Direct and indirect viral associations predict coexistence in wild plant virus communities

Anna Norberg[1], Hanna Susi[2], Suvi Sallinen[2], Pezhman Safdari[2], Nicholas Clark[3], and Anna-Liisa Laine[1]

[1]University of Zurich
[2]University of Helsinki
[3]The University of Queensland

March 07, 2024

Abstract

Integration of community ecology with disease biology is viewed as a promising avenue for uncovering determinants of pathogen diversity, and for predicting disease risks. Plant-infecting viruses represent a vastly underestimated component of biodiversity with potentially important ecological and evolutionary roles. We performed hierarchal spatial analysis of wild plant populations to characterise the diversity and coexistence structure of within-host virus communities, and their predictors. Our results show that these virus communities are characterised by single infections of few, dominating virus taxa as well as diverse, non-random coinfections. Using a novel graphical modelling framework we demonstrate that after accounting for environmental heterogeneity at the level of both individual host plants and populations, most virus co-occurrence patterns can be attributed to virus-virus associations. Moreover, we show that conditioning variables changed virus association networks especially through their indirect effects. This highlights a previously underestimated mechanism of how human-driven environmental change can influence disease risks.

Introduction

Viruses, like all organisms, occur as communities with other species (Díaz-Muñoz 2017). Importantly, the diversity of co-occurring viruses may have profound impact on virus evolution and epidemiology (Alcaide *et al.* 2020). Hence, understanding the conditions and scales at which virus communities vary is critical when predicting how virus communities respond to environmental change or to disease control measures (Gilman *et al.* 2010; Malmstrom *et al.* 2011; Massart *et al.* 2017; Halliday *et al.* 2020). Over the past decade metagenomic surveys exploring virus diversity in wild hosts ranging from plants to insects and mammals (Wren *et al.* 2006; Roossinck 2010, 2012; Ng *et al.* 2011; Letko *et al.* 2020) have revealed a tremendous, largely undescribed, virus diversity across environments. It is becoming clear that the diversity of virus taxa even within the same host can be highly variable, and that viral co-occurrences are common in nature (Roux *et al.* 2015; Díaz-Muñoz 2017; Munson-Mcgee *et al.* 2018; Alcaide *et al.* 2020).

Predicting how pathogen communities respond to environmental change requires a principled community ecology framework to disentangle possible drivers of coexistence patterns (Johnson *et al.* 2015; Seabloom *et al.* 2015). As Popovic *et al.* (2019) recently outlined, there are three main drivers of observed species co-occurrence patterns (Figure 1). First, two species might share similar responses to environmental variables such as temperature (Suzuki *et al.* 2014; Obrępalska-Stęplowska *et al.* 2015; Alcaide *et al.* 2021). Second, two species may exhibit similar responses to the occurrence of a third species: for example, some viruses require another virus species to *mediate* their replication within a host (Pirone & Blanc 1996; Syller 2012), resulting in an indirect association between the dependent species. Third, two species might exhibit a direct, biotic

association: a virus can for example facilitate the establishment of another (Six & Klug 1973; Waterhouse & Murant 1983). Hence, the sequence of arrival (Fukami 2015), can have far reaching implications for the structure of pathogen communities (Karvonen *et al.* 2019).

Spatial structure and abiotic habitat conditions are considered important for structuring pathogen communities (Bergner *et al.*2020) via their effect on both hosts and pathogens (Makiola *et al.* 2019) (Figure 1). While demographic stochasticity is important at smaller spatial scales (Tilman 2004), abiotic environmental heterogeneity increases with increasing spatial scale, promoting greater coexistence through species-specific environmental responses (Chase & Leibold 2003). Dispersal presumably decreases coexistence by forcing species to interact and by homogenising intra- and interspecific interactions (Snyder & Chesson 2003). Most plant viruses are vector-dispersed, and thus the distribution of vectors can also influence the structure of plant pathogen communities (Schröder *et al.* 2017). A special characteristic of pathogen communities is that their immediate environment is a living organism in itself, and hence the influence of the abiotic environment (e.g. weather) on pathogens can be either direct or mediated by the host and/or vectors (Figure 1). Moreover, the distinction between abiotic and biotic effects becomes blurred, as the interaction between a host and a pathogen can be considered both an environmental effect as well as a biotic association (Figure 1). Pathogens can only occur where they have susceptible hosts and thus spatial variation in resistance is expected to have direct impacts on pathogen (co-)occurrence patterns (Jousimo *et al.*2014; Carlsson-Granér & Thrall 2015).

Ecological knowledge often relies on observational methods, but inferring signals of biotic interactions from co-occurrence data is challenging (Blanchet *et al.* 2020). However, studying species associations through the perspective of conditional probabilities helps to overcome some of these challenges. Conditional probability refers to the probability that two species will be found together *after* controlling for the other species in the network. Implementations for ecological applications have been recently developed (Harris 2016; Clark *et al.* 2018; Popovic *et al.* 2019). *Markov random fields* (MRFs) are a group of graphical network models which enable the estimation of conditional dependencies from networks of interacting variables (Sutton & McCallum 2011; Clark *et al.* 2018). In addition to analysing conditionally dependent species co-occurrence patterns after accounting for the occurrences of all other species in the community, also additional covariates can be included, resulting in *conditional random fields* (CRFs) (Azaele *et al.* 2010; Harris 2016; Clark *et al.* 2018; Popovic *et al.* 2019). MRFs and CRFs allow us to discover how species are associated with each other and their environment, and importantly also how the environment influences these associations (Clark *et al.* 2020).

Here, we aim to determine the relative roles of spatial environmental heterogeneity and both direct and indirect virus-virus associations in determining the coexistence structure of within-host virus communities of naturally occurring *Plantago lanceolata* populations, identified via deep sequencing of small RNAs from field collected plant samples (Kreuze *et al.* 2009). To overcome the challenges of inferring signals of biotic interactions from co-occurrence data (Blanchet *et al.* 2020), we incorporate relevant environmental covariates at the scale of both viruses and the hosts, analyse co-occurrence patterns through conditional probabilities, and use a reasonably-sized data set of 400 plants, sampled hierarchically at biologically relevant scales (across a population network and within populations), covering a wide environmental range. Specifically, we ask:

Q1. What are the relevant spatial scales of virus diversity and co-occurrences, and can we detect signals of coexistence mechanisms influencing the relationship between viral diversity and co-occurrence at different spatial scales? Q2. Do host and habitat characteristics, and spatial structure of the host populations influence virus community structure, or can we explain the structure solely using the direct and indirect associations between the viruses? Q3. After accounting for the effects of the host and habitat characteristics, and spatial patterns on virus distributions, are there any remaining non-random negative or positive direct associations between viruses? Q4. Do associations between viruses change when comparing conditional network models (CRFs) with *only* host- and/or habitat-related and/or spatial explanatory variables included in the model, demonstrating how these different sources of environmental heterogeneity explain the structure of the virus community?

Our results demonstrate that there are non-random co-occurrence patterns between viruses, which are only partly resolved by the host and habitat characteristics and spatial structure. We find that the majority of the explained virus co-occurrence patterns can be attributed to direct and indirect associations among the viruses.

Material and Methods

Study system and sampling

Our study focused on viruses infecting natural populations of*Plantago lanceolata* in the Åland Island, SW of Finland. These populations have been monitored since 1993, and they form a highly fragmented network of ~4000 populations (Ojanen *et al.* 2013). We used a stratification process to select 20 focal populations that were sampled in early June 2017 for their virus communities. We collected leaf samples from 20 randomly selected plants in each population, resulting in altogether 400 samples for small RNA sequencing. For more details on the study system and sampling, see Supplementary Material and Methods.

VIRUS DETECTION

For virus identification, RNA was extracted from the samples using phenol-chloroform-isoamylalcohol extraction (Chang *et al.* 1993). The small RNA (sRNA) was sequenced using Illumina Hiseq (2009). After quality check and adaptor removal, the obtained reads were analysed using the VirusDetect pipeline (Zheng *et al.* 2017) to obtain Operational Taxonomical Units (OTUs). The OTUs were classified according to their host range and plant associated OTUs were selected for our analysis (Hulo *et al.* 2011). The OTUs were named at the virus family level and are henceforward referred to as "viruses". We compiled a presence-absence matrix describing the virus community in each host individual and host population.

EXPLANATORY VARIABLES

Host- and habitat-related variables

During sampling we recorded the locations of the plants with GPS and measured the size of the plants (Table 1). We recorded signs of herbivore damage, and surveyed surrounding vascular plant communities between 1930 June 2017 by counting the number of vascular plant species within multiple one-square-meter quadrats in each population. We used the vegetation data to calculate the Shannon diversity index (Shannon 1948) for each population (Table 1).

We estimated the proportion of agricultural area surrounding each study population from Corine Land Cover (CLC, version 2020_20u1) with QGIS (QGIS Development Team 2019) by creating a one-kilometre buffer zone around each study population following the patch borders and calculating the proportion of 20 m $\times$ 20 m pixels falling under agricultural land use category within this buffer zone (Table 1). We also estimated the coverage of *P. lanceolata* foliage in square meters in each population. We quantified the connectivity of a host population with respect to other populations by calculating the Euclidian distances between populations and calibrating these measures by the species dispersal capacity (Hanski 1999).

We obtained weather observations for the study populations from the Finnish Meteorological Institute (Aalto *et al.* 2016). We calculated the number of severe winter days during the winter before the sampling season (2016-2017), and the sum of temperatures of the effective summer days during previous summer (2016). See Table 1 and Supplementary Material and Methods for more details, including biological justification for the selected variables.

Spatial variables

We included spatial variables implemented as Moran's eigenvector maps (MEMs, Dray *et al.* 2006). We calculated MEMs based on the GPS coordinates of the sampled host plants, with the assumption of positive autocorrelation. We estimated the significance of individual eigenvectors with a permutation test for the Moran's I statistic and included all the eigenvectors with Moran's I $>$0.7, as well as the last significant

one, thus representing both coarse and fine scale signal of significant spatial autocorrelation. We used the R environment (R Core Team 2020) and packages 'spdep' and 'adespatial'.

STATISTICAL METHODS

Descriptive analyses

We began with descriptive analyses and illustration of the virus co-occurrence structure. We assessed the nestedness of the data by organising the virus community matrix based on overlap in virus presences among plants and decreasing fill (Almeida-Neto *et al.*2008). We calculated C-scores (Stone & Roberts 1990) for the virus community at the level of individual host plants and populations (Figure 2). We calculated the numbers of pairwise virus co-occurrence combinations in host plants, as well as the pairs of viruses that never co-occur. We used the full dataset of 25 viruses (See Results) for these descriptive analyses.

We quantified the relationship between the cumulative virus richness and co-occurrence patterns with respect to increasing area sampled by calculating the mean species–area–curve and mean coexistence–curve (Hart *et al.* 2017). Both curves were constructed by randomly selecting one sampled host plant, and then increasing the spatial scale and sample by including the next closest plant to the species richness or co-occurrence calculation. We calculated the mean curves by repeating this 100 times, selecting a different initial plant for each round. We also calculated the maximum number of co-occurring virus pairs from the number of taxa observed thus far. For more details, see Supplementary Material and Methods and Supplementary Results Figure S2-3.

Markov Random Field networks

To understand the role of host- and habitat-related variables and spatial configuration of the hosts, as well as virus interactions (Figure1, Table 1) in explaining the network structure, we fitted both Unconditional (MRF) and Conditional Markov Random Field models (CRF) to the virus community data (Clark *et al.* 2018). Markov Random Fields (MRFs) are graphical models, which can represent complex distributions as network graphs. These networks consist of *nodes*and *edges* , corresponding to the observed variables within the data, and to the probabilistic interactions between variables that need to be estimated. The edge associations are undirected, meaning that the effect of one node on another is reciprocal. If there is no edge between two nodes in the estimated graph, these nodes are conditionally independent from one another, whereas if there is an edge, these nodes are conditionally dependent, *after* accounting for the other node effects in the graph model (Cheng *et al.* 2014). Conditional Random Fields allow for these dependencies among nodes to be further conditional on other covariates (Cheng *et al.* 2014; Clark*et al.* 2018). Hence, the values for the edge associations can change in the presence of these covariates, and the resulting graph model illustrates the pairwise associations between viruses in host plants, conditional not only on the rest of the virus community, but also on the covariates included in the model (Table 1).

The modelling framework is described in detail by Clark *et al.*(2018). Briefly, we modelled the log-odds of detecting virus $i$given covariate x and occurrence of virus j with

$$\log\left(\frac{P\left(y_i=1\middle|y_{\setminus i},\ x\right)}{1-P\left(y_i=1\middle|y_{\setminus i},\ x\right)}\right)=\ \alpha_{i0}+\beta_i^T x+\sum_{j:j\neq i}\left(\alpha_{ij0}+\beta_{\mathrm{ij}}^T x\right)\ \mathrm{y}_j\ ,$$

where y_i is the vector of presences and absences of virus$i$; $y_{\setminus i}$ denotes the presences and absences of all other viruses except i ; α_{i0} is the virus-level intercept; and β_i^T is the effect of covariate x on the occurrence probability of virus i . Parameters $\alpha_{ij0}$and β_{ij}^T represent the associations between viruses, conditional on the occurrences of all the non-focal viruses (other than the focal virus i).

We used data on viruses with at least 10 occurrences (16 viruses) in the entire virus community matrix (i.e. minimum prevalence of 2.5% of sampling units). To understand how environmental characteristics and the host affect the virus community, we included several explanatory, conditioning variables in the model (Table 1), describing: 1) The level of s*patial autocorrelation* of the host populations (implemented

as Moran's eigenvectors) 2) *habitat-related* characteristics, namely the *quality* of the habitat of the host plants (the connectedness (S) of the focal *P. lanceolata* population to other populations, agricultural land use (percentage of the surrounding landscape) and the Shannon diversity of the local plant community, which have been demonstrated to influence virus occurrences in this system (Susi & Laine 2020), and as the *weather* conditions of local populations (severity of the previous winter and temperature sum over the effective summer days during previous summer); as well as 4)*host-related* characteristics of the focal host plants (host population size, host plant size and signs of herbivory). See Table 1 for full details.

Altogether our dataset used for modelling consisted of 16 virus taxa and 16 explanatory variables (Table 1), resulting in 272 coefficients in each regression. To avoid overfitting, regularisation has been implemented in the method through least absolute shrinkage and selection operator (LASSO), forcing some regression coefficients to zero, and thus performing variable selection and reducing the risk of overfitting (Clark *et al.* 2018). To achieve an undirected network and symmetry within the coefficients of conditional dependence (i.e.$\alpha_{ij0} = \alpha_{ji0}$ and$\beta^T_{\mathrm{ij}} = \beta^T_{\mathrm{ji}}$) we take the mean of the corresponding estimates, which is the default setting of the applied algorithm (Clark *et al.* 2018).

We fitted six model variants in total: 1) an Unconditional Markov Random Field model (referred to as 'MRF'), with only virus occurrences included, 2) a Conditional Markov Random Field model (CRF) with only habitat- and host-related (collectively referred as 'environmental', see Table 1) variables included as additional constrains ('CRFenv'), 3) a CRF model with only host-related variables included as additional constraints ('CRFhost'), 4) a CRF model with only spatial variables and variables related to habitat (quality and weather) included as additional constraints ('CRFhabitat'), 5) a CRF model with only spatial variables included as additional constrains ('CRFspat'), 6) a Conditional Markov Random Field model with both all environmental as well as spatial variables included as additional constrains ('CRFfull'). We will refer to the variants (2-6) collectively as 'CRF models' or 'CRFs'.

We evaluated the model fit by calculating the Area Under Curve values (Hanley & Mcneil 1982) using the full data set. We used cross validation (with four folds) to estimate model generality by comparing predicted and observed outcomes simultaneously for all taxa. To account for parameter uncertainty of the final model, we modelled 100 bootstrapped replicates for the model. If the 90% confidence interval of bootstrapped coefficients did not overlap with zero, we considered the variable to have a statistically significant effect. To test whether the viral associations were phylogenetically conservative, we compared the direct associations drawn from all our network models to the phylogenetic relationships of the viruses, constructed from taxonomy, by conducting a Mantel test between the matrices.

All results were produced with R (version 4.0.2, R Core Team 2020), and packages 'vegan' (Oksanen *et al.* 2019), 'MFRcov' (Clark *et al.* 2018), 'igraph' (Csardi & Nepusz 2006), along with their dependencies. An R package called 'meta17-network' including the analytical pipeline, data, and documentation for full reproduction of the results can be found in Github (aminorberg/meta17network-pkg).

Results

Description of the data

The plant small RNA sequencing yielded on average 18 222 557 reads (min 11148864, max 63 508 829 and sd 3 944 297) per sample. VirusDetect assembled 512 908 contigs in total (min 39 nt, mean 63 nt, max 5408 nt and sd 45 nt). There were 5504 contigs (min 40 nt, mean 90 nt, max 1232 nt) with hits to known virus taxa (Gorbalenyai & Siddell 2021) with average sequence similarity of 82% (min 23%, max 100%). Our analyses are focused on the 25 identified plant-associated viruses.

Out of the 400 sampled host plants, four samples were discarded due to missing explanatory variable data. Thus, we had 396 sampled plants, resulting in 9 900 unique virus observations. The prevalence of the individual viruses varied from 0.5 to 36% in the whole data. Of the 396 plants, 29% were uninfected, 32% hosted a single infection, and 39% of the plants hosted multiple infections, of which 7% consisted of five or more viruses (Figure 2A).

Descriptive analyses

The virus communities exhibit a significantly nested structure: the mean C-score over the viruses was 290.52 at the individual plant level (with the maximum of 87155 possible checkerboard units); and 2.97 at the population level (with the maximum of 890 possible checkerboard units) (Figure 2B). Almost all viruses (24 OTUs) were present in the most virus-rich host population (population 861; Figure 2B and 2D), while other populations hosted smaller subset of the viruses, with the simplest metacommunity consisting of six viruses (population 3222; Figure 2B and 2C).

All the viruses co-occurred with another virus at least once (Figure 3A). The most abundant viruses, Closteroviridae and Caulimoviridae, occurred as single infections in 39 and 33 host plants, respectively, which made up nearly one third of their total occurrences (133 and 122, respectively). The number of unique virus co-occurrence combinations (of two viruses or more) in the whole data set was 111. Only Avsunoviroidae and Alphasatellitidae never co-occurred (of which the latter was so rare that it was not included in the network models, see below).

The species-area curve shows that all 25 viruses have been encountered when on average 264 plants have been sampled (Figure 3B). The difference between the species-area and coexistence-area curves shows how the cumulative sum of co-occurring viruses is lower than the cumulative viral richness would predict, until the coexistence curve reaches its maximum at 387 sampled host plants. For example, when 100 host plants have been sampled, the number of viruses observed was 20, which would enable 190 co-occurring pairs of viruses while the average number of observed co-occurring virus pairs at that point was 168.

Markov Random Field networks

The CRF models (with additional conditioning variables) clearly outperformed the MRF (with only virus occurrences included): the AUC for MRF was 0.69 while for the CRFs it varied between 0.87 and 0.89. Based on cross-validation, there were no pronounced discrepancies between the different CRFs, but the overall performance of the CRFs was better than that of the MRF model: the 50% quantile for the mean for predicting both true positives and negatives correctly for the MRF was 0.76, whereas the corresponding value for the CRF variants was around 0.91. The MRF predicted more false positives, whereas the CRFs predicted more false negatives. The mean values for different performance measures are reported in the Supplementary Results Table S1.

To understand the changes in the network resulting from the addition of conditioning variables, we compared the virus-virus-associations between viruses based on the MRF and the different CRF variants. The MRF revealed mostly positive associations between the viruses (Figure 4A). After including spatial, habitat and host-related variables (Table 1), some of the associations between the viruses diminished or disappeared, and all of the conditional associations were positive (Figure 4B). The number of significant virus co-occurrences captured by the MRF model was 50 (Figure 4A). The corresponding number for the CRFfull model was 16 (Figure 4B). The CRFs incorporating subsets of conditioning variables identified intermediate amounts of associations: 30 for CRFhost, 38 for CRFspat, 28 for CRFhabitat, and 18 for CRFenv.

Although several associations could be explained exchangeably with habitat- or host-related variables, many associations were also explained solely by either habitat- or host-related variables (Figure 4C-D). For example, Bromoviridae showed a high number of associations with other viruses (Figure 4A), but was not explained by host-related or spatial variables (Figure 4C and E). However, several of its strong associations with other viruses were explained by the habitat-related effects (Figure 4D). The 11 association links captured by the CRFfull model were captured by all the other conditional and unconditional model variants as well (Figure 4B). We will refer to these associations as 'permanent'. In this network, Bromoviridae and especially Secoviridae appeared as hubs, with five association links to other viruses. These permanent associations represented direct interactions between viruses that could not be explained with indirect effects of the rest of the virus community nor any combination of additional conditioning variables.

Next, to understand how host- and habitat-related variables and spatial configuration of the hosts influence

virus community structure, we investigated the direct effects of the additional conditioning variables. All the significant direct effects of the environmental and spatial variables were for either Caulimoviridae or Geminiviridae (Table 2): e.g., increasing agricultural land use in the surrounding landscape increased the occurrence probability of Caulimoviridae, and host population size predicted higher occurrence probability for Geminiviridae.

None of the indirect effects of the additional conditioning variables influenced the associations between viruses so that the direction of the direct virus-virus association would change from positive to negative or vice versa. However, all conditioning variables except host plant size and agricultural land use had some effect(s) on some viral associations (Table 2). In terms of the number of virus-virus-associations influenced, the most influential indirect effects were the spatial structure of the host populations (MEMs) and host population connectivity. All the effects of the first, coarse scale spatial variable (MEM1) were positive, whereas the effects of the spatial variables at increasingly finer scales (MEM2-4) were all negative. Increasing connectivity of the host population had both negative and positive effects on the virus-virus associations: for example, higher connectivity lowered the occurrence probability of Avsunoviridae in the presence of Bromoviridae (and vice versa, symmetrically). There were altogether eight significant herbivory-related effect, all of them positive (Table 2).

Discussion

Here, we use community modelling to uncover determinants of wild plant virus diversity, a vastly unexplored component of biodiversity (Roossinck 2011). We show that virus communities exhibit a clear nested structure with abundant co-occurrences, both pairwise and higher-order. The observed coexistence patterns are mediated by host plant characteristics, abiotic environment and spatial structure, as well as associations among the viruses. Many present-day threats posed by infectious diseases involve interactions that are manifested across nested scales of biological organisation (Johnson *et al.* 2015), and our findings shed light on how plant virus coexistence is maintained at different spatial scales.

The simplest within-host viral communities detected in our data consisted of single infections, but nearly half (46%) of the infected plants hosted multiple infections. Single infections were typically observed only for the few most common viruses, and half of all the viruses never occurred as single infections. Our findings are in line with other studies that have found high levels of coinfection (Al Rwahnih *et al.* 2009; Rey *et al.* 2012; Tugume *et al.*2016). However, our approach enabled the detection of virus communities in a substantial number of host individuals, with capacity to detect viromes consisting of up to 24 distinct taxa. This suggests that virus communities are highly variable within and among hosts.

From the difference between species-area and coexistence curves, we see that a larger number of hosts is needed to maintain the coexistence patterns that could be derived from the overall virus richness. Indeed, increasing environmental heterogeneity and varying responses of viruses to this heterogeneity promote coexistence by reducing competition, as predicted by the classic species sorting paradigm, which has been shown to be influential for microbial communities (see e.g. Székely & Langenheder 2014 and refs. within). Combined with the weakening effects of demographic stochasticity and the homogenising effects of dispersal, these mechanisms lead to more stable coexistence at increasing spatial scales (Hart *et al.* 2017; Levine & Hart 2020).

We found that the conditional network models (CRFs), outperformed the model incorporating only associations between viruses (MRF) in explaining virus community structure. The differences between the CRF model variants were less pronounced, as seen from their equal model fit and performance. Hence, we conclude that there is environmental variation and processes operating at different spatial scales that influence the wild virus community coexistence structure both directly and indirectly. The environmental characteristics and the spatial variables explain community structure almost interchangeably. Previously, e.g. within-host diversity of pathogens has been shown to increase with latitude, while pathogen turnover follows an opposite trajectory, suggesting limited transmission in lower latitudes (Seabloom *et al.* 2010).

Our analysis revealed spatial variables to influence associations between viruses, resulting in indirect effects

on their co-occurrence. These viruses are vector-dispersed, and most likely dispersal-limited at larger spatial scales (Pleydell *et al.* 2018), leading to less homogenisation at these scales. However, at the scale of a host populations, variation among host genotypes (Sallinen *et al.*2020) and demographic stochasticity (through e.g. herbivory, as seen from the indirect effects of herbivore damage on the associations between viruses) are expected to be important, as well as abiotic interactions mediating coexistence (Kozanitas *et al.* 2017).

After accounting for the effects of the host and habitat characteristics, as well as spatial structure on virus distributions, non-random, significant positive associations between viruses remain. Although these associations are based on purely observational co-occurrence patterns, given the spatial scale of our sampling and our method for detection that targets the host's defence response (Kreuze*et al.* 2009), we consider these indicative of biotic interactions between the viruses within a host plant (Wintermantel *et al.*2008). The interactions among coinfecting viruses may involve positive effects on replication (Pruss *et al.* 1997; Taiwo *et al.*2007), or even obligate dependencies as some viruses require their specific helper virus in order to complete their lifecycle (DaPalma*et al.* 2010). Viruses may also suppress host immunity allowing subsequent infections to escape recognition by host immunity (González-Jara *et al.* 2005). As expected, the inclusion of explanatory variables reduced the number of direct associations between viruses identified by the models, suggesting that shared environmental responses play an important role in the assembly of the communities (see e.g. Leathwick *et al.* 2006; Ovaskainen & Soininen 2011). After accounting for host- or habitat-related or spatial variables, support for several direct associations (e.g. Tombusviridae-Alphaflexiviridae, Betaflexiviridae-Potyviridae and Tombusviridae-Betaflexiviridae) was no longer detected.

Importantly, we found that the inclusion of the additional conditioning variables changed the association networks especially through their indirect effects. For example, we found a significant indirect effect of the coarsest-scale spatial variable (MEM1) on the association between Fimoviridae and Alphaflexiviridae, while the direct effect of this spatial variable was not significant. Hence, the occurrence probability of Alphaflexiviridae is jointly affected by the occurrence of Fimoviridae as well as spatial structure. Due to symmetry, Fimoviridae is similarly affected jointly by the occurrence of Alphaflexiviridae and spatial structure. These indirect significant effects were more frequent (in total 32 effects) in comparison to direct effects (in total 10). The indirect links can be indicative of either biotic interactions between viruses only manifesting under certain environmental conditions, or these environmental conditions having an effect only in interaction with the other virus (Kozanitas *et al.* 2017). Such indirect effects have traditionally been challenging to detect, yet not accounting for them can lead to over- or underestimation of signals of biotic interactions in co-occurrence data (Blanchet *et al.* 2020).

Although our virus community data set along with its environmental explanatory variables is extensive and of high quality, our results are limited by our sample size and its effects on the parameter estimation of our modelling method. We use regularisation to avoid overfitting, but we note that the estimated parameters could in theory change with more data included. However, as seen from the species-area and coexistence-area curves, our sampling effort captures the detected virus diversity already before all the samples have been included, indicating a promising sample size.

Our results demonstrate that natural plant virus communities are characterised by single infections of few, dominating virus taxa as well as diverse, non-random coinfections. Virus diversity can be explained by coexistence-promoting mechanisms, some of which we could tease apart with our modelling. We show that host and habitat characteristics, as well as spatial structure, resolve some of the observed co-occurrence patterns, to some degree interchangeably. Importantly, we find that some virus-virus associations are mediated by either host or habitat characteristics, or the spatial structure of the host populations. However, a substantial part of the explained virus co-occurrence patterns can be attributed to positive, direct associations among the viruses. Moreover, we show that additional conditioning variables changed virus association networks especially through their indirect effects. Thus, our study contributes to increasing understanding on how plant virus coexistence and thus biodiversity is maintained at different spatial scales. Our results highlight a previously underestimated mechanism of how human-driven environmental change can influence disease risks by changing biotic associations between viruses that are conditional on their environment.

Acknowledgements

We thank Krista Raveala, Mikko Jalo, Pauliina Hyttinen, and Mikko Immonen for assisting in the field work, Laura Häkkinen for assisting in the laboratory work, and Elina Numminen for helping with the use of population database. We also thank Guillaume Blanchet for helpful discussions. This work was funded by the European Research Council (4100097 RESISTANCE) and Academy of Finland (334276) to A-LL, Academy of Finland (321441) to HS, and LUOVA doctoral program fellowship to SS.

Figure and table legends

Figure 1. A schematic representation of direct and indirect associations between viruses in the context of their hosts and environment. Within plant populations, individual plants host viruses. The viruses have associations with each other, either through direct mutualistic or antagonistic interactions (bold font), or indirect associations mediated by another virus, shared vector, and/or host immunological responses (small font). The environment can influence the (co-)occurrences either directly or through host effects at the level of host populations or individuals, or via the vectors (capital font). The solid bold arrow represents the direct association between two viruses, whereas the dashed arrows show the confounding indirect associations. The additional conditioning explanatory variables included in our constrained models are illustrated. The indirect effects (mediator species, changes and variation in host immune response) are included in the models implicitly through spatial variables and the model structure.

Figure 2 . **Virus communities by population.** In A, each pie chart represents the virus community of a host plant population, overlayed on the map of Åland. Infection load is shown with greyscale, as indicated in the legend. Sliced portions of same shade of grey show plants with equal number of viruses, but in different combinations. B) Nestedness of the virus communities when aggregated to host population level. Two most contrasting communities are shown in more detail in C and D. In these barplots, the within-host virus community composition of a virus-rich plant host population 861 (D) and virus-poor population 3222 (C) is shown with different colours indicating different virus taxa, as shown in legend.

Figure 3. Viral co-occurrence structure. A) Virus occurrences (diagonal) and co-occurrences (off-diagonals), organised according to the first principal component of the incidence matrix. The colouring of the font intensifies with the value. The diagonal elements show the overall number of occurrences of the focal virus, whereas the off-diagonal elements show the numbers of host plants where the two viruses co-occur. The numbers inside brackets next to virus names indicate the number of single infections. B) Species-area curve and coexistence curves. The uppermost black dotted line indicates the species-area curve, the grey dashed line indicates the maximum amount of possibly coexisting pairs (calculated from the species-area-curve), and the black solid line shows the actual average coexistence curve.

Figure 4. Virus networks. A) Markov random field co-occurrence network, representing direct pairwise associations between viruses, after accounting for the rest of the community (MRF). B) Constrained Markov random field co-occurrence network, representing direct pairwise associations between viruses, after accounting for the rest of the community as well as all environmental and spatial effects (CRFfull). C) Direct associations *explained* by host-related variables (CRFhost); D) direct associations *explained* by habitat-related variables (CRFhabitat); and E) direct associations *explained* by spatial variables (CRFspat). The viruses closer in space have stronger positive associations with each other and edge thickness is scaled by the strength of association. In summary, when the associations in panels C-D are subtracted from all the associations in panel A, the associations in panel B are left. Hence, the dashed lines in C-D represent associations that disappear from panel A after adding the explanatory variable group in question, resulting in the permanent associations illustrated with solid lines in panel B.

Table 1. Explanatory conditioning variables. See Supplementary Material and Methods for more thorough description and justification of the explanatory variables.

Table 2. The direct effects of the additional explanatory variables on the virus occurrence probabilities and the indirect effects of the additional explanatory variables through their

influence on the association between a focal pair of viruses. Effects$\geq$ |0.25| are highlighted with **bold** font. All mean coefficient estimates are significant (based on bootstrapping and a 90% confidence interval).

References

Aalto, J., Pirinen, P. & Jylhä, K. (2016). New gridded daily climatology of Finland: permutation-based uncertainty estimates and temporal trends in climate. *J. Geophys. Res. Atmos.* , 121.

Alcaide, C., Rabadán, M.P., Moreno-Pérez, M.G. & Gómez, P. (2020). Implications of mixed viral infections on plant disease ecology and evolution. *Adv. Virus Res.* , 106, 145–169.

Alcaide, C., Sardanyés, J., Elena, S.F. & Gómez, P. (2021). Increasing temperature alters the within-host competition of viral strains and influences virus genetic variability. *Virus Evol.* , 2020.07.06.190173.

Almeida-Neto, M., Guimarães, P., Guimarães, P.R., Loyola, R.D. & Ulrich, W. (2008). A consistent metric for nestedness analysis in ecological systems: Reconciling concept and measurement. *Oikos* , 117, 1227–1239.

Azaele, S., Muneepeerakul, R., Rinaldo, A. & Rodriguez-Iturbe, I. (2010). Inferring plant ecosystem organization from species occurrences.*J. Theor. Biol.* , 262, 323–329.

Bergner, L.M., Orton, R.J., Benavides, J.A., Becker, D.J., Tello, C., Biek, R., *et al.* (2020). Demographic and environmental drivers of metagenomic viral diversity in vampire bats. *Mol. Ecol.* , 29, 26–39.

Blanchet, F.G., Gazelles, K. & Gravel, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecol. Lett.*

Carlsson-Granér, U. & Thrall, P.H. (2015). Host resistance and pathogen infectivity in host populations with varying connectivity.*Evolution (N. Y).* , 69, 926–938.

Chang, S., Puryear, J. & Cairney, J. (1993). A Simple and Efficient Method for Isolating RNA from Pine Trees. *Plant Mol. Biol. Report.* , 11, 113–116.

Chase, J.M. & Leibold, M.A. (2003). *Ecological Niches: Linking Classical and Contemporary Approaches* . University of Chicago Press, Chicago, IL.

Cheng, J., Levina, E., Wang, P. & Zhu, J. (2014). A sparse Ising model with covariates. *Biometrics* , 70, 943–953.

Clark, N.J., Kerry, J.T. & Fraser, C.I. (2020). Rapid winter warming could disrupt coastal marine fish community structure. *Nat. Clim. Chang.* , 2040.

Clark, N.J., Wells, K. & Lindberg, O. (2018). Unravelling changing interspecific interactions across environmental gradients using Markov random fields. *Ecology* , 99, 1277–1283.

Csardi, G. & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Syst.*

Cuellar, W.J., Kreuze, J.F., Rajamäki, M.L., Cruzado, K.R., Untiveros, M. & Valkonen, J.P.T. (2009). Elimination of antiviral defense by viral RNase III. *Proc. Natl. Acad. Sci. U. S. A.* , 106, 10354–10358.

DaPalma, T., Doonan, B.P., Trager, N.M. & Kasman, L.M. (2010). A systematic approach to virus-virus interactions. *Virus Res.* , 149, 1–9.

Díaz-Muñoz, S.L. (2017). Viral coinfection is shaped by host ecology and virus-virus interactions across diverse microbial taxa and environments.*Virus Evol.* , 3, 1–14.

Dray, S., Legendre, P. & Peres-Neto, P.R. (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecol. Modell.* , 196, 483–493.

Fukami, T. (2015). Historical contingency in community assembly: integrating niches, species pools, and priority effects. *Annu. Rev. Ecol. Evol. Syst.* , 46, 1–23.

Gilman, S.E., Urban, M.C., Tewksbury, J., Gilchrist, G.W. & Holt, R.D. (2010). A framework for community interactions under climate change. *Trends Ecol. Evol.* , 25, 325–331.

González-Jara, P., Atencio, F.A., Martínez-García, B., Barajas, D., Tenllado, F. & Díaz-Ruíz, J.R. (2005). A single amino acid mutation in the Plum pox virus helper component-proteinase gene abolishes both synergistic and RNA silencing suppression activities. *Phytopathology* , 95, 894–901.

Gorbalenyai, A.E. & Siddell, S.G. (2021). Recognizing species as a new focus of virus research. *PLoS Pathog.* , 17, 1–7.

Halliday, F.W., Rohr, J.R. & Laine, A.L. (2020). Biodiversity loss underlies the dilution effect of biodiversity. *Ecol. Lett.* , 23, 1611–1622.

Hanley, J.A. & Mcneil, B.J. (1982). The Meaning and Use of the Area under a Receiver Characteristic. *Radiology* , 143, 29–36.

Hanski, I. (1999). *Metapopulation Ecology . Oxford Ser. Ecol. Evol.* 1st edn. Oxford University Press, Oxford, UK, Oxford, UK.

Harris, D.J. (2016). Inferring species interactions from co-occurrence data with Markov networks. *Ecology* , 97, 3308–3314.

Hart, S.P., Usinowicz, J. & Levine, J.M. (2017). The spatial scales of species coexistence. *Nat. Ecol. Evol.* , 1, 1066–1073.

Hulo, C., De Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I., *et al.* (2011). ViralZone: A knowledge resource to understand virus diversity. *Nucleic Acids Res.* , 39, 576–582.

Johnson, P.T.J., de Roode, J.C. & Fenton, A. (2015). Why infectious disease research needs community ecology. *Science (80-.).* , 349, 1259504.

Jousimo, J., Tack, A.J.M., Ovaskainen, O., Mononen, T., Susi, H., Tollenaere, C., *et al.* (2014). Ecological and evolutionary effects of fragmentation on infectious disease dynamics. *Science (80-.).* , 344, 1289–1293.

Karvonen, A., Jokela, J. & Laine, A.-L. (2019). Importance of Sequence and Timing in Parasite Coinfections. *Trends Parasitol.* , 35, 109–118.

Kozanitas, M., Osmundson, T.W., Linzer, R. & Garbelotto, M. (2017). Interspecific interactions between the Sudden Oak Death pathogen Phytophthora ramorum and two sympatric Phytophthora species in varying ecological conditions. *Fungal Ecol.* , 28, 86–96.

Kreuze, J.F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I., *et al.* (2009). Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses. *Virology* , 388, 1–7.

Leathwick, J.R., Elith, J. & Hastie, T. (2006). Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecol. Modell.* , 199, 188–196.

Letko, M., Seifert, S.N., Olival, K.J., Plowright, R.K. & Munster, V.J. (2020). Bat-borne virus diversity, spillover and emergence. *Nat. Rev. Microbiol.* , 18, 461–471.

Levine, J.M. & Hart, S.P. (2020). The Dimensions of Species Coexistence. In: *Unsolved Problems in Ecology* (eds. Dobson, A., Tilman, D. & Holt, R.D.). Princeton University Press, Princeton, New Jersey, US, pp. 145–159.

Makiola, A., Dickie, I.A., Holdaway, R.J., Wood, J.R., Orwin, K.H. & Glare, T.R. (2019). Land use is a determinant of plant pathogen alpha- but not beta-diversity. *Mol. Ecol.* , 28, 3786–3798.

Malmstrom, C.M., Melcher, U. & Bosque-Perez, N.A. (2011). The expanding field of plant virus ecology: Historical foundations, knowledge gaps, and research directions. *Virus Res.* , 159, 84–94.

Massart, S., Candresse, T., Gil, J., Lacomme, C., Predajna, L., Ravnikar, M., *et al.* (2017). A framework for the evaluation of biosecurity, commercial, regulatory, and scientific impacts of plant viruses and viroids identified by NGS technologies. *Front. Microbiol.* , 8.

Munson-Mcgee, J.H., Peng, S., Dewerff, S., Stepanauskas, R., Whitaker, R.J., Weitz, J.S., *et al.* (2018). A virus or more in (nearly) every cell: Ubiquitous networks of virus-host interactions in extreme environments. *ISME J.* , 12, 1706–1714.

Ng, T.F.F., Duffy, S., Polston, J.E., Bixby, E., Vallad, G.E. & Breitbart, M. (2011). Exploring the diversity of plant DNA viruses and their satellites using vector-enabled metagenomics on whiteflies. *PLoS One* , 6.

Obrępalska-Stęplowska, A., Renaut, J., Planchon, S., Przybylska, A., Wieczorek, P., Barylski, J., *et al.* (2015). Effect of temperature on the pathogenesis, accumulation of viral and satellite RNAs and on plant proteome in peanut stunt virus and satellite RNA-infected plants. *Front. Plant Sci.* , 6, 1–14.

Ojanen, S.P., Nieminen, M., Meyke, E., Pöyry, J. & Hanski, I. (2013). Long-term metapopulation study of the Glanville fritillary butterfly (Melitaea cinxia): Survey methods, data management, and long-term population trends. *Ecol. Evol.* , 3, 3713–3737.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., *et al.* (2019). vegan: Community Ecology Package. R package version 2.5-6.

Ovaskainen, O. & Soininen, J. (2011). Making more out of sparse data: hierarchical modeling of species communities. *Ecology* , 92, 289–295.

Pirone, T.P. & Blanc, S. (1996). Helper-Dependent Vector Transmission of Plant Viruses. *Annu. Rev. Phytopathol.* , 34, 227–247.

Pleydell, D.R.J., Soubeyrand, S., Dallot, S., Labonne, G., Chadœuf, J., Jacquot, E., *et al.* (2018). Estimation of the dispersal distances of an aphid-borne virus in a patchy landscape. *PLoS Comput. Biol.* , 14, 1–24.

Popovic, G.C., Warton, D.I., Thomson, F.J., Hui, F.K.C. & Moles, A.T. (2019). Untangling direct species associations from indirect mediator species effects with graphical models. *Methods Ecol. Evol.* , 10, 1571–1583.

Pruss, G., Ge, X., Shi, X.M., Carrington, J.C. & Vance, V.B. (1997). Plant viral synergism: The potyviral genome encodes a broad-range pathogenicity enhancer that transactivates replication of heterologous viruses. *Plant Cell* , 9, 859–868.

QGIS Development Team. (2019). QGIS Geographic Information System.

R Core Team. (2020). R: A language and environment for statistical computing.

Rey, M.E.C., Ndunguru, J., Berrie, L.C., Paximadis, M., Berry, S., Cossa, N., *et al.* (2012). Diversity of dicotyledenous-infecting geminiviruses and their associated DNA molecules in Southern Africa, including the South-west Indian Ocean Islands. *Viruses* , 4, 1753–1791.

Roossinck, M.J. (2010). Lifestyles of plant viruses. *Philos. Trans. R. Soc. B Biol. Sci.* , 365, 1899–1905.

Roossinck, M.J. (2011). The big unknown: Plant virus biodiversity. *Curr. Opin. Virol.* , 1, 63–67.

Roossinck, M.J. (2012). Plant Virus Metagenomics: Biodiversity and Ecology. *Annu. Rev. Genet.* , 46, 359–369.

Roux, S., Hallam, S.J., Woyke, T. & Sullivan, M.B. (2015). Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *Elife* , 4, 1–20.

Al Rwahnih, M., Daubert, S., Golino, D. & Rowhani, A. (2009). Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. *Virology* , 387, 395–401.

Sallinen, S., Norberg, A., Susi, H. & Laine, A.L. (2020). Intraspecific host variation plays a key role in virus community assembly. *Nat. Commun.* , 11, 1–11.

Schröder, M.L., Glinwood, R., Ignell, R. & Krüger, K. (2017). The role of visual and olfactory plant cues in aphid behaviour and the development of non-persistent virus management strategies. *Arthropod. Plant. Interact.* , 11, 1–13.

Seabloom, E.W., Borer, E.T., Gross, K., Kendig, A.E., Lacroix, C., Mitchell, C.E., *et al.* (2015). The community ecology of pathogens: Coinfection, coexistence and community composition. *Ecol. Lett.* , 18, 401–415.

Seabloom, E.W., Borer, E.T., Mitchell, C.E. & Power, A.G. (2010). Viral diversity and prevalence gradients in North American Pacific Coast grasslands. *Ecology* , 91, 721–732.

Shannon, C.E. (1948). A Mathematical Theory of Communication. *Bell Syst. Tech. J.* , 27, 623–656.

Six, E.W. & Klug, C.A.C. (1973). Bacteriophage P4: a satellite virus depending on a helper such as prophage P2. *Virology* , 51, 327–344.

Snyder, R.E. & Chesson, P. (2003). Local dispersal can facilitate coexistence in the presence of permanent spatial heterogeneity. *Ecol. Lett.* , 6, 301–309.

Stone, L. & Roberts, A. (1990). The Checkerboard Score and Species Distributions. *Oecologia* , 85, 74–79.

Susi, H. & Laine, A. (2020). Agricultural land use disrupts biodiversity mediation of virus infections in wild plant populations. *New Phytol.* , 230, 2447-.

Sutton, C. & McCallum, A. (2011). An introduction to conditional random fields. *Found. Trends Mach. Learn.* , 4, 267–373.

Suzuki, N., Rivero, R.M., Shulaev, V., Blumwald, E. & Mittler, R. (2014). Abiotic and biotic stress combinations. *New Phytol.* , 203, 32–43.

Syller, J. (2012). Facilitative and antagonistic interactions between plant viruses in mixed infections. *Mol. Plant Pathol.* , 13, 204–216.

Székely, A.J. & Langenheder, S. (2014). The importance of species sorting differs between habitat generalists and specialists in bacterial communities. *FEMS Microbiol. Ecol.* , 87, 102–112.

Taiwo, M.A., Kareem, K.T., Nsa, I.Y. & D'A Hughes, J. (2007). Cowpea viruses: Effect of single and mixed infections on symptomatology and virus concentration. *Virol. J.* , 4, 1–5.

Tilman, D. (2004). Niche tradeoffs, neutrality, and community structure: a stochastic theory of resource competition, invasion, and community assembly. *Proc. Natl. Acad. Sci. U. S. A.* , 101, 10854–61.

Tugume, A.K., Mukasa, S.B. & Valkonen, J.P.T. (2016). Mixed infections of four viruses, the incidence and phylogenetic relationships of Sweet potato chlorotic fleck virus (Betaflexiviridae) isolates in wild species and sweetpotatoes in Uganda and evidence of distinct isolates in East Africa. *PLoS One* , 11, 1–27.

Waterhouse, P.M. & Murant, A.F. (1983). Further evidence on the nature of the dependence of carrot mottle virus on carrot red leaf virus for transmission by aphids. *Ann. Appl. Biol.* , 103, 455–464.

Wintermantel, W.M., Cortez, A.A., Anchieta, A.G., Gulati-Sakhuja, A. & Hladky, L.L. (2008). Co-infection by two criniviruses alters accumulation of each virus in a host-specific manner and influences efficiency of virus transmission. *Phytopathology* , 98, 1340–1345.

Wren, J.D., Roossinck, M.J., Nelson, R.S., Scheets, K., Palmer, M.W. & Melcher, U. (2006). Plant virus biodiversity and ecology. *PLoS Biol.* , 4, 0314–0315.

Zheng, Y., Gao, S., Padmanabhan, C., Li, R., Galvez, M., Gutierrez, D.,*et al.* (2017). VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virology* , 500, 130–138.

Hosted file

`tables.docx` available at https://authorea.com/users/732366/articles/710782-direct-and-indirect-viral-associations-predict-coexistence-in-wild-plant-virus-communities