buffer (Promega Corp., Madison, WI, USA). Samples were incubated at 37 °C for 30 min, and the reaction was stopped by adding 10 μl of the DNase stop solution (Promega Corp., Madison, WI, USA) and incubating the samples at 65 °C for 10 min. DNA was then extracted from the viral fraction, using the DNeasy PowerSoil Pro kit (Qiagen, Hilden, Germany), following the manufacturer's instructions, with an added step of a 10-minute incubation at 65 °C before the bead-beating step. Libraries were constructed using the DNA Hyper Prep library kit (Kapa Biosystems-Roche, Basel, Switzerland). Paired-end 150 bp sequencing was done using the NovaSeq S4 platform (Illumina), to an approximate depth of 10 Gbp per virome.

5.4.3. Total DNA extraction, amplicon library construction, and sequencing.

Total DNA was extracted from 0.25 g of soil with the DNeasy PowerSoil Pro kit (Qiagen, Hilden, Germany), following the manufacturer's instructions, with an added step of a 10-minute incubation at 65 °C before the bead-beating step. For rhizosphere samples 0.25 g of soil was brushed off the roots into the extraction tube. Construction of the amplicon libraries followed a previously described dual-indexing strategy [**60**, **61**]. To target the V4 region of the 16S rRNA gene, universal primers 515F and 806R were used, using the following PCR protocol: an initial denaturation step at 98 °C for 2 min, followed by 30 cycles of 98 °C for 20 s, 50°C for 30 s and 72 °C for 45 s, and a final extension step at 72 °C for 10 min. To amplify the ITS1 region, we used the universal primers ITS1-F and ITS2 [**62**, **63**, **64**] and the following PCR program: an initial denaturation step at 95°C for 2 min, followed by 35 cycles of 95°C for 20 s, 50°C for 30 s, and 72°C for 50 s, followed by a final extension at 72°C for 10 min. All PCR reactions were performed using the Platinum Hot Start PCR Master Mix (Invitrogen). Libraries were cleaned using AmpureXP magnetic beads (Beckman Coulter), quantified (Qubit 4 fluorometer), and pooled in equimolar concentrations. Paired-end sequencing (250 bp) was performed on the MiSeq platform (Illumina), using a standard flow cell per library (one for 16S and one for ITS.

5.4.4. RNA extraction, library construction, and sequencing.

Rhizosphere samples for RNA extraction were immediately put into liquid nitrogen in the field and stored at -80 °C until further processing. Total RNA was extracted using the RNeasy PowerSoil Pro Kit (Qiagen, Hilden, Germany), following the manufacturer's instructions. RNA was submitted to Genewiz (San Francisco, CA, USA) for ribodepletion, cDNA preparation, and library construction via the NEBNext Ultra II RNA library prep kit (New England Biolabs, Ipswitch, MA, USA). Paired-end sequencing (150 bp) was done using the NovaSeq 6000 platform (Illumina) to an approximate sequencing depth of 10 Gbp per sample.

5.4.5. Virome bioinformatic processing.

Sequencing reads were trimmed using Trimmomatic v0.39 [**65**], removing Illumina adapters and quality trimming reads, using paired-end trimming, a sliding window size of 3:40, and a minimum read length of 50 bp. PhiX sequences were removed using bbduk from from the BBMap v38-72 package [**66**], using k=31 and hdist=1, and host plant reads were removed using Bowtie2 v2.4.2, using the sensitive setting, mapping against the genome of tomato (S. lycopersicum), GenBank accession number GCA_012431665. The remaining reads were assembled into contigs, using MEGAHIT 1.0.6 [**67**], with settings k-min 27, minimum contig length of 10 kb, and presets meta-large. Resulting contigs were renamed using the rename function from BBMap, and viral contigs were predicted using VIBRANT v1.2.0 [**68**], in virome mode. All predicted viral contigs were used for subsequent analysis. Viral contigs were dereplicated into vOTUs, using dRep v3.2.0 [**69**], at 95% ANI with a minimum coverage threshold of 85%, using the ANImf algorithm. Reads were mapped to the vOTUs and to the PIGEONv2.0 database using Bowtie2 v2.4.2 [**70**], using sensitive mode, and vOTUs from the PIGEONv2.0 database that were recovered in this dataset were clustered with the vOTUs assembled in this dataset using dRep, using the same settings as above. Reads were then again mapped to this non-redundant dataset of vOTUs, using Bowtie2, and the resulting samfiles were converted to bamfiles using SAMtools v1.15.1 [**71**]. A coverage table was created using CoverM v0.6.1 [**72**], using CoverM contig with the mean coverage and a minimum covered fraction of 75%. The resulting coverage table was used for statistical analysis, unless otherwise noted. All scripts for bioinformatic processing are available on Github (https://github.com/AnneliektH/TomatoRhizo).

5.4.6. Metatranscriptome bioinformatic processing.

Sequencing reads were trimmed and PhiX and host plant reads were removed as described above. The remaining reads were assembled into contigs, using MEGAHIT v1.0.6 [**67**], with settings k-min 27, minimum contig length of 200 bp and presets meta-large. The resulting sequences were translated into proteins using prodigal v2.6.3 [**73**], using standard settings. HMMER v3.3.2 [**74**] was used to recover RNA-dependent RNA polymerase (RdRp) sequences, as described previously [**39**, **58**], using a p-value of 0.00001. DIAMOND v0.9.22.123 [**75**] was used for a protein-protein blast of the assembled proteins against the NCBI nr prokaryotic database (v203), using a p-value of 1e-6. Contigs with a prokaryotic gene were used for read mapping, and reads that mapped to these contigs were removed from the read pool, in order to reduce prokaryotic RNA presence. The remaining reads were re-assembled using MEGAHIT, and this process was iterated three times. The final set of contigs that had a predicted RdRp sequence was clustered using dRep v3.2.0 [**69**] at 95% ANI with a minimum coverage threshold of 85%, using the ANImf algorithm. Metatranscriptomic reads were mapped back to these sequences and to RefSeq v203 RNA viral sequences (n=4,472) using Bowtie2 as described above, and a coverage table was created for downstream analyses using CoverM.

5.4.7. Amplicon sequence bioinformatic processing.

Paired-end reads assembly into single sequences was done using PANDAseq v2.9 [**76**], chimeric sequence removal, dereplication, error rate inference, denoising and read merging was done using DADA2 v1.12.1 [**77**]. Taxonomy was assigned using the RDP classifier implementation in DADA2 [**78**] via the SILVA database v132 [**79**] for 16S rRNA gene sequences, and via the UNITE database v2021-05-10 for ITS sequences [**80**]. OTU tables with counts of each OTU in each sample were generated using the makeSequenceTable function in DADA2.

5.4.8. Statistical analysis.

All statistical analyses were performed using R v 4.1.0 [**81**], using the mean coverage vOTU abundance table or the other coverage tables prepared as described above (for RdRps, 16S rRNA OTUs, and ITS OTUs, respectively), unless otherwise noted. Bray-Curtis dissimilarities were calculated on log-transformed relative abundances, using the vegdist function from Vegan v2.6-2 [**82**], and principal coordinates analyses were performed with the pcoa() function from ape v5.4-2 [**83**]. All maps were made using the R package ggmaps [**84**] and all other plots were created using the R package ggplot2 v3.3.5 [**85**]. All scripts are available at https://github.com/AnneliektH/TomatoRhizo.

5.5. Data availability

Raw sequencing reads for viromes, 16S and ITS are available on NCBI under bioproject PRJNA937255 and vOTU sequences are available on Dryad within the PIGEONv2.0 database (https://doi.org/10.25338/B8C934)

5.6. Acknowledgements

This study was primarily supported by USDA NIFA AFRI award # 2021-144 67013-34815-0 (grant to JBE). Additional support was provided by USDA NIFA Hatch # CA-D145 PPA-2464-H. We thank Simon Roux and Rex Malmstrom (DOE LBNL JGI) for their exploratory density fractionation and sequencing of our viromic DNA from two Russell Ranch samples from 2018, which became part of our PIGEONv2.0 database and which contributed to vOTU recovery via read mapping in this study.

Bibliography

[1] Joseph Edwards, Cameron Johnson, Christian Santos-Medellín, Eugene Lurie, Natraj Kumar Podishetty, Srijak Bhatnagar, Jonathan A Eisen, and Venkatesan Sundaresan. Structure, variation, and assembly of the root-associated microbiomes of rice. *Proc. Natl. Acad. Sci. U. S. A.*, 112(8):E911–20, February 2015.

[2] Rodrigo Mendes, Paolina Garbeva, and Jos M Raaijmakers. The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. *FEMS Microbiol. Rev.*, 37(5):634–663, September 2013.

[3] Corné M J Pieterse, Ronnie de Jonge, and Roeland L Berendsen. The Soil-Borne supremacy. *Trends Plant Sci.*, 21(3):171–173, March 2016.

[4] Peter A H M Bakker, Corné M J Pieterse, Ronnie de Jonge, and Roeland L Berendsen. The Soil-Borne legacy. *Cell*, 172(6):1178–1180, March 2018.

[5] Davide Bulgarelli, Klaus Schlaeppi, Stijn Spaepen, Emiel Ver Loren van Themaat, and Paul Schulze-Lefert. Structure and functions of the bacterial microbiota of plants. *Annu. Rev. Plant Biol.*, 64:807–838, January 2013.

[6] Rodrigo Mendes, Marco Kruijt, Irene de Bruijn, Ester Dekkers, Menno van der Voort, Johannes H M Schneider, Yvette M Piceno, Todd Z DeSantis, Gary L Andersen, Peter A H M Bakker, and Jos M Raaijmakers. Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science*, 332(6033):1097–1100, May 2011.

[7] Jacqueline M Chaparro, Amy M Sheflin, Daniel K Manter, and Jorge M Vivanco. Manipulating the soil microbiome to increase soil health and plant fertility. *Biol. Fertil. Soils*, 48(5):489–499, July 2012.

[8] Andrea Nuzzo, Aditi Satpute, Ute Albrecht, and Sarah L Strauss. Impact of soil microbial amendments on tomato rhizosphere microbiome and plant growth in field soil. *Microb. Ecol.*, 80(2):398–409, August 2020.

[9] P Gosling, A Hodge, G Goodlass, and G D Bending. Arbuscular mycorrhizal fungi and organic farming. *Agric. Ecosyst. Environ.*, 113(1):17–35, April 2006.

[10] Megan H Ryan and James H Graham. Is there a role for arbuscular mycorrhizal fungi in production agriculture? *Plant Soil*, 244(1):263–271, July 2002.

[11] David Paez-Espino, Emiley A Eloe-Fadrosh, Georgios A Pavlopoulos, Alex D Thomas, Marcel Huntemann, Natalia Mikhailova, Edward Rubin, Natalia N Ivanova, and Nikos C Kyrpides. Uncovering earth's virome. *Nature*, 536(7617):425–430, August 2016.

[12] Gareth Trubl, Ho Bin Jang, Simon Roux, Joanne B Emerson, Natalie Solonenko, Dean R Vik, Lindsey Solden, Jared Ellenbogen, Alexander T Runyon, Benjamin Bolduc, Ben J Woodcroft, Scott R Saleska, Gene W Tyson, Kelly C Wrighton, Matthew B Sullivan, and Virginia I Rich. Soil viruses are underexplored players in ecosystem carbon processing, 2018.

[13] Joanne B Emerson, Simon Roux, Jennifer R Brum, Benjamin Bolduc, Ben J Woodcroft, Ho Bin Jang, Caitlin M Singleton, Lindsey M Solden, Adrian E Naas, Joel A Boyd, Suzanne B Hodgkins, Rachel M Wilson, Gareth Trubl, Changsheng Li, Steve Frolking, Phillip B Pope, Kelly C Wrighton, Patrick M Crill, Jeffrey P Chanton, Scott R Saleska, Gene W Tyson, Virginia I Rich, and Matthew B Sullivan. Host-linked soil viral ecology along a permafrost thaw gradient. *Nature Microbiology*, 3(8):870–880, July 2018.

[14] Christian Santos-Medellin, Laura A Zinke, Anneliek M Ter Horst, Danielle L Gelardi, Sanjai J Parikh, and Joanne B Emerson. Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *ISME J.*, 15(7):1956–1970, July 2021.

[15] Jackson W Sorensen, Laura A Zinke, Anneliek M ter Horst, Christian Santos-Medellin, Alena Schroeder, and Joanne B Emerson. DNase treatment improves viral enrichment in agricultural soil viromes. June 2021.

[16] Anneliek M Ter Horst, Christian Santos-Medellín, Jackson W Sorensen, Laura A Zinke, Rachel M Wilson, Eric R Johnston, Gareth Trubl, Jennifer Pett-Ridge, Steven J Blazewicz, Paul J Hanson, Jeffrey P Chanton, Christopher W Schadt, Joel E Kostka, and Joanne B Emerson. Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. *Microbiome*, 9(1):233, November 2021.

[17] Gareth Trubl, Natalie Solonenko, Lauren Chittick, Sergei A Solonenko, Virginia I Rich, and Matthew B Sullivan. Optimization of viral resuspension methods for carbon-rich soils along a permafrost thaw gradient. *PeerJ*, 4:e1999, May 2016.

[18] Gareth Trubl, Simon Roux, Natalie Solonenko, Yueh-Fen Li, Benjamin Bolduc, Josué Rodríguez-Ramos, Emiley A Eloe-Fadrosh, Virginia I Rich, and Matthew B Sullivan. Towards optimized viral metagenomes for double-stranded and single-stranded DNA viruses from challenging soils. *PeerJ*, 7:e7265, July 2019.

[19] Kurt E Williamson, Mark Radosevich, and K Eric Wommack. Abundance and diversity of viruses in six delaware soils. *Appl. Environ. Microbiol.*, 71(6):3119–3125, June 2005.

[20] Devyn M Durham, Ella T Sieradzki, Anneliek M ter Horst, Christian Santos-Medellín, C Winston A Bess, Sara E Geonczy, and Joanne B Emerson. Substantial differences in soil viral community composition within and among four northern california habitats. *ISME Communications*, 2(1):1–5, October 2022.

[21] Christian Santos-Medellín, Katerina Estera-Molina, Mengting Yuan, Jennifer Pett-Ridge, Mary K Firestone, and Joanne B Emerson. Spatial turnover of soil viral populations and genotypes overlain by cohesive responses to moisture in grasslands. *Proc. Natl. Acad. Sci. U. S. A.*, 119(45):e2209132119, November 2022.

[22] Ruonan Wu, Michelle R Davison, William C Nelson, Emily B Graham, Sarah J Fansler, Yuliya Farris, Sheryl L Bell, Iobani Godinez, Jason E Mcdermott, Kirsten S Hofmockel, and Janet K Jansson. DNA viral diversity, abundance, and functional potential vary across grassland soils with a range of historical moisture regimes. *MBio*, 12(6):e0259521, December 2021.

[23] Anneliek M ter Horst, Jane D Fudyma, Jacqueline L Sones, and Joanne B Emerson. Dispersal, habitat filtering, and eco-evolutionary dynamics as drivers of local and global wetland viral biogeography. April 2023.

[24] Li Bi, Dan-ting Yu, Shuai Du, Li-mei Zhang, Li-yu Zhang, Chuan-fa Wu, Chao Xiong, Li-li Han, and Ji-zheng He. Diversity and potential biogeochemical impacts of viruses in bulk and rhizosphere soils, 2021.

[25] Joanne B Emerson. Soil viruses: A new hope. *mSystems*, 4(3), May 2019.

[26] Akbar Adjie Pratama and Jan Dirk van Elsas. The 'neglected' soil virome – potential role and impact, 2018.

[27] M M Swanson, G Fraser, T J Daniell, L Torrance, P J Gregory, and M Taliansky. Viruses in soils: morphological diversity and abundance in the rhizosphere. *Ann. Appl. Biol.*, 155(1):51–60, August 2009.

[28] Laurent Philippot, Jos M Raaijmakers, Philippe Lemanceau, and Wim H van der Putten. Going back to the roots: the microbial ecology of the rhizosphere. *Nat. Rev. Microbiol.*, 11(11):789–799, November 2013.

[29] Jennifer E. Schmidt, Rachel L. Vannette, Alexandria Igwe, Rob Blundell, Clare L. Casteel, and Amélie C. M. Gaudin. Effects of agricultural management on rhizosphere microbial structure and function in processing tomato plants. *Applied and Environmental Microbiology*, 85(16):e01064–19, 2019.

[30] Jennifer E. Schmidt, Angela D. Kent, Vanessa L. Brisson, and Amélie C.M. Gaudin. Agricultural management and plant selection interactively affect rhizosphere microbial community structure and nitrogen cycling. *Microbiome*, 7(1):146, 2019.

[31] Peter A H M Bakker, Roeland L Berendsen, Rogier F Doornbos, Paul C A Wintermans, and Corné M J Pieterse. The rhizosphere revisited: root microbiomics. *Front. Plant Sci.*, 4:165, May 2013.

[32] Davide Bulgarelli, Matthias Rott, Klaus Schlaeppi, Emiel Ver Loren van Themaat, Nahal Ahmadinejad, Federica Assenza, Philipp Rauf, Bruno Huettel, Richard Reinhardt, Elmon Schmelzer, Joerg Peplies, Frank Oliver Gloeckner, Rudolf Amann, Thilo Eickhorst, and Paul Schulze-Lefert. Revealing structure and assembly cues for arabidopsis root-inhabiting bacterial microbiota. *Nature*, 488(7409):91–95, August 2012.

[33] Derek S Lundberg, Sarah L Lebeis, Sur Herrera Paredes, Scott Yourstone, Jase Gehring, Stephanie Malfatti, Julien Tremblay, Anna Engelbrektson, Victor Kunin, Tijana Glavina Del Rio, Robert C Edgar, Thilo Eickhorst, Ruth E Ley, Philip Hugenholtz, Susannah Green Tringe, and Jeffery L Dangl. Defining the core arabidopsis thaliana root microbiome. *Nature*, 488(7409):86–90, August 2012.

[34] Thomas R Turner, Euan K James, and Philip S Poole. The plant microbiome. *Genome Biol.*, 14(6):209, June 2013.

[35] Andrew Matthews, Sarah Pierce, Helen Hipperson, and Ben Raymond. Rhizobacterial community assembly patterns vary between crop species. *Front. Microbiol.*, 10:581, April 2019.

[36] Mia M Howard, Christian A Muñoz, Jenny Kao-Kniffin, and André Kessler. Soil microbiomes from fallow fields have Species-Specific effects on crop growth and pest resistance. *Front. Plant Sci.*, 11:1171, August 2020.

[37] Pankaj Trivedi, Jan E Leach, Susannah G Tringe, Tongmin Sa, and Brajesh K Singh. Plant–microbiome interactions: from community assembly to plant health. *Nat. Rev. Microbiol.*, 18(11):607–621, August 2020.

[38] Ning Ling, Tingting Wang, and Yakov Kuzyakov. Rhizosphere bacteriome structure and functions. *Nat. Commun.*, 13(1):836, February 2022.

[39] Evan P Starr, Erin E Nuccio, Jennifer Pett-Ridge, Jillian F Banfield, and Mary K Firestone. Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. *Proc. Natl. Acad. Sci. U. S. A.*, 116(51):25900–25908, December 2019.

[40] Nicole E Tautges, Jessica L Chiartas, Amélie C M Gaudin, Anthony T O'Geen, Israel Herrera, and Kate M Scow. Deep soil inventories reveal that impacts of cover crops and compost on soil carbon sequestration differ in surface and subsurface soils. *Glob. Chang. Biol.*, 25(11):3753–3766, November 2019.

[41] Kristina Wolf, Emma Torbert, Dennis Bryant, Martin Burger, R. Denison, Israel Herrera, Jan Hopmans, William Horwath, Stephen Kaffka, Angela Kong, R. Norris, J. Six, Thomas Tomich, and Kate Scow. The century experiment: the first twenty years of uc davis mediterranean agroecological experiment. *Ecology*, 99, 01 2018.

[42] Simon Roux, Evelien M Adriaenssens, Bas E Dutilh, Eugene V Koonin, Andrew M Kropinski, Mart Krupovic, Jens H Kuhn, Rob Lavigne, J Rodney Brister, Arvind Varsani, Clara Amid, Ramy K Aziz, Seth R Bordenstein, Peer Bork, Mya Breitbart, Guy R Cochrane, Rebecca A Daly, Christelle Desnues, Melissa B Duhaime, Joanne B Emerson, François Enault, Jed A Fuhrman, Pascal Hingamp, Philip Hugenholtz, Bonnie L Hurwitz, Natalia N Ivanova, Jessica M Labonté, Kyung-Bum Lee, Rex R Malmstrom, Manuel Martinez-Garcia, Ilene Karsch Mizrachi, Hiroyuki Ogata, David Páez-Espino, Marie-Agnès Petit, Catherine Putonti, Thomas Rattei, Alejandro Reyes, Francisco Rodriguez-Valera,

Karyna Rosario, Lynn Schriml, Frederik Schulz, Grieg F Steward, Matthew B Sullivan, Shinichi Sunagawa, Curtis A Suttle, Ben Temperton, Susannah G Tringe, Rebecca Vega Thurber, Nicole S Webster, Katrine L Whiteson, Steven W Wilhelm, K Eric Wommack, Tanja Woyke, Kelly C Wrighton, Pelin Yilmaz, Takashi Yoshida, Mark J Young, Natalya Yutin, Lisa Zeigler Allen, Nikos C Kyrpides, and Emiley A Eloe-Fadrosh. Minimum information about an uncultivated virus genome (MIUViG). *Nat. Biotechnol.*, 37(1):29–37, January 2019.

[43] Ho Bin Jang, Benjamin Bolduc, Olivier Zablocki, Jens H Kuhn, Simon Roux, Evelien M Adriaenssens, J Rodney Brister, Andrew M Kropinski, Mart Krupovic, Rob Lavigne, Dann Turner, and Matthew B Sullivan. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.*, 37(6):632–639, June 2019.

[44] Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S Joardar, Vamsi K Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M McGarvey, Michael R Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H Rangwala, Daniel Rausch, Lillian D Riddick, Conrad Schoch, Andrei Shkeda, Susan S Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E Tully, Anjana R Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D Murphy, and Kim D Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, 44(D1):D733–45, January 2016.

[45] Jess W Sorensen, Anneliek M ter Horst, Laura A Zinke, and Joanne B Emerson. Soil viral communities differed by management and over time in organic and conventional tomato fields. May 2023.

[46] Christian Santos-Medellín, Steven J Blazewicz, Jennifer Pett-Ridge, and Joanne B Emerson. Viral but not bacterial community succession is characterized by extreme turnover shortly after rewetting dry soils. February 2023.

[47] Ruonan Wu, Michelle R Davison, Yuqian Gao, Carrie D Nicora, Jason E Mcdermott, Kristin E Burnum-Johnson, Kirsten S Hofmockel, and Janet K Jansson. Moisture modulates soil reservoirs of active DNA and RNA viruses. *Commun Biol*, 4(1):992, August 2021.

[48] Beth F T Brockett, Cindy E Prescott, and Sue J Grayston. Soil moisture is the major factor influencing microbial community structure and enzyme activities across seven biogeoclimatic zones in western canada. *Soil Biol. Biochem.*, 44(1):9–20, January 2012.

[49] Sarah E. Evans, Steven D. Allison, and Christine V. Hawkes. Microbes, memory and moisture: Predicting microbial moisture responses and their impact on carbon cycling. *Funct. Ecol.*, 36(6):1430–1441, June 2022.

[50] Yuntao Li, Jonathan Adams, Yu Shi, Hao Wang, Jin-Sheng He, and Haiyan Chu. Distinct soil microbial communities in habitats of differing soil water balance on the tibetan plateau. *Sci. Rep.*, 7:46407, April 2017.

[51] Jos M Raaijmakers, Timothy C Paulitz, Christian Steinberg, Claude Alabouvette, and Yvan Moënne-Loccoz. The rhizosphere: a playground and battlefield for soilborne pathogens and beneficial microorganisms. *Plant Soil*, 321(1):341–361, August 2009.

[52] Barbara Reinhold-Hurek, Wiebke Bünger, Claudia Sofía Burbano, Mugdha Sabale, and Thomas Hurek. Roots shaping their microbiome: global hotspots for microbial activity. *Annu. Rev. Phytopathol.*, 53:403–424, 2015.

[53] Kelly S Ramirez, Jonathan W Leff, Albert Barberán, Scott Thomas Bates, Jason Betley, Thomas W Crowther, Eugene F Kelly, Emily E Oldfield, E Ashley Shaw, Christopher Steenbock, Mark A Bradford, Diana H Wall, and Noah Fierer. Biogeographic patterns in below-ground diversity in new york city's central park are similar to those observed globally. *Proc. Biol. Sci.*, 281(1795), November 2014.

[54] Sean M Gibbons, J Gregory Caporaso, Meg Pirrung, Dawn Field, Rob Knight, and Jack A Gilbert. Evidence for a persistent microbial seed bank throughout the global ocean. *Proceedings of the National Academy of Sciences*, 110(12):4651–4655, 2013.

[55] Lourens Gerhard Marinus Baas Becking. *Geobiologie of inleiding tot de milieukunde*. Number 18-19. WP Van Stockum & Zoon, 1934.

[56] Anneliek M ter Horst, Temiloluwa V Adebiyi, Daisy A Hernandez, Jane D Fudyma, and Joanne B Emerson. Almond rhizosphere viral, prokaryotic, and fungal communities differed significantly among four california orchards and in comparison to bulk soil communities. June 2023.

[57] Luke S Hillary, Evelien M Adriaenssens, David L Jones, and James E McDonald. RNA-viromics reveals diverse communities of soil RNA viruses with the potential to affect grassland ecosystems across multiple trophic levels. *ISME Commun*, 2:34, April 2022.

[58] Anneliek M ter Horst, Jane D Fudyma, Aurélie Bak, Min Sook Hwang, Christian Santos-Medellín, Kristian A Stevens, David M Rizzo, Maher Al Rwahnih, and Joanne B Emerson. RNA viral communities are structured by host plant phylogeny in oak and conifer leaves. *Phytobiomes Journal*, pages PBIOMES–12–21–0080–R, April 2022.

[59] Pauline C Göller, Jose M Haro-Moreno, Francisco Rodriguez-Valera, Martin J Loessner, and Elena Gómez-Sanz. Uncovering a hidden diversity: optimized protocols for the extraction of dsDNA bacteriophages from soil, 2020.

[60] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder,

Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld, and Rob Knight. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, 7(5):335–336, May 2010.

[61] Joseph Edwards, Christian Santos-Medellín, and Venkatesan Sundaresan. Extraction and 16S rRNA sequence analysis of microbiomes associated with rice roots. *Bio Protoc*, 8(12):e2884, June 2018.

[62] Matthew T Agler, Jonas Ruhe, Samuel Kroll, Constanze Morhenn, Sang-Tae Kim, Detlef Weigel, and Eric M Kemen. Microbial hub taxa link host and abiotic factors to plant microbiome variation. *PLoS Biol.*, 14(1):e1002352, January 2016.

[63] M Gardes and T D Bruns. ITS primers with enhanced specificity for basidiomycetes–application to the identification of mycorrhizae and rusts. *Mol. Ecol.*, 2(2):113–118, April 1993.

[64] Thomas J White, Thomas Bruns, Sjwt Lee, John Taylor, and Others. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR protocols: a guide to methods and applications*, 18(1):315–322, 1990.

[65] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, August 2014.

[66] Brian Bushnell. BBMap: a fast, accurate, splice-aware aligner. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2014.

[67] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31(10):1674–1676, May 2015.

[68] Kristopher Kieft, Zhichao Zhou, and Karthik Anantharaman. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, 8(1):90, June 2020.

[69] Matthew R Olm, Christopher T Brown, Brandon Brooks, and Jillian F Banfield. drep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.*, 11(12):2864–2868, December 2017.

[70] B Longmead and S L Salzberg. Fast gapped-read alignment with bowtie2. *Nat. Methods*, 9(4):357–359, 2012.

[71] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[72] Ben J Woodcroft. CoverM: Read coverage calculator for metagenomics.

[73] Doug Hyatt, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11:119, March 2010.

[74] Sean R Eddy. Accelerated profile HMM searches. *PLoS Comput. Biol.*, 7(10):e1002195, October 2011.

[75] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, 12(1):59–60, January 2015.

[76] Andre P Masella, Andrea K Bartram, Jakub M Truszkowski, Daniel G Brown, and Josh D Neufeld. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics*, 13:31, February 2012.

[77] Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. DADA2: High-resolution sample inference from illumina amplicon data. *Nat. Methods*, 13(7):581–583, July 2016.

[78] Qiong Wang, George M Garrity, James M Tiedje, and James R Cole. Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, 73(16):5261–5267, August 2007.

[79] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, 41(Database issue):D590–6, January 2013.

[80] Kessy Abarenkov, Allan Zirk, Timo Piirmann, Raivo Pöhönen, Filipp Ivanov, R Henrik Nilsson, and Urmas Kõljalg. UNITE QIIME release for fungi, May 2021.

[81] Team RCore. R: A language and environment for statistical computing. R foundation for statistical computing, vienna, austria, 2016.

[82] J Oksanen, F G Blanchet, M Friendly, R Kindt, P Legendre, D McGlinn, P R Minchin, R B O'Hara, G L Simpson, P Solymos, and Others. vegan: Community ecology package. R package version 2.5-2. 2018, 2018.

[83] Emmanuel Paradis and Klaus Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3):526–528, February 2019.

[84] D Kahle and H Wickham. ggmap: Spatial visualization with ggplot2. the R journal, 5 (1), 144-161. *URL https://journal. r-project. org/archive/2013-1/kahle.*

[85] H Wickham. ggplot2: elegant graphics for data analysis. data. 2016.

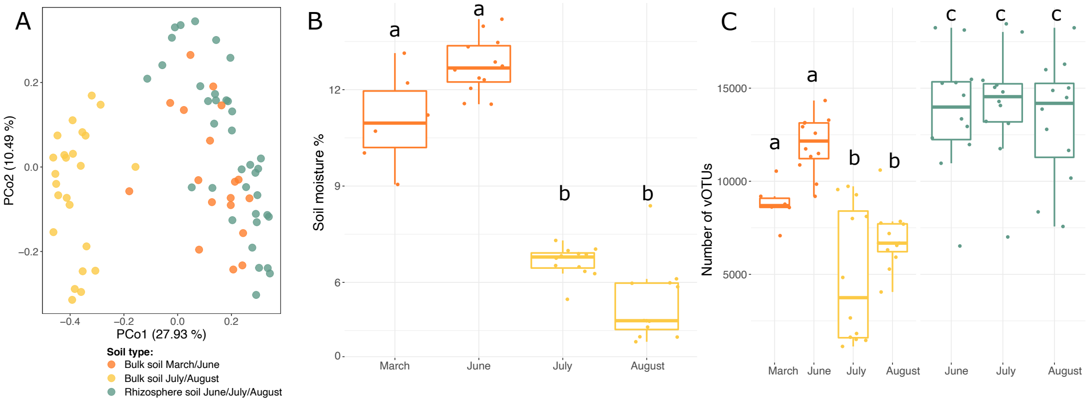


Figure 5.1: Soil viral community and vOTU abundance patterns with soil moisture A: Principal coordinates analysis (PCoA), based on Bray-Curtis dissimilarities derived from the table of vOTU abundances of viral community composition in the 78 viromes. B: Soil moisture percentage in bulk soils for each time point C: Total number of vOTUs recovered per time point, in bulk soils versus rhizosphere soils (colors correspond to the legend in panel A).

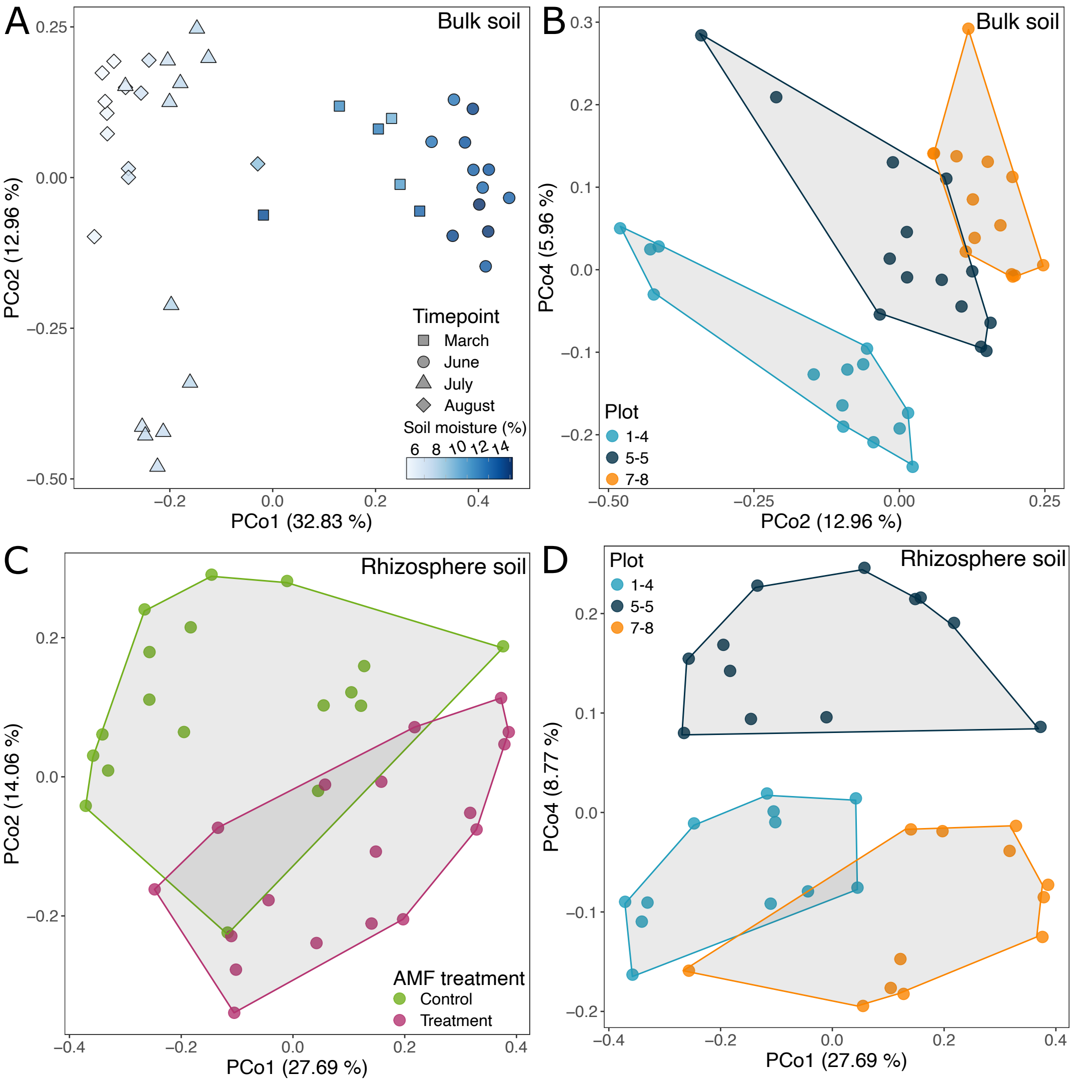


Figure 5.2: Overarching compositional patterns for tomato bulk soil and rhizosphere viral communities. PCoA of viral community composition in A: the 42 bulk soils, colored by soil moisture percentage. B: the 42 bulk soils, colored by plot location. C: the 36 rhizosphere soils, colored by treatment with AMF. D: the 36 rhizosphere soils, colored by plot location in the field.

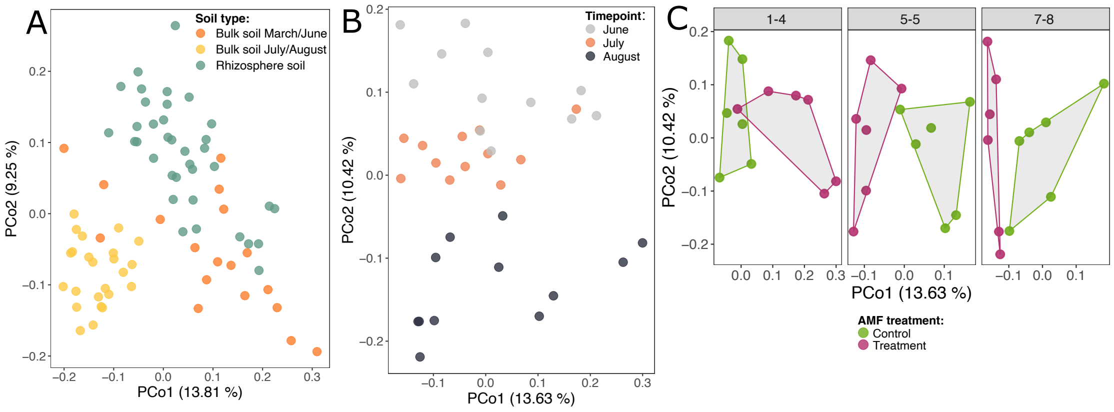


Figure 5.3:Overarching compositional patterns for tomato bulk soil and rhizosphere prokaryotic communities. A: PcOA based on Bray-Curtis dissimilarities derived from the table of OTU abundances of prokaryotic community composition in the 78 samples. B: PcOA of the 36 rhizosphere samples, colored by time point of sampling. C: PcOA of the 36 rhizosphere samples, separated per plot location and colored by treatment

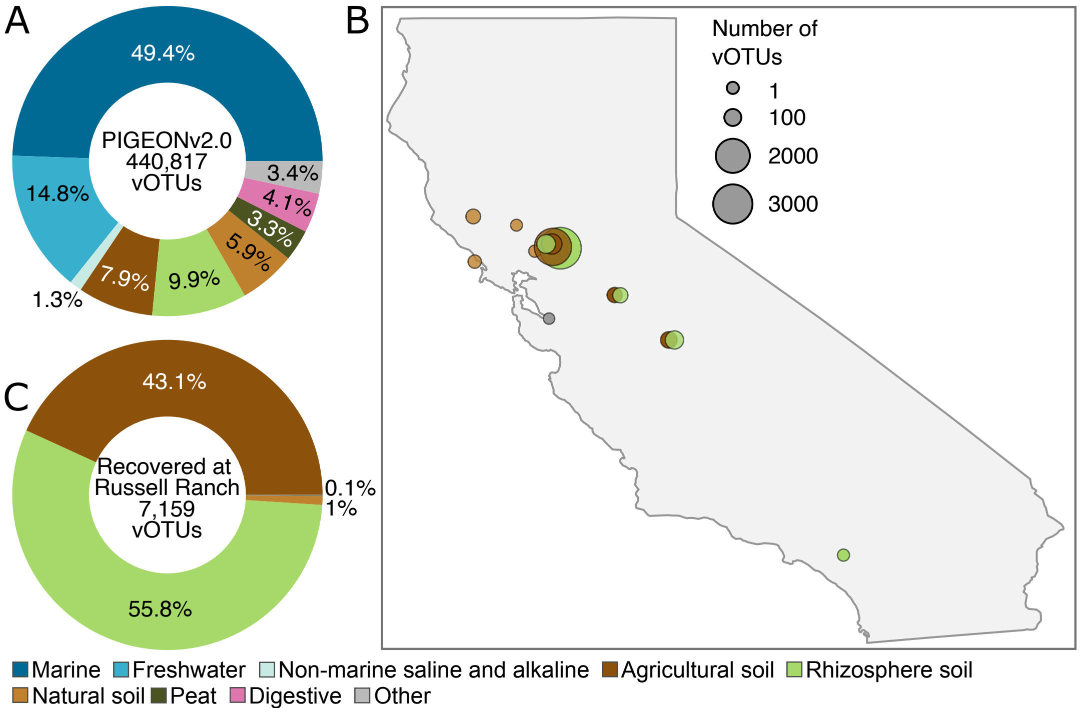


Figure 5.4: California wide distribution of Russell Ranch vOTUs, using the PIGEON v2.0 database. A: Composition of the PIGEONv2.0 database of 440,817 vOTU sequences, colored by environment, excluding vOTUs from Russell Ranch. B: Relative proportions of all vOTUs recovered from PIGEONv2.0 at Russell Ranch, colored by original environment. C: vOTUs (n=7,159) from PIGEONv2.0 recovered at Russell Ranch by read mapping, according to the location where they were first recovered, colored by the environment they were first recovered in

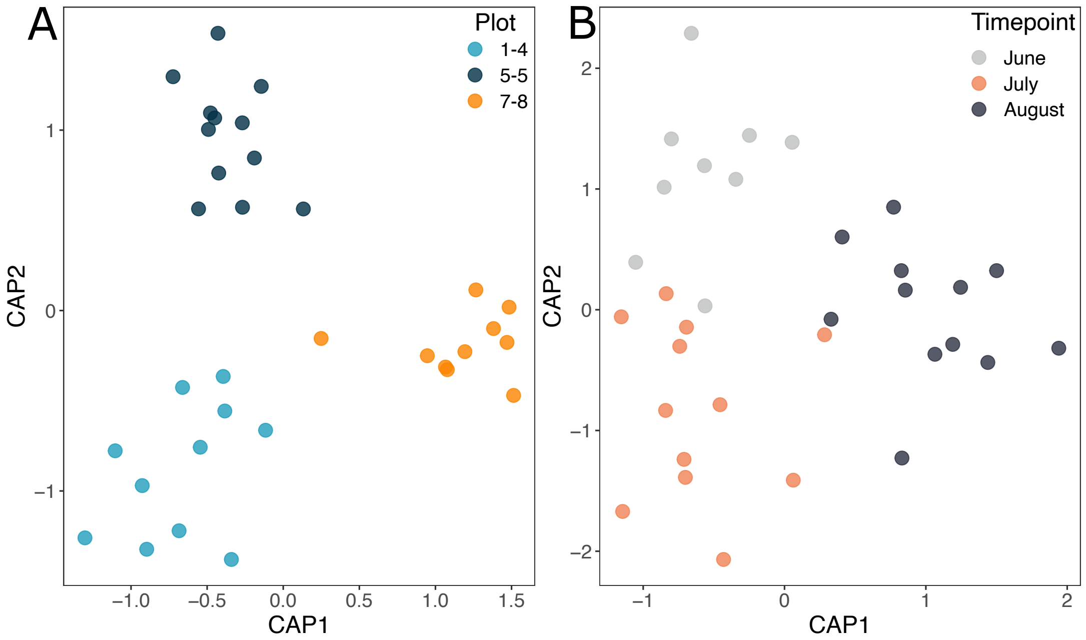


Figure 5.5: Community composition patterns of tomato rhizosphere RNA viral communities. A,B: Canonical analyses of principal coordinates (CAP) of the 33 rhizosphere samples, A: colored by plot location in the field and B: colored by time point of sampling