UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA AGRARIA, ALIMENTARIA Y DE BIOSISTEMAS

Departamento de Biotecnología-Biología Vegetal

Centro de Biotecnología y Genómica de Plantas

(UPM-INIA)

Characterisation of the virome of an agricultural ecosystem to understand virus ecology and evolution in a heterogeneous environment

Ph.D. Thesis:

ADRIÁN PELÁEZ LADERAS

Graduado en Biología, Máster en Virología, becario FPI

2021

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA AGRARIA, ALIMENTARIA Y DE BIOSISTEMAS

Departamento de Biotecnología-Biología Vegetal

Centro de Biotecnología y Genómica de Plantas

(UPM-INIA)

Ph.D. thesis:

Characterisation of the virome of an agricultural ecosystem to understand virus ecology and evolution in a heterogeneous environment

Author:

Adrián Peláez Laderas

Graduado en Biología, Máster en Virología

Directors:

Fernando García-Arenal Rodríguez

Dr. Ingeniero Agrónomo

Michael J. McLeish

Dr. in Molecular Ecology and Evolution

Madrid, 2021

Acknowledgements

Esta Tesis Doctoral se ha realizado en el laboratorio de Interacción y Coevolución Planta-Virus bajo la dirección y supervisión del jefe de grupo Prof. Fernando García-Arenal a quien me gustaría transmitirle mi más sincero agradecimiento por el continuo apoyo a mi estudio e investigación de doctorado, por su paciencia y su inspirador liderazgo, durante el transcurso de esta tesis. Su inestimable asesoramiento me ayudó en todo el tiempo de investigación y redacción de esta tesis. Sus críticas constructivas y consejos me permitieron ampliar los horizontes de mis conocimientos, sentando las bases de mi trabajo de investigación, desarrollando mi pensamiento crítico y alimentando mi entusiasmo por estudiar la evolución de los patógenos.

También le agradezco a Michael J. McLeish, codirector de esta tesis. Agradezco enormemente su ayuda prestada, su asistencia en los análisis bioinformáticos y bioestadísticos y en la escritura de esta tesis. También me convenció de usar R, contagiándome su pasión por la programación, la cual me ayudó a llevar a cabo esta tesis y a decidir mi futuro en este campo.

No podría haber imaginado tener mejores directores y mentores para mi estudio de doctorado.

Me gustaría agradecer a las Prof. Aurora Fraile Pérez y a María Ángeles Ayllón por sus utilísimas sugerencias, optimismo, incesante estímulo, ayudarme y enseñarme las técnicas clásicas en el campo de la virología vegetal. Aprecio que siempre hayan hecho un esfuerzo adicional por mí en momentos de necesidad.

También quiero reconocer la labor de la Prof. Rosario G. Gavilán con su ayuda en la identificación taxonómica y al Prof. Miguel A. Aranda en la prestación del ARN purificacido del aislado de WMV.

Agradezco a Miguel y Antolín su excelente asistencia técnica y ayudarme a extraer los miles de plantas. Gracias a su ayuda la carga de trabajo se redujo mucho y me permitió invertir tiempo en hacer más análisis estadísticos.

Esta Tesis Doctoral no habría sido posible sin la colaboración y el apoyo de los miembros (pasados o actuales) de nuestro laboratorio, así que un sincero agradecimiento a Israel, Soledad, Livia, Sayanta, Ayushi, Ricky, Manuel Guillermo, Cristina, Viji, Marisa, Laura, Ayo, Carlos, Bhumika, Fran, María Miriam, Bisola, Adrián Danut, Rafael, Cristian, Nico y Ana que mantuvieron un ambiente alegre pero científico en el laboratorio.

Un agradecimiento muy especial a Alberto por ser un amigo, compañero de laboratorio y de piso. Su ayuda y actitud a la hora de prepararme las reuniones de laboratorio y charlas me ayudaron a llevar mejor la presión.

Estoy tremendamente agradecido a Estefanía Conde por ser una fuente constante de apoyo, optimismo durante el transcurso de mi tesis. No hace falta decir cuánto me ha ayudado, pudiendo contar con ella para tomar un respiro (muy necesarios), cuando los experimentos o la escritura no iban según los planes. Le estoy muy agradecido por mostrarme siempre la luz al final del túnel y contagiarme su jovial perspectiva de la vida.

Estoy agradecido a mis amigos de Alcázar de San Juan y Jaén, que me enviaron buenos deseos y suerte siempre que hablé con ellos, y me motivaron para alcanzar mis objetivos.

Estoy en deuda con mis padres, hermano y familia de Madrid, que en todo momento me apoyaron y creyeron en mí. Sin ellos no habría sido posible atravesar los momentos difíciles.

Table of Contents

List of Abbreviations ..... **XI**

List of Acronyms and Full Names of Virus Species ..... **XIII**

Summary ..... **XV**

Resumen ..... **XIX**

1. INTRODUCTION **1**

1.1. The importance of emerging plant diseases **3**

1.2. The process of virus emergence **4**

1.3. Evolutionary factors in virus emergence **8**

1.3.1. The genetics of host range evolution 8

1.3.2. Fitness trade-offs in host range evolution 10

1.3.3. Causes of across-host fitness trade-offs 12

1.4. Ecological factors in virus emergence **14**

1.4.1. The relationship between ecosystem anthropisation and disease risk 14

1.4.2. The relationship between host range and host community composition 17

1.5. Ecological diversity in central Spain **20**

1.5.1. Oakwood habitat 21

1.5.2. Wasteland habitat 22

1.5.3. Edge and Crop habitats 22

1.6. Virus detection at the landscape scale **23**

1.7. *Watermelon mosaic virus* **27**

1.7.1. Host range, ecology, and epidemiology 31

2. OBJECTIVES **35**

3. MATERIAL AND METHODS **39**

3.1. Field work and sample collection **41**

3.1.1. Identification of sampling sites 41
3.1.2. Sampling design and sample preservation 44
3.2. Characterisation and comparison of plant communities 45
3.2.1. Characterisation of plant communities 45
3.2.2. Assessment of diversity estimators 45
3.3. RNA extraction and nucleotide sequence determination 46
3.3.1. Total RNA extraction from plant tissues 46
3.3.2. Rationalisation of plant species for RNA extraction and HTS analysis 49
3.3.3. Library preparation and high throughput sequencing 50
3.3.4. Nucleotide sequence determination by Sanger method 51
3.4. Sequence detection and validation 52
3.4.1. Trimming and storing of HTS libraries 52
3.4.2. Detection of viral sequences in the HTS libraries 52
3.4.3. Validation of virus sequence detections by HTS 53
3.4.4. Evaluation of the quality of HTS libraries 54
3.4.5. Virus detection in individual RNA extracts by RT-PCR 55
3.5. Computational and statistical analyses 57
3.5.1. Host range, incidence, and population genetics analyses 57
3.5.2. Phylogenetic analyses 60
3.5.3. Virus richness relationships with habitat and host 61
3.5.4. Statistical significance tests 63
4. RESULTS 65
4.1. Sampling collection and sample effort assessment 67
4.2. Characterisation of plant communities 69
4.2.1. Differentiation of plant communities 69
4.2.2. Comparison of plant species diversity among habitats 71
4.2.3. Rationalisation of plant species selection for HTS 74

4.3.Sample preparation and obtention of nucleotide sequences 79

4.3.1. Optimisation of RNA preparations..... 79

4.4. Detection and validation of viral OTUs 84

4.4.1. Quality control of plant RNA extracts for detection of viral sequences..... 84

4.4.2. Validation of viral OTUs 85

4.4.3. Viral OTUs detected in HTS libraries 87

4.4.4. Viral OTUs distributions over hosts and host range 93

4.5. A plant virus in a heterogeneous environment 106

4.5.1. WMV host range and incidence 107

4.5.2. Analysis of WMV genetic diversity 110

4.5.3. Analysis of WMV population structure 112

4.5.4. Phylogenetic analysis of WMV..... 113

5. DISCUSSION 119

5.1. Characterisation of sampled habitats in an agricultural ecosystem in central Spain..... 123

5.2. Assessment of biases in viral detection 125

5.3. Characterisation of the virome 128

5.4. Diversification of WMV at the landscape scale 131

6. CONCLUSIONS 135

7. BIBLIOGRAPHY 139

8. ANNEXES 179

List of Abbreviations

AI:	Association Index
AIC:	Akaike's Information Criterion
bp:	Base pair
cDNA:	Complementary DNA
CI:	Cylindrical inclusion Protein
CP:	Coat Protein
CTAB:	Cetyltrimethylammonium bromide
D_{AE}**:**	Asymptotic estimator of Shannon diversity
DCA:	Detrended correspondence analysis
DEPC:	Diethyl pyrocarbonate DIECA
DIECA:	Sodium diethyldi-thiocarbamate
d_N*:*	Non-synonymous nucleotide substitution
DNA:	Deoxyribonucleic acid
d_S*:*	Synonymous nucleotide substitution
EtBr:	Ethidium bromide
GFG model:	Gene-for-gene model
H:	Number of haplotypes
HC-Pro:	Helper Component Protease
Hd:	Haplotype diversity
HTS:	High-Throughput Sequencing
MA model:	Matching-alleles model
MUSCLE:	Multiple Sequence Comparison by Log- Expectation
NIa:	Small Nuclear Inclusion Protein
NIb:	Large Nuclear Inclusion Protein
P1:	First protein
P3:	Third protein
RdRp:	RNA-dependent RNA polymerase
RNA:	Ribonucleic acid
rRNA:	Ribosomal RNA

RT-PCR:	Reverse Transcription Polymerase Chain Reaction
***Sr*:**	Species richness
SDS:	Sodium dodecyl sulfate
S_q:	Tsallis entropy estimator of diversity
(+)ssRNA	Positive sense single-stranded RNA
(-)ssRNA	Negative sense single-stranded RNA
TAE:	Tris-acetate-Ethylenediaminetetraacetic acid
TYLCD:	Tomato yellow leaf curl disease
VPg:	Viral Protein genome-linked

List of Acronyms and Full Names of Virus Species

CGMMV:	*Cucumber green mottle mosaic virus*
HCRSV:	*Hibiscus chlorotic ringspot virus*
HCSV:	*Hibiscus chlorotic spot virus*
PMMoV:	*Pepper mild mottle virus*
PFBV:	*Pelargonium flower break virus*
PPV:	*Plum pox virus*
PVY:	*Potato virus Y*
RYMV:	*Rice yellow mottle virus*
SMV:	*Soybean mosaic virus*
TMGMV:	*Tobacco mild green mosaic virus*
ToMV:	*Tomato mosaic virus*
TuMV:	*Turnip mosaic Virus*
TYLCV:	*Tomato yellow leaf curl virus*
WMV:	*Watermelon mosaic virus*

Summary

The socio-economic relevance of emergent viruses has driven a considerable research effort aimed at understanding the process of emergence. Virus emergence depends on a range of factors intrinsic to the virus, such as genetic traits that determine its fitness in different hosts (i.e., evolutionary factors), or extrinsic factors related to ecology and epidemiology that ultimately result in the virus encountering and infecting a new host population (i.e., ecological factors). Therefore, mechanisms underlying emergence involve the evolution of virus host range, which has received considerable research attention in the past. However, until recently, the focus on factors intrinsic to the virus has been at the expense of understanding ecological factors that affect host-range evolution. Previous work has highlighted the importance of knowing the host range of viruses, to understand how host and virus ecology influences the epidemiology of the virus, and thus the risk of infection. Nevertheless, it is unusual to find literature from studies that have characterised the host range of a virus in a given ecosystem. At the spatial scale of the landscape, the influence that environmental heterogeneity has on host and virus diversity and on the large numbers of plant-virus interactions, requires minimising biases in the quantification and comparison of communities. New high-throughput sequencing (HTS) approaches in combination with computational tools developed for big-data, are now the gold standard for addressing landscape scale dynamics and the joint study of the ecological and evolutionary factors that underlie virus emergence.

The goal of this thesis was to identify ecological factors that determine plant-virus interactions in the heterogeneous communities of an agricultural ecosystem. The ultimate ambition of determining the ecological factors that drive transmission dynamics and emergence at the landscape scale, exceeds the scope of a doctoral thesis. Nevertheless, the goal of the thesis sets the basis for reaching that ambitious target through fulfilling a series of objectives; the first steps towards understanding virus emergence. The first objective, to evaluate plant community variation of an agricultural ecosystem in central Spain, was performed by comparing four key habitats subject to increasing levels of human intervention. The loss of biodiversity in agricultural ecosystems is expected to change epidemiological dynamics of plant viruses. Therefore, variation in plant

communities from each of four habitat types, Oakwood, Wasteland, Edge and Crop, was evaluated to investigate, in the second objective, plant community structure-function relationships with virus diversity. To this aim, the second objective was to characterise the virome of plant communities in the four habitats using a HTS approach, to examine the contribution of ecological factors in structuring single-stranded RNA (ssRNA) virus communities. The third objective focused on watermelon mosaic virus (WMV), a widespread large host range generalist virus, where HTS was combined with RT-PCR amplification and Sanger sequencing to assess the effect of the four habitats on host range, incidence, and genetic diversity.

The results showed plant assemblages of each of the habitats clustered into discrete categories, with Crop and Edge the most similar, and Oakwood the most differentiated plant community. The diversity of the plant assemblages differed among the habitats, Crop with the lowest diversity and Wasteland with the highest. Plant diversity was affected by seasonality. Apart from Crop with an opposite trend, the plant diversity of each habitat decreased between spring and autumn. Of the 118 total plant species that were sequenced, 106 of them were validated as hosts of viruses. Edge had the highest host richness with 52 species, and Crop the lowest with 28 species. Oakwood and Wasteland had 45 and 41 host species respectively. The highest virus (operational taxonomic unit; OTU) richness was detected in Edge followed by Wasteland. Oakwood and Crop, which were polar opposites in the level of disturbance, supported less species than the other two habitats. The distribution of viruses among the habitats was associated strongly with host species abundance only within Edge, but in general, host use was not driven by the abundance of any given plant species. A large proportion (62 %) of the 90 OTUs detected had host ranges of less than five species, and a small proportion (19%) had host ranges of 20 or more species. The host range of a given virus differed significantly among the habitats, with Edge producing the highest realised mean host ranges. The effect of habitat and host ecology on plant viruses was also evident in the structuring of WMV incidence, genetic diversity, and host range. Although sensitivity in detection by HTS was greater than by RT-PCR, both host range estimates of 43 and 24 species respectively, showed WMV infected species occurring in each habitat. The incidence of WMV

was higher in Crop compared to Edge and Wasteland, but equivocal to that in Oakwood.

Incidence was correlated with host species relative abundance in Crop, but not in the other habitats, with no dependency between seasonality and habitat in structuring incidence. The genetic diversity of WMV was not significantly structured according to host species or habitat, but rather between *Cucumis melo* and wild host species. Both phylogenetic inferences and haplotype networks showed 4 main clades of WMV, each with mixed host associations, and with Oakwood supporting the only habitat-specific clade. Overall, habitat associations with the coat protein gene of WMV were stronger than host species associations, but both had an important role in differentiation between and within WMV populations.

This thesis demonstrated that viral detection by HTS is robust to quantifying variation at higher levels of organisation among habitats. Differences in community composition associated with each habitat at any given time, resulted in a subset of possible biotic interactions within each of them, and contributed to structuring plant-virus and virus-virus interactions at the ecosystem scale. Ecological differentiation among habitats was shown in the distinctions between their plant and virus community compositions, and the responses viruses had to host plant ecology. For instance, Edge was conducive to infections by many viruses, and supported host species that were infected on average by more viruses with wider realised host ranges compared to the other habitats, which revealed important ecological features of a reservoir community. Ecological interactions were also partly contingent on seasonality and the level of disturbance. The infection by viruses of a large majority of the plant samples suggests the association of symbionts with most host species is frequent. Thus, ecological factors affect transmission between and within plant communities of agricultural ecosystems, and influence the processes of host range evolution and, hence virus emergence.

Resumen

La relevancia socioeconómica de los virus emergentes ha impulsado un esfuerzo considerable en la investigación destinada a comprender el proceso de emergencia. La emergencia de los virus depende de una serie de factores intrínsecos al virus, como los rasgos genéticos que determinan su adaptación a diferentes huéspedes (es decir, factores evolutivos), o factores extrínsecos relacionados con la ecología y la epidemiología que, en última instancia, hacen que el virus encuentre e infecte a una nueva población de huéspedes (es decir, factores ecológicos). Por lo tanto, los mecanismos que subyacen a la emergencia implican la evolución de la gama de huéspedes del virus, tema que ha recibido una considerable atención en el pasado. Sin embargo, hasta hace poco el énfasis en el estudio de los factores intrínsecos del virus ha sido en detrimento de la comprensión de los factores ecológicos que afectan a la evolución de la gama de huéspedes. Los trabajos publicados han puesto de relieve la importancia de conocer la gama de huéspedes de los virus para entender cómo la ecología del huésped y del virus influye en la epidemiología del virus y, por tanto, en el riesgo de infección. Sin embargo, hay pocos estudios que hayan caracterizado la gama de huéspedes de un virus en un ecosistema determinado. A escala espacial del paisaje, la influencia que la heterogeneidad ambiental ejerce sobre la diversidad de huéspedes y virus y sobre el gran número de interacciones planta-virus, exige minimizar los sesgos en la cuantificación y comparación de las comunidades. Los nuevos enfoques de secuenciación de alto rendimiento (HTS) en combinación con las herramientas computacionales desarrolladas para big-data, son las mejores herramientas para abordar la dinámica a escala del paisaje y el estudio conjunto de los factores ecológicos y evolutivos que subyacen a la emergencia viral. El objetivo de esta tesis es identificar los factores ecológicos que determinan las interacciones planta-virus en las comunidades heterogéneas de un ecosistema agrícola. Determinar los factores ecológicos que afectan a la dinámica de transmisión y la emergencia a escala del paisaje excede del alcance de una tesis doctoral. Sin embargo, el trabajo realizado en esta tesis sienta las bases para alcanzar esa ambiciosa meta mediante el cumplimiento de una serie de objetivos como primeros pasos hacia la comprensión de la emergencia de los virus.

El primer objetivo ha sido evaluar la variación de la comunidad vegetal de un ecosistema agrícola del centro de España, y se llevó a cabo comparando cuatro hábitats clave sujetos a niveles crecientes de intervención humana. Se espera que la pérdida de biodiversidad en los ecosistemas agrícolas cambie la dinámica epidemiológica de los virus de plantas. Por lo tanto, se evaluó la variación de las comunidades vegetales de cada uno de los cuatro tipos de hábitat, Encinar, Erial, Linde y Cultivo, para investigar, en el segundo objetivo, las relaciones entre estructura y función de la comunidad vegetal y la diversidad de virus. Para ello, el segundo objetivo ha sido caracterizar el viroma de las comunidades vegetales de los cuatro hábitats mediante HTS, para examinar la contribución de los factores ecológicos en la estructuración de las comunidades de virus con genomas de RNA monocatenario (ssRNA). El tercer objetivo se centró en el virus del mosaico de la sandía (watermelon mosaic virus, WMV), un virus generalista con una amplia gama de huéspedes, en el que se combinó la HTS con la amplificación por RT-PCR y la secuenciación Sanger para evaluar el efecto de los cuatro hábitats en la gama de huéspedes, la incidencia y la diversidad genética.

Los resultados mostraron que las comunidades de plantas de cada uno de los hábitats se agruparon en categorías discretas, siendo las del Cultivo y la Linde las más similares, y el Encinar la comunidad vegetal más diferenciada. La diversidad de las comunidades de plantas difiere entre los hábitats, siendo el Cultivo el de menor y el Erial el de mayor diversidad. La diversidad de plantas se vio afectada por la estacionalidad. A excepción del Cultivo, que muestra una tendencia opuesta, en cada hábitat la diversidad de plantas disminuyó entre la primavera y el otoño. De las 118 especies vegetales que se secuenciaron, 106 fueron validadas como huéspedes de virus. La Linde tuvo la mayor riqueza de huéspedes, con 52 especies, y el Cultivo la menor, con 28. Los Encinares y los Eriales tenían 45 y 41 especies de huéspedes, respectivamente. La mayor riqueza de virus (unidad taxonómica operativa; OTU) se detectó en la Linde, seguido del Erial. El Encinar y el Cultivo, que representan extremos opuestos en cuanto al nivel de perturbación, albergaban menos especies de virus que los otros dos hábitats. La distribución de los virus en los hábitats sólo se asoció con la abundancia de especies de huéspedes en la Linde, pero en general, el uso de

los huéspedes no se relaciona con la abundancia de ninguna especie de planta determinada. Una gran proporción (62%) de los 90 OTUs de virus detectados tenía una gama de huéspedes de menos de cinco especies, y una proporción mucho menor (19%) tenía una gama de huéspedes de 20 o más especies. La gama de huéspedes de un virus determinado difería significativamente entre los hábitats, siendo la Linde el que presentaba de media una mayor gama de huéspedes. El efecto de la ecología del hábitat y del huésped sobre los virus de plantas también fue evidente en la estructuración de la incidencia, la diversidad genética y la gama de huéspedes de WMV. Aunque la sensibilidad de detección por HTS fue mayor que por RT-PCR, ambas estimas de la gama de huéspedes, de 43 y 24 especies respectivamente, mostraron especies infectadas por WMV en cada hábitat. La incidencia de WMV fue mayor en el Cultivo que en la Linde y el Erial, pero similar a la del Encinar. La incidencia se correlacionó con la abundancia relativa de las especies de huéspedes en los Cultivos, pero no en los otros hábitats, y no se encontró dependencia entre la estacionalidad y el hábitat en la estructuración de la incidencia. La diversidad genética de WMV no se estructuró de forma significativa en función de la especie de huésped o de hábitat, sino más bien entre *Cucumis melo* y las especies de huéspedes silvestres. Tanto las inferencias filogenéticas como las redes de haplotipos mostraron 4 clados principales de WMV, y cada uno de ellos incluye haplotipos que infectan a distintos huéspedes. Sólo el Encinar aloja un clado específico de hábitat. En general, la diversidad del gen de la proteína de la cápsida de WMV se asocia más con el hábitat que con la especie de huésped, pero ambos factores tienen un papel importante en la diferenciación entre y dentro de las poblaciones del WMV.

Los resultados de esta tesis demuestran que la detección de virus mediante HTS es robusta en cuanto a cuantificar la variación a niveles superiores de organización, es decir, entre hábitats. Las diferencias en la composición de las comunidades asociadas con cada hábitat en un momento dado dieron lugar a un subconjunto de posibles interacciones planta-virus en cada uno de ellos, y contribuyeron a estructurar las interacciones planta-virus y virus-virus a escala del ecosistema. La diferenciación ecológica entre los hábitats se puso de manifiesto por la distinta composición de sus comunidades de plantas y de virus,

y por las respuestas de los virus a la ecología de las plantas huésped. Por ejemplo, la Linde favorece las infecciones por muchos virus, y contiene huéspedes que de media se infectan por más virus con gama de huéspedes más amplia que en los otros hábitats. Estas características ecológicas de la Linde son importantes, y son propias de una comunidad reservorio. Las interacciones ecológicas también dependen de la estacionalidad y del nivel de perturbación del hábitat. El hecho de que una gran mayoría de las especies de plantas esté infectada por virus indica una alta frecuencia de asociaciones simbióticas. En conclusión, los resultados demuestran que los factores ecológicos afectan a la transmisión de virus entre y dentro de las distintas comunidades vegetales de los ecosistemas agrícolas, e influyen en la evolución de la gama de huéspedes y, por tanto, en la emergencia viral.

1. INTRODUCTION

1.1. The importance of emerging plant diseases

The intensification of agriculture in the second half of the 20th century has largely been a response to the need to feed a growing human population. However, agriculture in the 21st century faces multiple challenges. The rapid pace at which the world's population is growing implies a significant increase in demand for food (FAO, 2009). The increase in global demand poses a huge challenge to food production and to sustaining ecosystems (Tilman *et al.,* 2002). Plant diseases are important ecological agents that modulate the composition of ecological communities, and the services ecosystems provide (Cooper & Jones, 2006; Fraile & Garcia-Arenal, 2010; Jones, 2009; Malmstrom *et al.,* 2011). A major challenge is to reduce the impact of plant diseases in agricultural production without compromising environmental integrity (i.e., biophysical processes that support plant and animal life in the environment (Tilman *et al.,* 2002)).

The social and economic effects of crop losses caused by plant diseases have been well documented. Plant diseases are responsible for a 13-16% yearly reduction in global crop yields, being a serious threat to food security (Oerke, 2006; Savary *et al.,* 2006). These losses are due to disease effects that range from systemic changes in growth and development in individual plants, to the complete devastation of entire crops (Savary *et al.,* 2006). The losses due to plant diseases are often greatest for emerging diseases. Emerging infectious diseases are defined as those that have appeared in a new host population or have previously existed in a host population but are rapidly increasing in incidence or geographic range as a result of long-term changes in its underlying epidemiology (McArthur *et al.,* 2019; Vurro *et al.,* 2010; Woolhouse 2002, Woolhouse *et al.,* 2005). The negative impact of emerging diseases is greatest since they affect host populations that have not been previously challenged by the causal agent (Vurro *et al.,* 2010; Woolhouse *et al.,* 2005), implying the absence of evolved constraints on susceptibility and pathogenicity. This could be accompanied by altered pathogenesis resulting in increased disease severity, allowing outbreaks of diseases of high impact and unusual severity (Anderson *et al*., 2004).

Most plant diseases in crops are caused by viruses and fungi (Waterworth & Hadidi, 1998), viruses being the second most important group of plant pathogens (Agrios, 2005; García-Arenal & McDonald, 2003) after fungi (Hsu, 2002). In fact, almost half of the emerging diseases of plants reported in the last four decades were caused by viruses (Elena *et al*., 2014). For example, *Cucumber green mottle mosaic virus* (CGMMV), which was described in 1935 in England (Ainsworth, 1935), has recently become one of the most economically important cucurbit pathogens worldwide (Crespo *et al.,* 2017), as it has expanded its distribution to infect various cucurbit crops in different regions (Ellouze *et al.,* 2020). Another example is tomato yellow leaf curl disease, caused by a cluster of closely related virus species including *Tomato yellow leaf curl virus (*TYLCV), which belongs to the genus *Begomovirus* of the *Geminiviridae* family. Phylogenetic analysis indicated that the virus arose in tomato crops in the Middle East, most likely between 1930 and 1950 from where it spread worldwide since the late 1980s and 1990s (García-Arenal & Zerbini 2019). TYLCV has been primarily associated with tomato, however it might also infect several alternative hosts in a sporadic way (García-Arenal & Zerbini 2019). The widespread occurrence of TYLCV and its tremendous rate of spread to new regions, makes it an important pathogen with a huge economic importance for tomato production (Pérez-Padilla *et al.,* 2020).

Thus, emerging plant diseases have an enormous socio-economic relevance that has motivated a considerable research effort to understand the emergence process (Elena *et al*., 2014), with the aim of being able to predict new emergences and to develop new strategies to decrease their impact (Johnson *et al*., 2015; Ostfeld *et al.,* 2012).

1.2. The process of virus emergence

Plant virus emergence is a complex, incompletely understood process involving multiple ecological and evolutionary factors that may be considered as acting sequentially (Elena *et al.,* 2014; Jones, 2009; Roossinck & García-Arenal.,

2015). It is broadly accepted that virus emergence involves three phases **(Figure 1.1)**: (I) in the first phase the virus contacts the new host populations. New virus-host contacts are favoured by changes in the ecology of a host, a virus or a vector, so that pre-existing genetic variants of the virus that are well established in a certain host population (reservoir host) are transmitted to a new geographical area or new host species due to contact between reservoir host populations and new hosts (focal host); (II) in the second phase the virus must adapt to the new host or geographic area to ensure productive infections (i.e., the virus must multiply in hosts of the new population to levels that would allow transmission to the new host) and therefore allow effective transmission within the new host population or environment (including the host population as a key component of a parasite's environment (Elena *et al.,* 2014; Roossinck & García-Arenal., 2015)); and (III) last, the virus must develop effective transmission mechanisms in the new host population to ensure optimal transmission (Holmes, 2009; Hudson *et al.,* 2008).

Changes to species distributions and abundances (i.e., ecological) play an important role in the emergence process, which may favour novel virus-to-host encounters that takes place in the first phase (McLeish *et al.,* 2018). For instance, despite the scarcity of empirical and experimental evidence, it is broadly accepted that virus emergence is facilitated by ecosystem simplification, defined as the loss of biodiversity and increases in managed croplands (Roossinck & García-Arenal, 2015). Ecosystem simplification associated with agricultural practices may result in reductions of species richness (i.e., the number of species), which affects species interactions. Ecological changes also affect virus-vector interactions, which may lead to new epidemiological dynamics (Islam *et al.,* 2020). For example, it is hypothesised that ecosystem simplification will affect vector populations and their vagility, which may lead to increased virus infection rates (Fraile & García-Arenal, 2016). In addition to ecological factors, virus evolution also plays an important role in the emergence process. The ability of a virus to

adapt to a new host, is relevant to the second and third stages of the emergence process (Elena *et al.,* 2014).

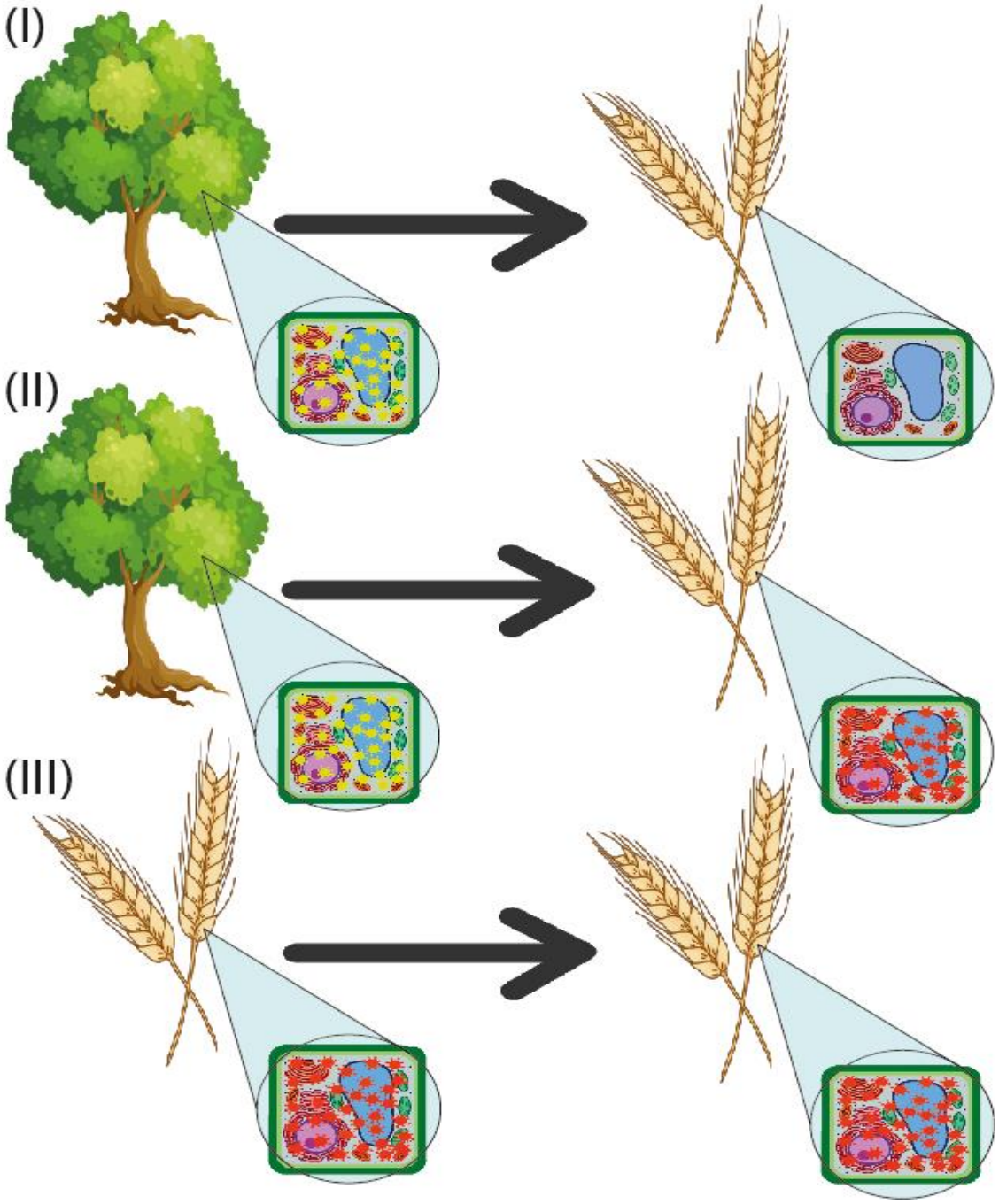

Figure 1.1. Graphical representation of the emergence process. In the first phase, the virus (yellow) present in the reservoir host (tree) comes into contact with populations of the focal host (barley). In the second phase, the virus has adapted (red) to the new host and is able to multiply in the focal host to levels that allow transmission to the new host (i.e.: productive infections). In the third phase, the virus has developed effective transmission mechanisms in the new host, ensuring optimal transmission within the new host population.

In line with the description of the phases of emergence given above, viruses that emerge in a new host necessarily have an origin in another host or reservoir, often a wild host (Roossinck & García-Arenal, 2015). However, there is evidence that crops have a role as reservoirs for emergence in wild plants. The first encounter that results in infections of the new host is called spillover. Spillover of the virus from its natural host to its new host (wild or crop) may result in viral adaptation to the new host, which results in infection dynamics becoming independent of its natural host. Eventually the virus may spillback to the natural host (Jones, 2020) **(Figure 1.2.).**

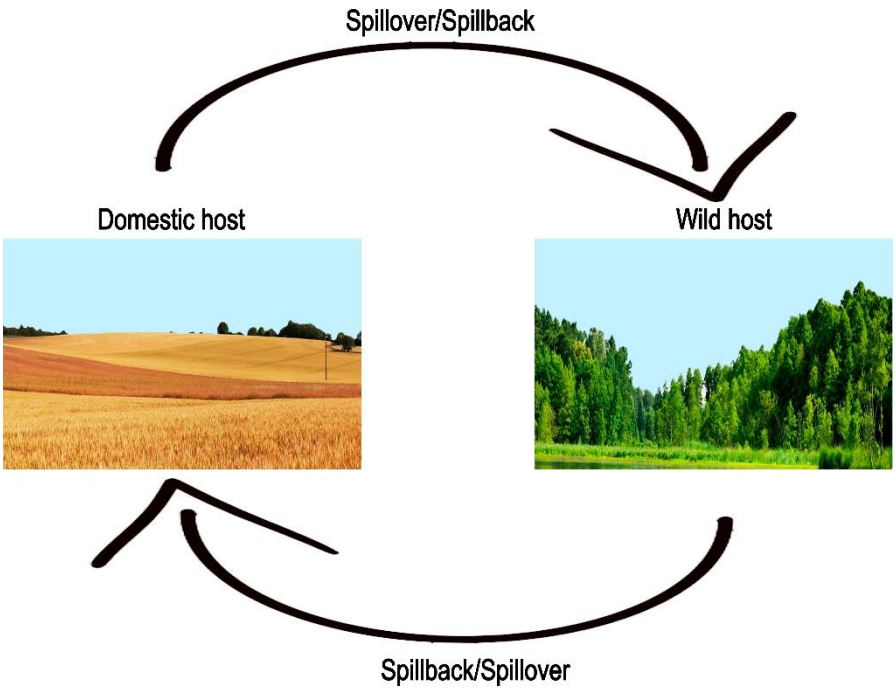

Figure 1.2. Diagram of virus transmission from the natural host (domestic or wild) to the new host (Spillover), as well as virus transmission from the new host to the natural host (Spill back) (Modified from Roossinck & García-Arenal, 2015).

Emergence often results in changes to a virus' host range; defined as the number of species in which it can reproduce (Haydon *et al.,* 2002). Host range evolution may result in an exchange of one host for another (host switch), or the acquisition of an additional host (host range expansion) (Elena *et al.,* 2014).

Emerging viruses are generally multi-host pathogens (Woolhouse, 2002) that spread to a new host population from epidemiologically connected reservoirs in which the pathogen is permanently maintained (Cronin *et al.,* 2010; Roche *et al.,* 2013). Reservoirs of plant viruses are expected to range from single-host populations to multi-species communities of mixed maintenance and non-maintenance populations (Viana *et al.,* 2014).

The virus host range is determined by factors intrinsic to the virus such as genetic traits that determine its fitness in different hosts (i.e., evolutionary factors) or extrinsic to the virus related to its ecology and epidemiology that ultimately result in the virus encountering a new host population (i.e., ecological factors) (Jones, 2009; McLeish *et al.,* 2019).

The extent of knowledge concerning evolutionary factors that allow viruses to adapt to new hosts or environments extrinsic to the host, far outweighs that about the effects of ecological factors on viral emergence (Jones 2009; Lefeuvre *et al.,* 2019). In recent years, the rise of emerging infectious diseases and the advent of high-throughput sequencing (HTS) approaches for virus detection and discovery, have promoted the joint study of ecological and evolutionary factors that facilitate the ability of viruses to infect their hosts (infection risk) and mechanisms that trigger emergence (McLeish *et al.,* 2020b).

1.3. Evolutionary factors in virus emergence

1.3.1. The genetics of host range evolution

The search for evolutionary determinants of virus host ranges has led to the identification of virus traits that correlate with host range. Several studies have analysed the relationship between host range breadth and several viral traits. One example was the analysis of a set of 480 plant viruses with host ranges reported in the VIDE database (Brunt *et al.,* 1996). This analysis showed that viruses with single-stranded genomes, those with three genome segments, and nematode transmitted viruses had broader host ranges (Moury *et al.,* 2017). A broad host range was also associated with seed transmission, however, the

association between seed transmission and host range breath depended on the nature of the genome. Additionally, this study showed that virus host ranges are constrained by plant taxonomy at the family taxonomic rank (Moury *et al.,* 2017).

Since the specificity of infection is an important feature of host-virus interactions, genetic studies of host range evolution have focused on the co-evolution of hosts and viruses. Genetic specificity, i.e., infection and multiplication in a given host can be produced by only a subset of pathogens, and for a given host-virus interaction, a subset of virus genotypes can infect and multiply in a subset of host genotypes, is a major determinant of host range (McLeish *et al.,* 2018). The outcome of these interactions can be integrated into co-evolutionary models, which consider the interaction between loci for resistance and infectivity in the host and virus, respectively (Agrawal & Lively 2002; Fraile & García-Arenal 2010). The two main models explaining pathogen genotype-host-genotype interactions are the gene-for-gene (GFG) and the matching-alleles (MA) models, initially proposed for plant-pathogen and invertebrate-pathogen interactions (Agrawal & Lively, 2002; 2003).

The GFG model proposes a hierarchy of alleles that determine host resistance or viral infectivity; i.e., some resistance alleles are intrinsically better than others, which confers resistance to a broader set of virus genotypes, and likewise some viral infectivity alleles are intrinsically more efficient than others, which allows infection of a broader subset of host genotypes. In contrast, the MA model assumes that there is no such hierarchy of resistance/infectivity alleles (Agrawal & Lively, 2002). Instead, a particular resistance allele is more effective at resisting a subset of virus genotypes and less effective at resisting the rest; the same is true for some viral infectivity alleles, which are inherently better at infecting a subset of host genotypes. The GFG or MA evolutionary models are related to different molecular mechanisms of host-virus interaction. For the GFG model, successful infection occurs when the host genotype does not recognise the virus genotype and for the MA model, infection requires molecular matches between host or virus (Agrawal & Lively, 2002; Dybdahl *et al.,* 2014). These two models condition the evolution of resistance/infectivity alleles. In particular, the GFG model proposes that viruses tend to evolve to infect all host genotypes

(universal infectivity). However, this evolutionary outcome is not possible for the MA model (Agrawal & Lively, 2002; 2003; Frank, 1996). These models can also explain the evolution of host range at the interspecific level. A GFG-like interaction allows a host range expansion and thus an evolution towards generalism (broad host range). An MA-like interaction, on the other hand, implies specialism (narrow host range), and it does not allow host range expansions but conditions host switches (Weitz *et al.,* 2013).

1.3.2. Fitness trade-offs in host range evolution

The role of fitness penalties (costs) associated with host resistance and viral infectivity in host-pathogen coevolution is another difference between the GFG and MA models. Theory predicts that for GFG-like interactions fitness penalties would be required to maintain polymorphisms of resistance and infectivity loci in the host or virus population, respectively. If the interaction is according to the MA model, these penalties are not required for polymorphisms to occur (Agrawal & Lively, 2002; 2003; Frank, 1996). The cost of infectivity on new host genotypes would hinder host range expansion in GFG-like interaction. However, in MA-like interaction, adaptation to a new host would result in host switching. MA-like interaction assumptions can be relaxed to allow for partial resistance/infectivity by incorporating partial resistance/infectivity costs. It has been proposed that pure GFG or MA-like interactions may not exist in nature, where real host-pathogen systems most likely lie in intermediate positions in a continuum between these two extremes (Agrawal & Lively, 2003).

Both pure GFG and MA models, and GFG or MA models with relaxed assumptions, imply that the fitness of the virus varies across all its potential hosts, leading to the concept of across-host adaptive trade-offs (McLeish *et al.,* 2018). As fitness is environment dependent, and the host is a major component of a virus' environment, fitness cannot be maximised simultaneously for all potential hosts **(Figure 1.3.)**. Therefore, adaptation to a specific host implies a fitness cost in other hosts, and adaptation to one or a few related hosts, where its fitness will be maximised, evolving towards specialisation. In this way costs of host adaptation will translate into an adaptive trade-off among hosts. Thus, a

generalist virus would optimise its fitness across all its hosts, but in each individual host its fitness will be lower than the corresponding maxima, favouring the evolution of specialisation rather than generalisation (Bedhomme *et al.,* 2015; Elena, 2017).

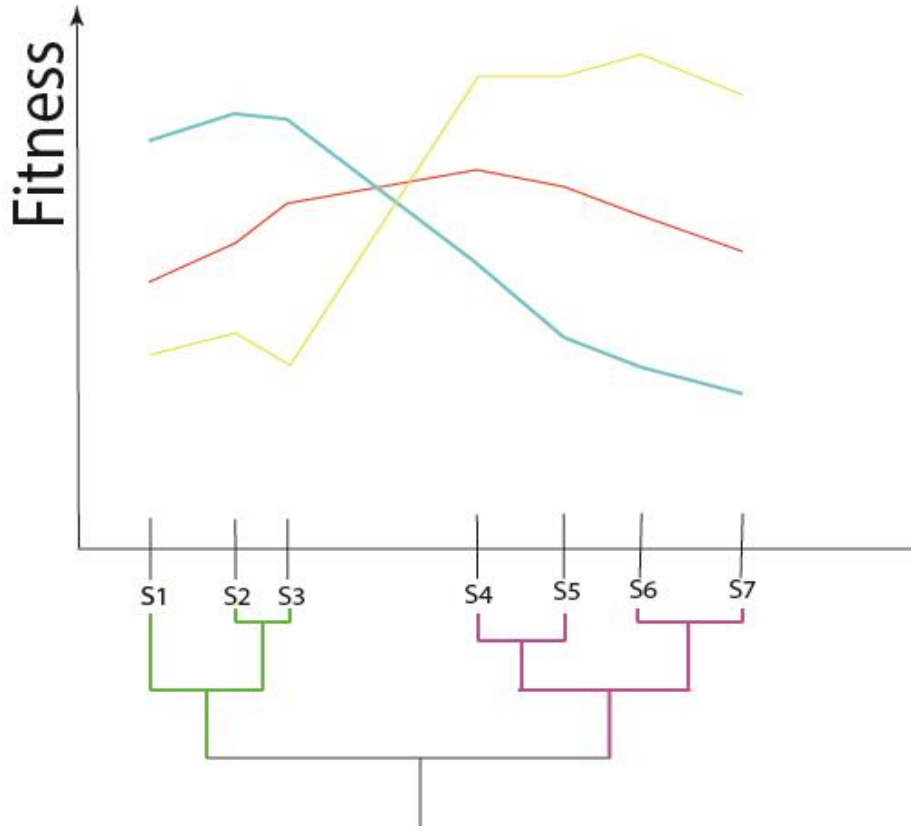

Figure 1.3. Theoretical example of the fitness of three virus genotypes across susceptible hosts that differ in their degree of genetic relatedness. The virus genotype represented by the yellow line has evolved on and adapted to species 6 (S6), in addition, it is a specialist for the host species from 'pink' clade but has very low fitness for the host species from 'green' clade. The genotype represented by the blue line has evolved on and adapted to species 2 (S2) and is a specialist of high fitness in the 'green' clade but pays a fitness cost in host species from the 'pink' clade. Finally, the red line shows the situation for a generalist virus that is paying a fitness costs all hosts. However, on average this virus genotype has optimised its fitness across all its hosts, but its fitness on each host is lower than the fitness showed by the other virus genotypes (Modified from Bedhomme *et al.*, 2015).

Because of the relevance of these general concepts for host range evolution, efforts to gather evidence for the occurrence of adaptive costs and across-host fitness trade-offs have been important. The observational (e.g., Sharp & Hahn, 2010) and experimental (e.g., Agudelo-Romero *et al.,* 2008)

evidence of across-host trade-offs has shown that emergence of plant and animal viruses often occurs because the virus jumps from its original host (reservoir) to another host that is taxonomically related (Longdon *et al.,* 2014). Further evidence has shown that generalist plant viruses do not perform equally well in all their hosts. For example, an analysis of 5 generalist viruses in 21 wild plant species showed significant host-virus associations where high prevalence corresponded to high host selectivity (i.e., host specialisation) (Malpica *et al.,* 2006). These results suggest that multiplication and transmission of generalist viruses differs between hosts with possible adaptive trade-offs (McLeish *et al.,* 2018). Experimental evidence is abundant of fitness penalties associated with adaptation to a new host as expressed by a reduced fitness in the primary host (Elena *et al*., 2009; 2014; García-Arenal & Fraile, 2013; Miyashita *et al*., 2016). For instance, adaptation of *Hibiscus chlorotic spot virus* (HCSV) to *Chenopodium quinoa* Willd. caused a loss of fitness in its original hibiscus host (Liang *et al.,* 2002). Another example was shown in serial passages in peas of *Plum pox virus* (PPV) isolates from peach, which resulted in increased infectivity and multiplication within the new host (i.e., adaptation) and decreased transmission efficiency in the original host (Wallis *et al.,* 2007). For instance, a penalty of replicative fitness in wild plants was quantified after expanding the host range of wild isolates of turnip mosaic virus (TuMV) from wild type turnips to plants carrying the *TuRB01* resistance gene (Jenner *et al.,* 2002). Another example was provided by Poulicard *et al.* (2010) where the accumulation of wild-type (WT) and resistance breaking (RB) genotypes of two *Rice yellow mottle virus* (RYMV) pathotypes showed fitness costs associated with resistance-breaking. Other examples can be found in the reviews cited in this paragraph.

1.3.3. Causes of across-host fitness trade-offs

The mechanistic causes of across-host fitness trade-offs have been broadly investigated. The simplest and most intuitive cause is antagonistic pleiotropy, i.e., mutations that have a positive fitness effect on a given host are deleterious in an alternative one (Whitlock, 1996). Antagonistic pleiotropy has been evidenced through various studies of many virus species belonging to

different taxa, such as potato virus Y (PVY), tobacco mild green mosaic virus (TMGMV), TuMV, soybean mosaic virus (SMV), HCRSV, pepper mild mottle virus (PMMoV), tomato mosaic virus (ToMV), and pelargonium flower break virus (PFBV) (Ayme *et al.,* 2006; Bera *et al.,* 2018; Fraile *et al.,* 2011; Ishibashi *et al.,* 2012; Janzac *et al.,* 2010; Jenner *et al.,* 2002; Liang *et al.,* 2002; Montarry *et al.,* 2012; Moreno-Pérez *et al.,* 2016; Moury & Simon, 2011; Poulicard *et al.,* 2010, 2012; Rico *et al.,* 2006; Wang & Hajimorad, 2016). Antagonistic pleiotropy of host range mutations may be particularly relevant to RNA viruses, as most plant viruses are, as an inevitable consequence of their small, compact genomes that code for multifunctional proteins. This genomic architecture makes it difficult to enhance one function without impairing another (McLeish *et al.,* 2018).

Another mechanism that generates across-host trade-offs is the epistatic interactions among mutations involved in processes of host adaptation, (Bedhomme *et al.,* 2015). The effects of interactions among mutations involved in epistasis are not additive (Bedhomme *et al.,* 2015). Epistasis has been studied in detail in plant viruses through experimental evolutionary approaches (Sanjuán *et al.,* 2004).

In a trade-off scenario, the evolution of generalism requires high rates of transmission between different hosts that prevents adaptation to any one host (McLeish *et al.,* 2018). If among-host transmission is infrequent, the virus will adapt to the current host, otherwise the virus population will behave as if the fitness landscape (i.e., visualisation of a high-dimensional map of the relationship between genotype and fitness (Singhal *et al.,* 2019)) were constant and equivalent to the average for each host (Bedhomme *et al.,* 2015). This evidence has been supported by different experiments with plant viruses (Lalić & Elena 2012; Montarry *et al.,* 2011; Poulicard *et al.,* 2012), bacteriophages (Cabanillas *et al.,* 2013), and human viruses (Da Silva *et al.,* 2010). In addition, a phylogenetic study carried out by Streicker *et al.* (2010) showed that the intensity of rabies transmission rates among 23 bat species was higher between taxonomically related species, and that cross-species trade-offs between geographical distant hosts were more pronounced that between geographically close hosts, and is consistent with lower rates of transmission.

1.4. Ecological factors in virus emergence

1.4.1. The relationship between ecosystem anthropisation and disease risk

Geophysical factors and human management of the landscape through practices used in agriculture, husbandry, or fuel production, have significant effects on ecosystems that result in their fragmentation. Land-use practices that result in mosaics of different vegetation types, or habitats (defined as an ecological or environmental area inhabited by an organism or a population of a species (Thomas, 2019)), may result in ecosystem simplification and increase connectivity between plant and virus populations (Elena *et al.,* 2014). Consequently, changes in host ecology are linked to novel host-virus encounters, changes in transmission and disease spread, and therefore, causes of disease risk (Jones, 2009), defined as the probability that a host will develop a disease (McLeish *et al.,* 2018; Roossinck & García-Arenal., 2015).

Host diversity has been considered a fundamental factor in the study of disease risk and virus evolution (Keesing *et al.,* 2006; Ostfeld & Keesing 2012). In fragmented ecosystems, diverse mosaics of wild and anthropic habitats may share climatic and historical characteristics, and biological invasions, which are related to the spatial distribution of species over large and small spatial scales (Whittaker *et al.,* 2001). Species interactions such as parasitism and mutualism, may be partly contingent on climate (Cowling *et al.,* 1996), the topology of a study area (Tischendorf & Fahrig, 2000), or landscape disturbance (Büchi *et al.,* 2009). Local scale patterns tend to be driven by species interactions (McLeish *et al.,* 2021). Landscape-scale studies of disease risk often characterise compositional differences among communities in terms of host distribution and abundance, to generalise about ecosystem processes (Tilman *et al.,* 1997). For example, the proportion of crop land compared to unmanaged land has been shown by several studies to have a significant effect on the richness of vector communities, and ultimately on the risk of disease (Claflin *et al.,* 2017; Tamburini *et al.,*2016). However, environmental heterogeneity influences biological processes at different scales (spatial and temporal). Observations of compositional variation among plant communities made by studies at different scales, may not coincide

with the same underlying causal processes. Therefore, biological processes that are inferred from patterns of species distributions and abundance alone, are not scale-insensitive. Information on the distribution of plant viruses at the landscape scale, which could show scale-dependencies of pattern-process, is scarce (Malmstrom *et al.,* 2011; Roossinck & García-Arenal, 2015). This makes detecting and interpreting all the causal factors of disease risk and virus evolution a difficult task (Johnson *et al.,* 2015).

Two popular and contrasting hypotheses have emerged as generalisations for explaining disease risk as related to host community diversity. The first, the dilution effect hypothesis (Ostfeld & Keesing 2000; Schmidt, & Ostfeld 2001; Van Buskirk & Ostfeld 1995), predicts that an increase in diversity is negatively correlated with disease risk, as an increase in diversity would result in a decrease of the abundance of the focal host, reducing the number of contacts between susceptible and infected individuals (Keesing *et al.,* 2010; Van Buskirk & Ostfeld 1995). For example, some studies have analysed the relationship between diversity and disease risk in the wild pepper or chiltepin (*Capsicum annuum* var. *glabriusculum*), a species that grows in plant communities with different levels of human management in Mexico (Pagán *et al.,* 2012; Rodelo-Urrego *et al.,* 2013; 2015). Results showed that a low diversity was positively correlated with infection risk of the begomoviruses *Pepper golden mosaic virus* (PepGMV) and *Pepper huasteco yellow vein virus* (PHYVV) (Pagán *et al.,* 2012). Thus, this study supported the Dilution effect hypothesis. The second, the amplification effect hypothesis, postulates that an increase in diversity is positively correlated with disease risk, as an increase in host community diversity may result in a higher abundance of host species, increasing the inoculum sources for a focal host (Ostfeld & Keesing, 2012). Contrary to the dilution effect hypothesis, other studies have supported the amplification effect hypothesis (Borer *et al.,* 2010; Lacroix *et al.,* 2014; Power & Mitchell 2004; Power *et al.,* 2011).

Recent studies suggest that the diversity-disease risk relationship is not generalisable (Hosseini et al., 2017; Young et al., 2017). Different traits of individual species predict whether they are likely to be competent hosts or non-competent hosts, such as defence strategies (Han *et al.,* 2015) among others

(Lloyd-Smith *et al.,* 2005). Because the susceptibility and ability to sustain the replication or transmit a specific virus is expected to vary among host species/genotypes, changes in species composition, rather than diversity per se, have a greater impact on disease risk than diversity. The presence of certain host species might have a positive or negative effect on disease risk (Levi *et al.,* 2016). For example, an increment in community diversity may lead to a higher non-competent hosts abundance that may reduce disease risk by diluting vector encounters, competing with competent hosts in the community and decreasing their abundance (Luis *et al.,* 2018). Thus, management of specific species rather than overall diversity could have a greater effect on disease risk.

The extent to which causal mechanisms are affected by species composition may depend on community (dis)assembly; i.e., how species (traits) are added to or removed from a community (Ostfeld & LoGiudice 2003). Often in a community there is both the replacement of individuals of existing competent host species by a new competent host species (i.e., substitutive assembly), as well as the addition of individuals of a new competent host species (i.e., additive assembly) (Mihaljevic *et al.,* 2014) that reduce the density of less competent host species. Amplification or dilution effects on communities may occur when competent hosts or non-competent hosts, respectively, are added to or subtracted from communities via the sampling effect (that is, more diverse communities are more likely to contain a host species that either increases or decreases disease risk).

Establishing the relationship between biodiversity and disease risk is not a trivial task as it is not a linear relationship and may be affected by other traits that contribute to the maintenance of biodiversity and ecosystem functioning, which have been identified as fundamental to understanding when declines in biodiversity will increase, decrease or have no effect on disease risk (Thompson *et al*., 2017; Welsh *et al.,* 2020).

1.4.2. The relationship between host range and host community composition

Differences in biotic and abiotic environmental conditions including hosts, and the availability of resources in each location, determine viral community composition (McLeish *et al*., 2019). Community composition highlights the identity of all taxa that a community consists of, and their relative abundances (Hanson *et al.,* 2012). The role a species has as a host or pathogen, may depend on its ability to compete with other taxa already present in the community. For example, a mycovirus conferred heat tolerance in a fungal endophyte associated with a grass, and allowed the fungus and grass to grow at high soil temperatures (Márquez *et al.,* 2007). Variation in the composition of the host community may correspond to variation in available resources and ecological interactions (Roossinck & García-Arenal, 2015), and affect the host range of viruses and their evolution. The interactions among species in a community, or in a given environment, determine what organisms are present in it, or the community structure (Lavorel & Garnier 2002; Van der Heijden *et al.,* 1998). Species diversity (i.e., richness and evenness) and trophic structures are some of the methods commonly used in comparing community structure.

Morphological, physiological, and behavioral traits that suit a species to a particular habitat or resource and limit its ability to use other environments (i.e., ecological specialisation), is an important cause of biological diversity of communities (Via & Hawthorne, 2002). Two hypothetical communities may have the same host species diversity but may differ according to the phenotypes each community of hosts possesses (i.e., functional diversity). Resource use, especially the host species used by a virus, is usually generalised in terms of the number of hosts used to fulfil their reproductive needs (McLeish *et al.,* 2018). Thus, the specificity in the use of hosts is typically defined by host range. Alternatively, the range or variety of environmental conditions a species can exist or grow in may be captured in what is termed a niche (Sexton *et al.,* 2017). Ecological specialisation is similar in many respects to niche theory, which establishes the relationship between a species' resource exploitation strategy and the breadth of resources available to it (Hutchinson, 1957).

A major development in niche theory were the concepts of the realised niche, which refers to the subset of available resources that a species is able to use, and the fundamental niche (Futuyma & Moreno, 1988; Soberón & Peterson, 2005). The fundamental niche stresses the potential of a species to utilise an environment, and for viruses this typically includes susceptible host tissue and host genotype (Al-Naimi *et al.,* 2005). The realised niche of a virus may also include ecological interactions, such as predation, mutualism, herbivory and parasitism, and the hosts that are infected, the realised host range. Additionally, other factors such as co-infection, the coexistence of different parasites or strains in the same host, influence the virus niche (May & Nowak, 1994), and contribute to the breadth of the realised host range. Viral traits that are important in the evolution of the host range at either local- (e.g. competition) or broad-scales (e.g. transmission), are expected to be different as the number of resources available for specialisation increases (Kneitel & Chase, 2004). Therefore, spatial and temporal variation in habitat quality (i.e., environmental heterogeneity) has consequences for host range evolution.

Together with processes such as intra- and interspecific competitive interactions, the distributions and abundances of host, virus, and vector traits underlie the habitat quality for plant viruses (McLeish *et al.,* 2018). Variation in habitat quality might explain differences in host specificity. For example, the degree of host specialisation of populations of the pathogenic fungus *Colletotrichum cereale*, which infects exotic grasses in North America, corresponded to ecosystem and/or host plant (Crouch *et al.,* 2009). The study showed that low levels of gene flow among isolated populations and ecosystems contributed to the degree of host specialisation. Host specificity is a measure of host range, but also how closely related hosts are genetically to one another (Poulin & Mouillot, 2003) and provides information about the degree of specialisation. This is usually articulated as the distinction between generalist and specialist species. If genetic constraints on host preference by viruses are not strong among available host species (Bedhomme *et al.,* 2012; Elena, 2017), the primary factors acting on community composition might reflect stochastic fluctuations in local extinctions, speciation, and gene flow. In the absence of

strong genetic constraints, host range is influenced by environmental heterogeneity, which in turn underlies host range evolution (Pandit *et al.,* 2009; Seabloom *et al.,* 2015). For example, Malpica *et al*. (2006) showed that host selectivity (i.e., host specialisation) corresponded to the highest prevalence in an assemblage of generalist viruses, and suggested specialisation was a successful strategy. Another example by McLeish *et al*. (2017), tested the effect of habitat quality on the host range of 11 generalist plant virus species, and showed evidence of host range plasticity consistent with the definition of facultative generalism (Shipley *et al.,* 2009). Under facultative generalism, the host range of species with broad host ranges varies with the availability of resources and can be narrow when a host is favored and offers fewer obstacles (McLeish *et al.,* 2017).

Additionally, adaptation may not be required for negotiating environmental variation and changes to the availability of hosts (González *et al*., 2020). If the traits of the new host are similar to those of the original host, the virus, by chance, might "fit" the association because functional genetic architectures are compatible (McLeish *et al.,* 2020b). The capacity to cope with novel conditions based on pre-existing capacities, known as ecological fitting, plays an important role in the evolution of the host range (Brooks & Boeger 2019).

Environmental heterogeneity has an important role in the distribution of viruses, and determines both community composition (Bohlen *et al.,* 2001; Johnson *et al*., 2015; Randolph & Dobson, 2012) and host range variation (Strauss *et al.,* 2015; Streicker *et al*., 2013). Thus, in order to understand host range evolution and predict disease risk, it is necessary to study multi-host multi-virus systems at landscape-scales (McLeish *et al*., 2017). The inherent complexity of landscape-scale studies presents analytical challenges, such as multivariate datasets with dissimilar probability distributions (Warton et al., 2015), taxonomic anomalies (Blüthgen *et al.,* 2008; Novotny *et al.,* 2002), and spatial or temporal scale dependencies (Nash *et al.,* 2014; O'Dwyer & Cornell, 2018; Peterson *et al.,*1998). However, new approaches that combine HTS and new tools for big-data, have made it possible to link molecular biology and ecology, and the capacity to integrate heterogeneous data.

Coupled with HTS and new computational tools, large-scale sampling approaches have enabled the analysis of spatial and temporal variation in the distribution and abundance of plant hosts and viruses (McLeish *et al.,* 2018). The ability to rapidly obtain data on multi-host multi-virus communities at the landscape scale will further the understanding of relationships between species and the environment, and the influence this has on host range evolution. As a first step to understanding viral emergence, the major goal of this thesis is to characterise multi-host multi-virus communities in an agricultural ecosystem in central Spain, to quantify species interactions at the landscape scale in relation to patterns of viral infection and host range evolution.

1.5. Ecological diversity in central Spain

Distinctions among viruses based on habitat quality and resource use (i.e., host species), can be used to determine whether species are ecologically divergent (Fraser *et al.,* 2009). Dissimilarities in species so-called ecological diversity, in addition to their genetic diversity, are important to identifying constraints on host range evolution. Land-use practices in the agricultural region of the Tajo Basin, in the Vega del Tajo-Tajuña, result in changes to host species distributions and abundances, and habitat quality. In this ecosystem, habitats are distinguished by difference in the dominant vegetation cover types and in species diversity. Key habitat categories include evergreen oak forest communities, (Oakwood habitat), unused patch communities or “eriales” that we call wastelands (Wasteland habitat), communities on the edges (Edge habitat) between crops and the communities in crop fields (Crop habitat). In this thesis I focus on these four habitats **(Figure 1.4)**, which represent variation in species diversity due to land-use practices. The features of the studied habitats, Oakwood, Wasteland, Edge, and Crop are described below.

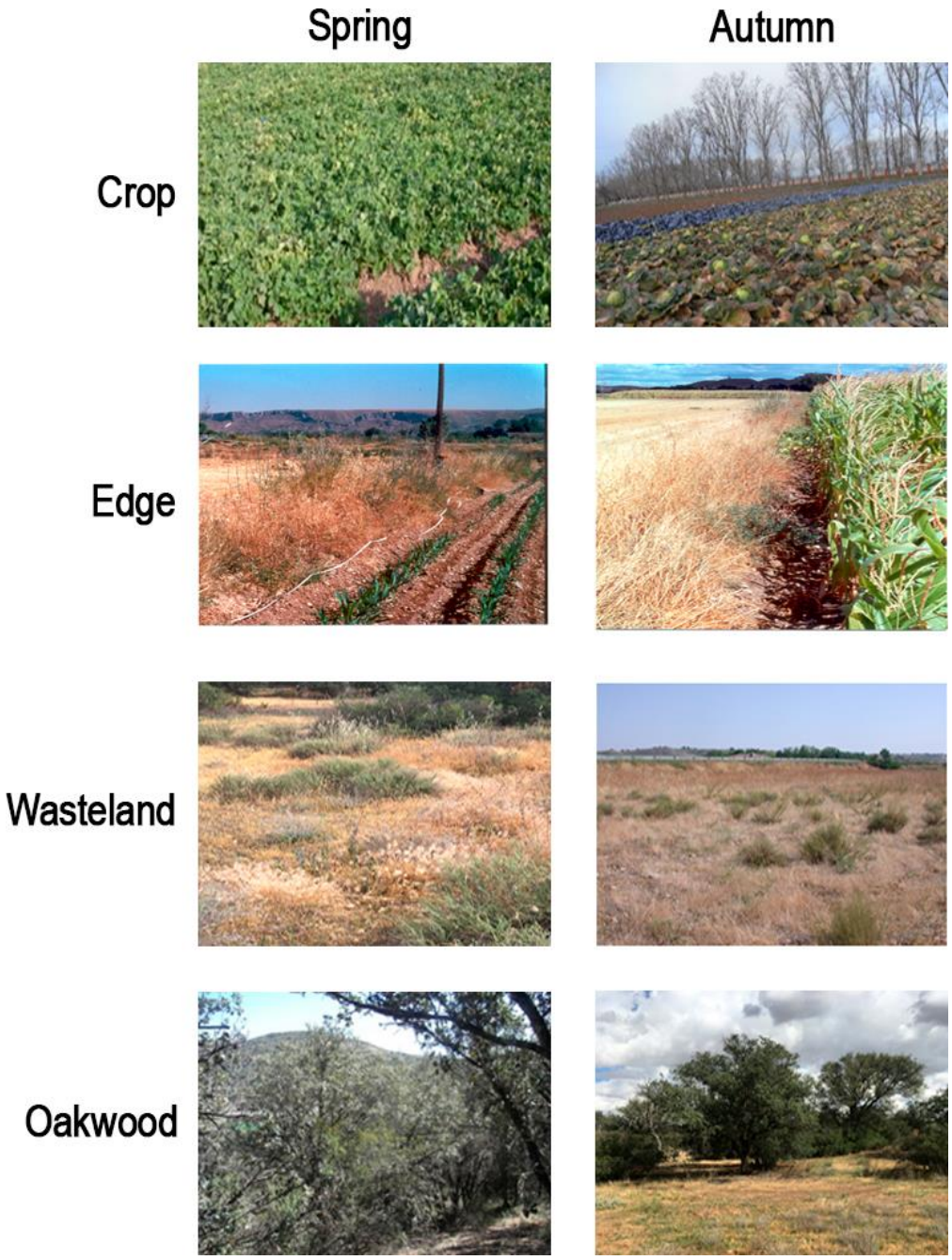

Figure 1.4. Representation of the habitats in in the Vega del Tajo-Tajuña agricultural region in autumn and spring seasons.

1.5.1. Oakwood habitat

Forests cover large areas of central Spain and a broad range of altitudes, climates, and substrates (Gavilán *et al.,* 2018). Their distributions occur in climates with highly variable temperatures and humidities (Granda *et al.,* 2014). The evergreen oak forest is considered as a natural ecosystem of this central

region in Spain, which has been variably disturbed by practices such as fuel production (prior to 1960s), hunting, or light grazing (Romero-Calcerrada & Perry, 2004). Oakwood habitat is made up largely of evergreen oaks, being *Quercus ilex* L., *Quercus coccifera* L., and *Quercus suber* L. the dominant species of this region, and the two first species typically occur on calcareous soils (Gavilán *et al.*, 2018) and frequently appear in our inventories.

Under the canopy of *Quercus* there is an understory that includes a broad variety of xerophilous and sclerophyllous bushes, such as *Daphne gnidium* L., *Thapsia villosa* L., *Cistus ladanifer* L., *Asparagus acutifolius* L., and *Retama sphaerocarpa* L. (Blanco *et al.,* 2005). However, if there is soil degradation, shrubs such as *Lavandula stoechas* LAM. and *Thymus mastichina* L. predominate along with *Cistus sp*. L. (Blanco *et al*., 2005). In addition, common perennial herbs such as *Tuberaria guttata* L., and annual species of grasses, legumes, and asters, both constitute the herbaceous stratum.

1.5.2. Wasteland habitat

Wasteland habitat represents the abandoned deforested areas that were traditionally grazed, often overgrazed, or were used for crop cultivation, which cessed progressively during the second half of the 20th century (Lana & Iriarte-Goñi 2015). Land use practices in Wasteland habitat resulted in poor soil quality and marginal productivity (Pandey *et al.,* 2020). Once human interventions were reduced, or ceased, Wasteland came to support a rich flora that includes rare species (Godefroid & Koedam, 2007; Muratet *et al.,* 2007; Öckinger *et al.,* 2009). Depending on the period and type of disturbance, Wasteland can support vegetation ranging from pioneer to pre-forest stages (Twerd & Banaszak-Cibicka, 2019) with a floristic composition that may include species such as *Artemisia vulgaris* L., *Cirsium arvense* L., *Plantago lanceolata* L., *Picris hieracioides* L., *Taraxacum campylodes* L., and *Rumex obtusifolius* L. (Muratet *et al.,* 2007).

1.5.3. Edge and Crop habitats

Crops in the study area are largely annual seasonal monocultures interspersed by weed species, which are rotated or left fallow between seasons.

In the study area there are both dry farming and irrigated crops. In the alluvial soils of the river valleys ("Vegas") crops are irrigated, except for some winter crops. Cultivated fields are fertilised, and crops and weed species are ploughed over between seasons. This results in cycles of colonisation and assemblies of vegetation that benefit from soil preparation, nutrients, and water applied to the crop (Sacristán *et al.*, 2004). The diversity of plants in Crop is conditioned on the species of crop that is grown, and the seed bank associated with the field (Wilson & Aebischer, 1995). Due to changes in agricultural practices over the last decades such as in the increased use of herbicides, there has been a dramatic decrease of weeds in this habitat (Baessler & Klotz, 2006; Hyvönen, 2007). However, a massive increase in the use of fertilisers that favour crop growth could also benefit weeds (Travlos *et al.,* 2018). We focus on irrigated summer crops, melon and maize, and winter crop cauliflower, and dry-farmed winter crop, barley. These crops represent a large fraction of the agricultural area of the region.

Crops are separated by narrow borders ("lindes") that constitute the Edge habitat, which is composed of relatively permanent plant assemblages where perennial species persist for years (McLeish *et al.,* 2021; Sacristán *et al.*, 2004). Plants that occur in Edge habitat benefit from irrigation and fertilisation applied to the adjacent crop. Edge habitat is generally left intact between seasonal cropping, with occasional low-level disturbance either by burning, partial ploughing, or herbicide treatments. The frequent plant species in Edge are *Chenopodium album* L., *Polygonum aviculare* L., *Convolvulus arvensis* L., *Papaver rhoeas* L., and *Stellaria media* L. (Fried *et al*., 2009)

1.6. Virus detection at the landscape scale

The use of HTS technologies has revolutionised traditional approaches in the detection of viruses and the study of emergence based on the interaction of a single virus and host (Woolhouse & Gowtage-Sequeria 2005). The versatility of HTS technologies has driven new research perspectives across scales from within cells and individuals, to multi-species communities at the landscape scale over an ecosystem (McLeish *et al*., 2020b; Roossinck, 2012; Schoelz *et al.,*

2018). The analysis of sequences derived from environmental samples that contain an unknown mixture of life forms, known as metagenomics (Roossinck, 2012), has been facilitated by the increase in accessibility to HTS technologies (Schoelz *et al.,* 2018). The characterisation of the "collective genome of a virus community within a given individual or a defined environment", or the virome (Maclot *et al*., 2020; Roossinck, 2012) provides a means to investigate simultaneously the evolutionary and ecological processes of entire populations and communities (Shates *et al.,* 2019), and allow a better understanding of the viral emergence process at multiple scales (Maclot *et al.,* 2020). A key benefit in the use of HTS applications is that no prior knowledge of the virus species or strain being targeted for analysis is required, and allows for a non-targeted detection approach (Adams & Fox, 2016). This is an important advantage over previous methods, such as the double-antibody sandwich enzyme-linked immunosorbent assay (DAS-ELISA; Clark & Adams, 1977) or RT-PCR, which necessitate a targeted detection approach (Boonham *et al*, 2014; Massart *et al,* 2014). Additionally, the ongoing development of analytical tools that facilitate the management, analysis, and visualisation of big-data obtained from HTS technologies (de Brevern *et al*., 2015) has improved the characterisation of the virome.

Genomic HTS approaches have been used to study molecular processes such as the silencing effects on RNA and chromatin (Donaire *et al.,* 2009), and more recently, used to reveal the diversity and functional attributes of viruses associated with wild environments (Bernardo *et al*., 2018; Melcher & Grover, 2011; Roossinck *et al.,* 2010). However, viruses in wild environments that have a low abundance can require a higher number of read sequences to detect in a given sample. Along with biological factors, technical biases that inhibit the detection of viruses may result from the quality of the nucleic acid in the sample (McLeish *et al.,* 2020a). At the landscape scale, variation in virus abundance as well as differences in host species' properties that may inhibit the extraction of nucleic acids, invalidate the application of a uniform protocol across samples prior to sequencing (McLeish *et al.,* 2020a).

There are different approaches used for virus detection by HTS. These approaches differ on the basis of the nucleic acid type targeted for sequencing (Roossinck, 2012). The main approaches focus on double-stranded RNA (dsRNA), virus-derived small interfering RNA (siRNA), virion-associated nucleic acids (VANA), and total RNA (totRNA); with or without ribosomal RNA depletion or the presence of polyadenylated RNA (poly(A) RNA) (Bejerman *et al.,* 2020; Maclot *et al.,* 2020; Pecman *et al*., 2017). Despite the great capacity to sequence any type of nucleic acid, there are advantages and disadvantages to the application of HTS, which depend largely on: i) the type of genome of the target virus (Barba *et al.,* 2014; Wu *et al.*, 2010) ; ii) the ability to reduce read sequences derived from the host or environment from where it was extracted, which are not from the target (Roossinck *et al.,* 2015); iii) the technique used for sample preparation (Maclot *et al.,* 2020); and iv) the availability of annotated reference sequences (Ng *et al.,* 2018).

The approach based on dsRNA enrichment, seeks to enrich viruses with dsRNA and ssRNA stages in the replication process. This approach has the advantage of producing purified and enriched virus RNA prior to sequencing, that allows for a specific and in-depth analysis of the metagenome. However, this approach is biased against (-)ssRNA virus sequences, since these do not accumulate large amounts of dsRNA during replication, and does not detect viruses with DNA genomes (Roossinck *et al.,* 2015). The siRNA approach is based on the detection of virus-derived small interfering RNAs (siRNAs) that are generated in the silencing response produced by the plant immune system to virus infection and target the viral RNAs for degradation (Ding & Voinnet, 2007; Mlotshwa *et al.,* 2008; Wu *et al*., 2010). The advantage of this approach is that it is very sensitive to viruses with either RNA or DNA genomes (Adams & Fox, 2016). However, it may be biased against viruses that do not trigger silencing responses or produce highly efficient silencing suppressors. Also, this approach involves a laborious nucleic acids preparation based in Trizol and CTAB-based protocols (Roossinck *et al.,* 2015). The approach based on VANA enriches viral particles by filtration and centrifugation (Melcher *et al.,* 2008), with the advantage of enriching low titre viruses, as well as viruses with either RNA or DNA genomes.

The disadvantage of this approach is that sample processing is complicated, and that purification may favour the detection of certain types of viruses according to the stability or other characteristics of the virus particle (Adams & Fox, 2016). The simplest and least costly approach uses extracts of total RNA (totRNA). The approach is not biased against the detection or discovery of any particular virus genome or particle. However, it has the disadvantage of generating a large proportion of extracted nucleic acids that are plant-derived, and reduced ability to detect low titre viruses. Ribosomal RNA depletion protocols (RiboZero, Illumina, USA) that are used to reduce the concentration of plant RNA in the extract, can improve the ratio of virus-to-plant derived RNA, and not bias against viral RNA (Adams & Fox, 2016), although it increases the cost and time required for sample preparation. Finally, the approach based on the enrichment of RNAs with poly(A) tails has the advantage of being able to detect viruses with DNA and RNA genomes. However, many virus-derived RNAs have genomes, or messenger RNAs (mRNA), that do not have poly(A) tails (Wu *et al.,* 2015). Each method has practical limitations that have consequences for the characterisation of the virome, and the selection of any approach must fit the goal of the study.

Another limitation of HTS is an outcome of reference-based alignment approaches for the detection of viruses that rely on homology-based annotation. A significant proportion of read sequences produced from metagenomic samples do not have references in the currently available sequence databases (Lefebvre *et al.,* 2019; Roossinck, 2012). Most of the known plant viruses have been isolated from domesticated plant species (Wren *et al.,* 2006), which may imply that many deeply divergent viruses, or those that lack common ancestry with known virus families remain undiscovered (Bejerman *et al.,* 2020). Moreover, other types of detection methods have been developed that aim to identify new viral sequences without the use of a sequence database. For example, by examining the characteristics of virus-derived small RNAs. The size profile of virus-derived small RNAs has been found to be a unique signature for each virus and can be used to detect viral sequences without homology present in reference databases (Aguiar *et al.,* 2015). Other approaches rely on automated annotation and machine learning classifiers of sequences, such as Markov profiling (Burks

& Azad 2020; Richardson *et al.,* 2020), or *k*-mer analyses (Alam & Chowdhury 2020; Ren *et al.,* 2017) that integrates unique features of virus genomes with information from similarity-based methods (Wood *et al.,* 2019). However, these methods are all partly dependent on either sufficient training data, the presence of universal markers, or computational power (Alam & Chowdhury, 2020; Burks & Azad, 2020; Richardson *et al*., 2020).

The detection of viruses by HTS is expected to be less prone to generate misleading results/ false positive (results caused by background reactions) or miss the results /false negative (results caused by unexplained interactions of reagents with target or host nucleic acids or proteins or lack of sensitivity) than other detection techniques (Ge *et al.,* 2003; Maree *et al.,* 2018; Roenhorst *et al.,* 2018). As with any method for the detection of viruses (Torre *et al*., 2020), attention needs to be given to the sensitivity and specificity of HTS techniques (Massart *et al.,* 2019). The sensitivity of the technique is directly related to the proportion of viral RNA among the cellular RNA in the sample, the efficiency of the enrichment strategy (if used), and the sequencing depth (Lambert *et al.,* 2018). In this context, precautions need to be taken to use controls to determine whether the sensitivity of the technique for viral detection is appropriate to the objectives of the study (Maree *et al.,* 2018). Another important criterion is the specificity of HTS in the detection of the targeted viruses. The specificity of sequence detection by HTS technology can be assessed by verifying the inclusivity and uniqueness of the sequences using a range of strategies. These validation procedures include bioinformatic solutions (Lambert *et al.,* 2018) and complementary methods, such as RT-PCR and primers specific to the viruses of interest (Massart *et al.,* 2017).

1.7. *Watermelon mosaic virus*

As a case study, *Watermelon mosaic virus* (WMV) was chosen to investigate the effect of habitat on the ecology and evolution of host range. WMV is a member of the largest genus of plant viruses, *Potyvirus* (family, *Potyviridae*),

and is known for causing substantial yield losses in cucurbit crops across its global distribution (Lecoq & Desbiez, 2012). As with all members of the genus *Potyvirus*, WMV is non-persistently transmitted by aphid species (Wylie *et al.,* 2017). It has been reported that WMV infects over 170 species of 26 families (Lecoq & Desbiez, 2012). WMV can produce a broad range of macroscopic symptoms. On the leaves these symptoms might be mosaic, vein banding, leaf deformations (including blisters, filiformism, and leaf size reduction), and on the fruit the symptoms might be severe discoloration, slight deformation, and necrosis, reducing the yield and quality of crops (Crescenzi *et al.,* 2001; Lecoq & Desbiez, 2012). The severity of WMV infection depends on the specific host genotype, the virus strain, and the environmental conditions affecting the physiology and development of the host (Revers *et al*., 1999). There are several studies that report synergistic interactions of WMV in co-infection with other viruses, which produce more severe symptoms than would be produced separately by each virus. For instance, Wang *et al.* (2002), studied the synergistic interactions of WMV in coinfection with CMV strains in Zucchini squash. The symptoms produced by these viruses were more severe when such viruses infected together than individually.

WMV, as all members of the genus *Potyvirus*, has non-enveloped, filamentous, and flexuous particles of circa 750 nm x 12 nm length and diameter respectively, and contain a single coat protein (CP) of 34.5 kilodalton (KDa), which comprises 282 amino acids (Gara *et al.,* 1997; Lecoq & Desbiez, 2012). Multiple copies of the CP are arranged in helical symmetry, encapsidating a monopartite positive-sense single-stranded RNA ((+)ssRNA) of 10,035 nucleotides, with a 5' noncoding region of 132 bp (Desbiez & Lecoq 2004; Martínez-Turiño & García, 2020). The 5' end of the ssRNA genome is covalently bound to a VPg (virus protein genome-linked) protein and the 3' end of the ssRNA is polyadenylated with a variable number of adenosines (Lecoq & Desbiez, 2012). The genomic RNA has a single open reading frame (ORF) that encodes a polyprotein of 3,217 amino acids with an estimated molecular weigh of 366.7 KDa that is subsequently processed by three virus-encoded proteases, which results in 10 functional proteins (Desbiez & Lecoq 2004), with most of them being

multifunctional. These products are: first protein (P1), helper component protease (HC-Pro), third protein (P3), 6K1 protein, cylindrical inclusion protein (CI), 6K2 protein, viral protein genome-linked (VPg), small nuclear inclusion protein (NIa), large nuclear inclusion protein (NIb) and coat protein (CP) **(Figure 1.5)**. Recently, a second ORF was identified, named PIPO that encodes the PIPO protein, embedded within the P3-enconding region in a different reading frame than the main polyprotein (Verma *et al.,* 2020).

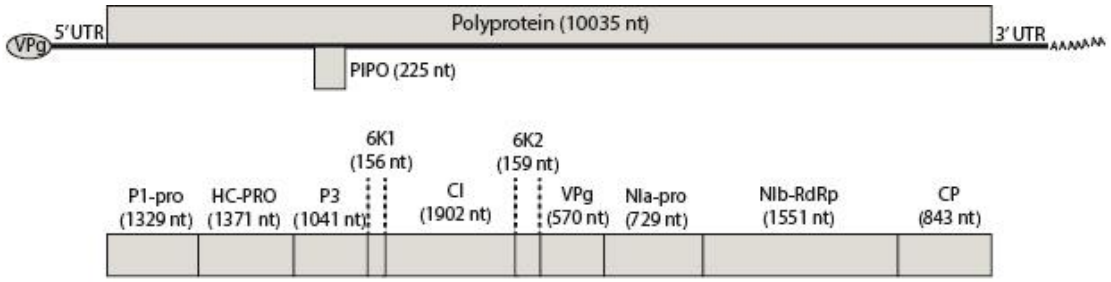

Figure 1.5. Genomic map of WMV. Open reading frames (top) and mature proteins (bottom) are represented with their bp sizes in parentheses. Reference genome taken from NCBI; NC_006262.

The P1 is the first proteinase that is required for cleaving the polyprotein at the P1/HC-Pro junction (Revers & García 2015). In addition, the P1 is involved in early replication, in the suppression of host defences, and enhancing virulence (Wylie *et al.,* 2017). The HC-Pro is the second proteinase that is required for cleaving the polyprotein at the HC-Pro/P3 junction (Revers & García 2015). Additionally, the HC-Pro is involved in multiple functions such as the suppression of RNA silencing, transmission by aphids, and within-host long-distance movement. The HC-Pro is also required to stabilize the CP and for the infectivity of potyviral progeny (Gadhave *et al.,* 2020). The P3 is involved in viral replication and there is abundant information supporting its relevance for viral pathogenicity and symptomatology (Revers & García 2015). The PIPO protein is essential for the intercellular movement of potyviruses (Wylie *et al.,* 2017). The 6k1 is present in the replication complex and is required for replication (Revers & García 2015). The CI is a multifunctional protein with helicase activity, involved in RNA replication and within-host movement. In addition, it is the avirulence factor for different resistant genes (Deng *et al.,* 2015; Sorel *et al.,* 2014). The 6K2 is a small

transmembrane protein anchoring the replication complex to the endoplasmic reticulum (Wylie *et al.,* 2017). The VPg is essential for virus replication, translation, and the suppression of RNA silencing (Tavert-Roudet *et al.,* 2017). The NIa-Pro is the third protease responsible for cleaving the remaining sites in the polyprotein (Wylie *et al.,* 2017). The NIb is the RNA-dependent RNA polymerase (RdRp) necessary for viral genome replication (Shen *et al.,* 2020). Finally, the CP is involved in different stages of the viral cycle: virus assembly (Zamora *et al.,* 2017), within-host movement, transmission mediated by aphids and host specificity (Gadhave *et al.,* 2020). The CP can be divided into three domains **(Figure 1.6.)**: two of them, the central or core domain and the carboxyl-terminal domain, are conserved among potyvirus species, while the amino-terminal domain is highly variable.

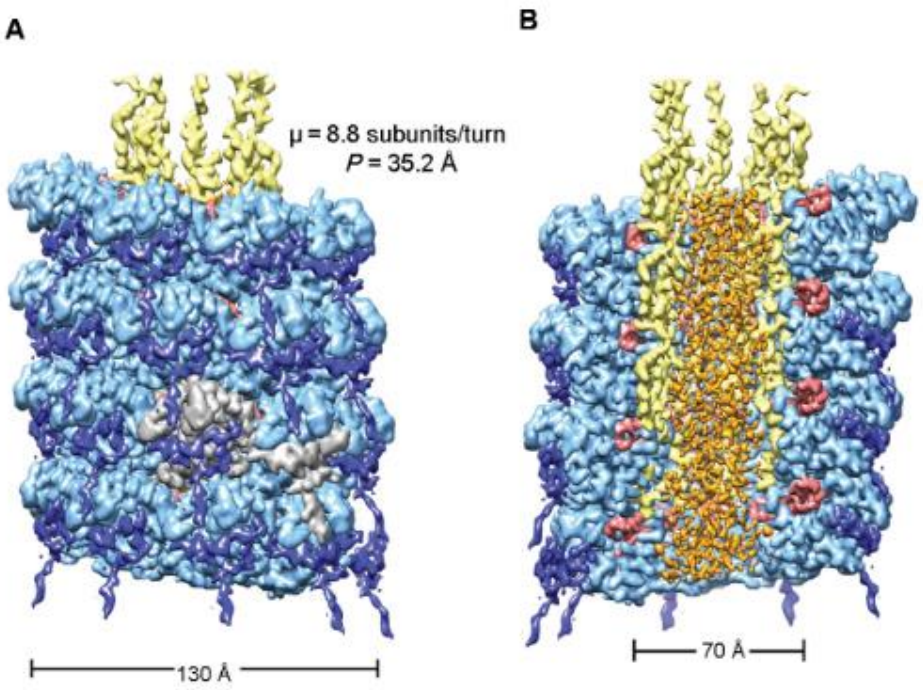

Figure 1.6. Overview of the cryo-EM structure of CP WMV virions. Renderings of the 3D cryoEM map [cutaway mode in **(B)**] calculated for WMV. The segmented densities are shown as follows: WMV CP central domain (light blue), amino-terminal domain (dark blue), carboxyl-terminal domain (yellow), scattered densities on the inner side of the helix (orange), and density for the ssRNA (red). In addition, one of the CP subunits is seen as grey **(A)** (Taken from Zamora, *et al.,* 2017).

All the domains control phosphorylation of the CP protein. Central and amino-terminal domains are involved in virion assembly and short-distance movement, while carboxyl and amino-terminal domains are implicated in seed transmission and long-distance movement. There are other functions that are controlled by unique domains, like RNA binding to the CP (central domain), regulation of the RNA translation (carboxy-terminal domain), aphid transmission, post-translational modification, and host adaptation (amino-terminal domain) (Martínez-Turiño & García, 2020).

1.7.1. Host range, ecology, and epidemiology

WMV has a worldwide distribution, with a relatively broad host range of 170 species of 26 families, and is widespread in cucurbit crops, especially in temperate and Mediterranean climates (Lecoq & Desbiez, 2008; 2012). Besides cucurbits, WMV can infect in natural conditions carrot, pea, and orchids (Kassem *et al.,* 2007; Lecoq *et al.,* 2011; Parry & Persley, 2005). Additionally, WMV may infects many weed species that may act as reservoirs throughout the year (Lecoq & Desbiez 2012).

WMV is horizontally transmitted in a non-persistent (stylet-borne) manner by several aphid species that vary in their transmission efficiency (Dombrovsky *et al.,* 2005). At least 35 different aphids have been associated to WMV transmission (Moya-Ruiz *et al.,* 2021), being the most efficient vectors the cowpea (*Aphis craccivora,* C.L. Koch), cotton (*Aphis gossypii* Glover), and green peach (*Myzus persicae* Sulzer) aphids (Lecoq & Desbiez, 2008). WMV is also transmissible by mechanical inoculation and by seed (Wylie *et al.,* 2017). Previous studies have shown the importance of the proteins HC-Pro and CP in the transmissibility by aphids (Atreya & Pirone 1993; Peng *et al.,* 1998). The attachment of virions to the aphid stylet has been described as involving two essential interactions: (I) A highly conserved DAG motif, located in the amino-terminal domain of the CP binds to the carboxyl-terminal domain of the HC-Pro protein, and, (II) the amino-terminal domain of the HC-Pro protein directly interacts with the aphid stylet, facilitating the binding by creating a reversible

"molecular bridge" between CP and aphid mouth parts (Gadhave *et al.,* 2020; Gal-On, 2007).

The epidemiology and genetic variation of WMV has been studied in France in higher detail than in other countries. WMV's was first described in France in 1974 in melon crops (Luis Arteaga *et al*., 1976). Despite its high prevalence, WMV was considered a minor pathogen as it caused mild symptoms, such as mottling and mosaics on pumpkin leaves, but without serious impacts on yield or product quality (Lecoq, 1992). However, since 1999, new WMV strains have emerged that induce more severe symptoms in cucurbit crops, which have led to serious yield and quality losses due to severe discoloration and deformation of leaves and fruits (Desbiez *et al.,* 2009). Indeed, the genetic variability of the Nib-CP region of WMV isolates differentiated into three distinct groups with intergroup nucleotide divergences of around 10 % (Desbiez *et al.,* 2007). Two of these groups (referred to as the classic group or group 1 (CL or G1) and group 2 (G2)) have been present in France and Spain for more than 30 years (Moreno et al., 2004; Lecoq & Desbiez, 2008), with the G1 being the predominant group until 1999. Group 3, also referred to as the "emerging" group (G3 or EM), includes isolates that have caused severe symptoms since 1999 (Desbiez *et al.,* 2011). Further analyses showed that new genetic variation within the WMV groups can be found (Moya-Ruiz *et al*., 2021). These analyses differentiated the isolates from the CL group in two different subgroups (CL-A and CL-B) (Desbiez *et al.,* 2009). In addition, genetic variability analysis of the CP showed that the WMV isolates of the emerging group clustered into the four genetic subgroups EM1 to EM4, which were associated with isolates distributed in southern France between 2004 and 2007 (Desbiez *et al.,* 2011). The subgroups EM1 and EM2 were present in certain regions of France where EM3 and EM4 were absent, and vice versa. This suggested that the introduction of the EM subgroups occurred in different locations (Desbiez *et al.,* 2009). Different Mediterranean countries have reported an increase in the distribution of the EM group that is replacing the CL and G2 groups that often coexisted in mixed infections (Bertin *et al.,* 2020; Joannon *et al.,* 2010). These results showed that the EM group had an epidemiological advantage over the CL and G2 groups. The transmission efficiency of the three

WMV groups by aphids was not significantly different, however the transmission of mixed CL–EM infections was significantly higher than that of CL alone, which highlighted the transmission efficiency of the EM group (Lecoq *et al.,* 2011).

2. OBJECTIVES

The general objective of this thesis is to understand virus-plant interactions and transmission dynamics at the landscape scale as a first step to understand plant virus emergence. To obtain the necessary information HTS and big-data technologies have been used.

Specific objectives developed in this thesis towards attaining the general objective above are:

1. Evaluation of the variation in plant community sampled among the multiple study sites for each of four habitat types.

2. Characterisation of the virome in an agricultural landscape in central Spain.

3. Analysis of the host range, incidence, and genetic diversity of a generalist virus as a case study.

3. MATERIAL AND METHODS

3.1. Field work and sample collection

3.1.1. Identification of sampling sites

Field work was mostly done before the author of this thesis joined the project, during 2015 - 2017. However, as the strategic sampling design formed the basis of this study, it is described here.

The field sampling was conducted in different sites selected *a priori,* taking them as replicates of four habitats, Oakwood, Q; Wasteland, E; Edge, L; and in Crop, Cr. In total, 23 different sites were sampled, of which 4 belong to Oakwood, 4 to Wasteland, 4 to Edge and 11 to Crop of which 4 sites correspond to melon fields, M; two sites to barley fields, H; three sites to cauliflower fields, C; and two to maize fields, Z. The sites were distributed in an area of 2303 km^2 at the centre of Spain in the region of Vega del Tajo-Tajuña. The list of locations of the sites is shown in **Table 3.1** and their location in the Iberian Peninsula is represented in **Figure 3.1.** Among the different habitats and sites sampled there is not a spatial correlation.

Collections at the wilder habitats (Wasteland and Oakwood) were done in two different seasons: spring, (P = “primavera”); and autumn, (F = “fall”), when the plant diversity was maximum and minimum, respectively. Anthropic habitats (Crops and adjacent Edge sites) were sampled according to the crop season. In melon and maize Crop collections were done in summer, (V = “verano”); barley Crop in spring and cauliflower Crop in autumn. Collections from 1 site of Crop and 2 sites of Edge could not be repeated in 2016 and 2017, either because the owners denied access to the fields or because the plant communities of the edges had been destroyed by ploughing. These affected sampling in melon Crop sites in the summer of 2016 and 2017, and from the Edge sites in the summer and spring of 2015 and 2016. The seasons when each habitat was sampled and the number of times it was visited is shown in **Table 3.2.**

Table 3.1. Locations of the sampling sites visited during the years 2015, 2016 and 2017 in Spain.

Location[1]	Site Code[2]	Province	Habitat [3]	Latitude (N)	Longitude (W)	Elevation (m)[4]
Aranjuez	C1	Madrid	Cr	40.051302	-3.593308	490
Aranjuez	C2	Madrid	Cr	40.043193	-3.599064	492
Colmenar de Oreja	H1	Madrid	Cr	40.049330	-3.477574	505
Villarejo de Salvanés	H2	Madrid	Cr	40.085899	-3.214026	561
Carabaña	H3	Madrid	Cr	40.245852	-3.253943	591
Villarrubia de Santiago	M1	Toledo	Cr	40.031840	-3.345220	557
Colmenar de Oreja	M2	Madrid	Cr	40.050578	-3.503040	501
Zarza de Tajo	M3	Cuenca	Cr	40.045900	-3.125000	506
Santa Cruz de la Zarza	M4	Toledo	Cr	40.032298	-3.182801	694
Colmenar de Oreja	Z1	Madrid	Cr	40.050052	-3.476076	488
Zarza de Tajo	Z2	Cuenca	Cr	40.044720	-3.131000	680
Villarrubia de Santiago	L1	Toledo	Ed	40.031850	-3.345300	557
Colmenar de Oreja	L2	Madrid	Ed	40.048758	-3.476462	506
Zarza de Tajo	L3	Cuenca	Ed	40.050352	-3.125794	508
Carabaña	L4	Madrid	Ed	40.245459	-3.252617	592
Villacañas	Q1	Toledo	Oa	39.595520	-3.391830	694
San Martín de la Vega	Q2	Madrid	Oa	40.225155	-3.530613	617
Villar del Olmo	Q3	Madrid	Oa	40.320294	-3.205903	668
Mondéjar	Q4	Guadalajara	Oa	40.345348	-3.137145	677
Colmenar de Oreja	E1	Madrid	Wl	40.059138	-3.500323	514
San Martín de la Vega	E2	Madrid	Wl	40.234966	-3.536191	524
Fuentidueña de Tajo	E3	Madrid	Wl	40.089637	-3.196057	549
Guadalajara	E4	Guadalajara	Wl	40.494167	-3.131139	881

[1]Name of the nearest municipality.
[2]The population names were: Q = Oakwood, E = Wasteland, L= Edge, C=Cauliflower field Crop, H=Barley field Crop, M=Melon field Crop, Z=Maize field Crop.
[3]The habitats were named with the acronym of the habitat name in English: Oa=Oakwood, Wl=Wasteland, Ed=Edge, Cr=Crop.
[4]Meters above sea level.

Table 3.2. Seasons in which the different habitats were sampled and the number of times they were visited.

Habitat		Season	Nº of collections[1]
Crop	(Melon[2])	Summer	10
	(Maize)		4
	(Barley)	Spring	6
	(Cauliflower)	Autumn	4
Edge		Summer	7
		Autumn	8
		Spring	7
Oakwood		Summer	8
		Autumn	8
Wasteland		Summer	8
		Autumn	8
Total			**78**

[1] Number of collections sampled in each season of each habitat during the years 2015-2017.
[2] Crop field sampled.

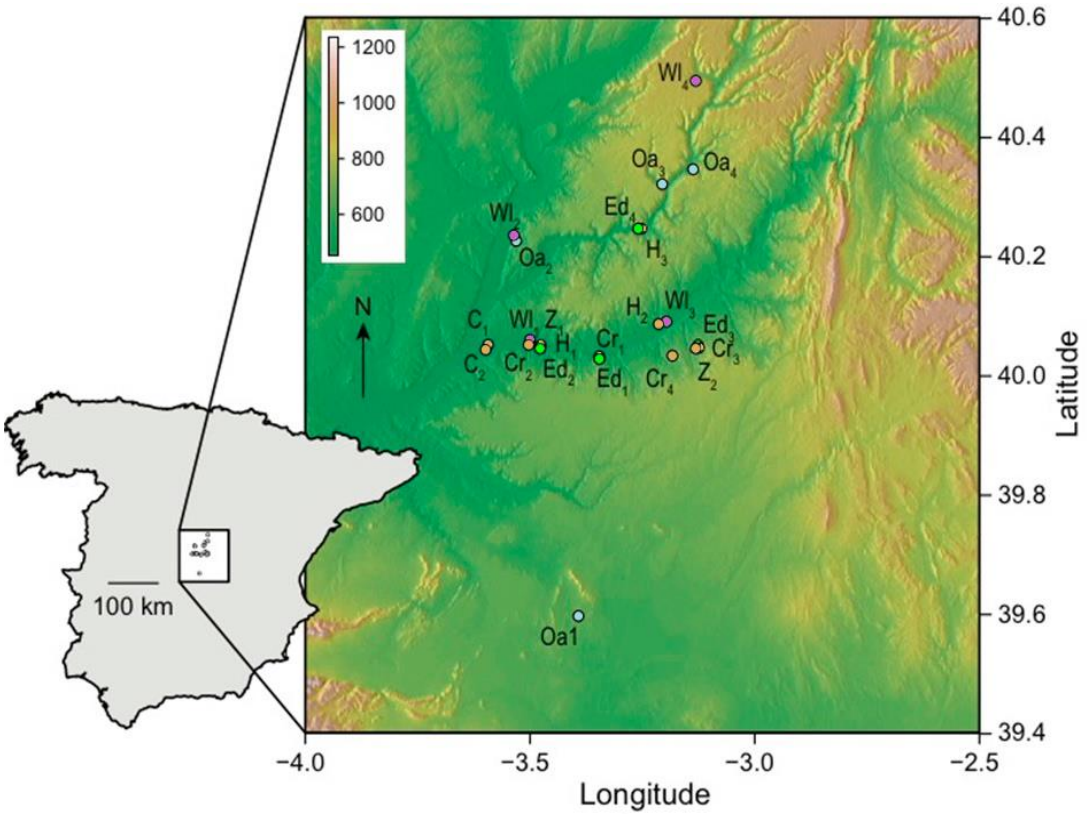

Figure 3.1. Geographic distribution of the sampling sites of this study. For each site, habitats are indicated: Cr (orange dots), Ed (green dots), Oa (blue dots), Wl (purple dots). The color scale indicates the elevation.

3.1.2. Sampling design and sample preservation

Plant samples were collected using a systematic sampling design that minimized the bias in species representation and relative abundance at each site (Choset, 2000). An area of 25 m x 25 m was georeferenced and marked out in each site of Oakwood and Wasteland. In Edge and Crop, a 25 m x 2 m area at each site was georeferenced. A boustrophedonic transect method (a line taken alternately from right to left and from left to right, and so on) was used at all sites except for Edge, which have highly linear configurations. Samples of leaves were harvested from plants at fixed points along the transect, regardless of whether they showed symptoms or not. One or more leaves were taken from different parts of an individual plant that represented a single sample of a collection. The rationale for our sampling strategy was based on the variation of the configuration and area of the different cover types associated with each habitat. Therefore, each habitat required different sampling efforts to ensure an unbiased estimate of species richness and relative abundance (Foster *et al.,* 2020). In Crop and Edge, 50 plant samples were collected per site at each resampling, and in Oakwood and Wasteland, 150 plant samples were collected per site at each resampling. All samples were placed into clear plastic bags and kept at 4ºC for transport to the laboratory. The samples were then processed in the laboratory by removing soil and any other non-plant material, to be placed in sealable plastic bags and frozen in liquid nitrogen and stored at -80°C. During each collection, an herbarium comprising all plant species sampled was compiled as a reference for subsequent taxonomic identification by Professor Rosario G. Gavilán (Universidad Complutense de Madrid, Madrid, Spain).

Rarefaction analyses (Hurlbert, 1971) were performed to assess the sample representation and sampling effort in each habitat. Given the number of individual samples of each species at each collection (n = 78) rarefaction analysis allows the calculation of the number of species (species richness) expected in smaller subsamples of n individuals. By means of making predictions for successively smaller subsamples of individuals in a collection, each community can be represented by “rarefaction curves” that illustrate the number of species richness as a function of the number of individual samples in each collection.

If the relationship between sampling effort and species richness reveals a steadily increasing curve, we may assume our sampling effort has been too small, so wider sampling may reveal more species (Nielsen *et al.,* 2007). On the other hand, if the value appears stable, or levels off and reaches an asymptote, one can say that there has been sufficient sampling for a determined collection, as an additional sampling would not reveal more species (Nielsen *et al.,* 2007). To perform rarefaction analyses the R package *vegan* (Oksanen *et al.,* 2015) was used.

3.2. Characterisation and comparison of plant communities

3.2.1. Characterisation of plant communities

The assess the homogeneity of relative species abundance estimates among the collections from sites of each habitat and whether each *a priori* habitat is supported by differences in plant community, a Detrended Correspondence Analysis (DCA) was conducted. This ordination method distinguishes communities based on the species occurrence and abundance. As such, species of high abundance are expected to influence the relationships among communities, which could particularly affect the relationship between Crop and the other habitats. Thus, this relationship was assessed by performing a DCA with the abundant crop species omitted from each Crop community and removing the Oakwood and Wasteland communities. To perform this analysis, the R package *vegan* was used.

3.2.2. Assessment of diversity estimators

The estimation of the plant species diversity was conducted to compare the habitats sampled and to assess ecosystem simplification.

It is known that the type of diversity estimator used can condition the interpretation of results (McLeish *et al.,* 2017). Thus, two diversity estimators were compared. One was the Tsallis entropy estimator of diversity (S_q*;* Mendes *et al., 2008*), which is sensitive to rare species that were expected from

incomplete samples and that generally characterise wild communities (McLeish *et al.,* 2021). The other estimator was an extrapolation (i.e., prediction) method that relies on sample completeness and not equal sample sizes (Hsieh *et al.,* 2016) to generate an asymptotic estimator (D_{AE}; Chao & Jost, 2012) of Shannon diversity (McLeish *et al.,* 2021)

In addition, the multiple collections at each site were part of a repetitive resampling strategy that enabled an assessment of variation in the community of each site and improved the sampling effort at each site. Furthermore, multiple sites from the same habitat were sampled as in a replication strategy expected to capture variation among the sites sampled of a given habitat. The way to treat these observations, aggregating or not aggregating data for analyses, is known to influence the diversity estimators (McLeish *et al.,*2021). Thus, to evaluate the effect of data aggregation in the quantification of differences among habitats, data were analysed with aggregation (at site level) or without aggregation (at collection level). Moreover, data aggregation by site will remove the temporal signal from data. To estimate how important was this effect we estimated the mean and standard deviations in seasonal diversity estimators among the habitats to infer the absence or presence of temporal factors. To perform these analysis R packages *iNEXT* (Hsieh *et al.,* 2016) and *vegan* were used.

3.3. RNA extraction and nucleotide sequence determination

3.3.1. Total RNA extraction from plant tissues

It is expected that high heterogeneity in the physical condition of individuals derived from environmental samples and differences between species potentially influence the performance of total RNA extraction procedures. The optimisation of the procedures used was conducted to obtain high quality RNA preparations necessary for nucleotide sequencing and virus detection. Three tissue homogenisation methods were compared to get homogenates from 0.3 g of plant tissues taken from each individual: (I) mortar with a pestle and glass powder, (II) Eppendorf tube without glass powder, and (III) Eppendorf tube with

glass powder. All tissue homogenisation methods were carried out using liquid nitrogen to prevent RNA degradation. The plant tissue homogenates were kept in liquid nitrogen until their subsequent extraction. Homogenates thus obtained were extracted by three different procedures implemented in commercial kits: (a) Direct-zol™ RNA MiniPrep Plus (Zymo Research, Irvine, CA, USA), (b) RNeasy Mini Kit (Qiagen, Hilden, Germany), (c) Agilent Plant RNA Isolation Mini Kit (Agilent Technologies, Santa Clara, CA, USA) **(Figure 3.2).**

The Direct-zol™ RNA MiniPrep Plus kit provides high-quality RNA purification directly from samples using TRIzol® reagent to lyse the homogenised samples and inactivate ribonucleases, while the RNeasy Mini Kit and Agilent Plant RNA Isolation Mini Kit use an extraction solution with β-mercaptoethanol. This extraction solution contains guanidine-thiocyanate (in the case of RNeasy Mini Kit) and guanidine hydrocholoride (in the case of Agilent Plant RNA Isolation Mini Kit) that immediately inactivates ribonucleases to ensure purification of intact RNA. In addition, RNeasy Mini Kit and Agilent Plant RNA Isolation Mini Kit incorporate two types of membranes; one to remove cell debris and other to bind total RNA, while Direct-zol™ RNA MiniPrep Plus kit including only the total RNA binding membrane.

For the three kits, the obtained RNA pellet was re-suspended in 50 µl of nuclease-free water. The total RNA concentration (ng/µl) and purity (according to the absorbance ratios 260/280 nm and 260/230 nm) of each extract were determined by optical density (OD) measurements, using a NanoDrop ® ND-1000 spectrophotometer (Thermo Scientific, Waltham, MA, USA). In addition, the integrity of the extracted RNA was verified by gel electrophoresis in 1% Agarose, 1x TAE buffer (40mM Tris, 20mM acetic acid, 1 mM EDTA) (Sambrook *et al.,* 1989) run for 90 min at 90V, and stained with ethidium bromide (EtBr).

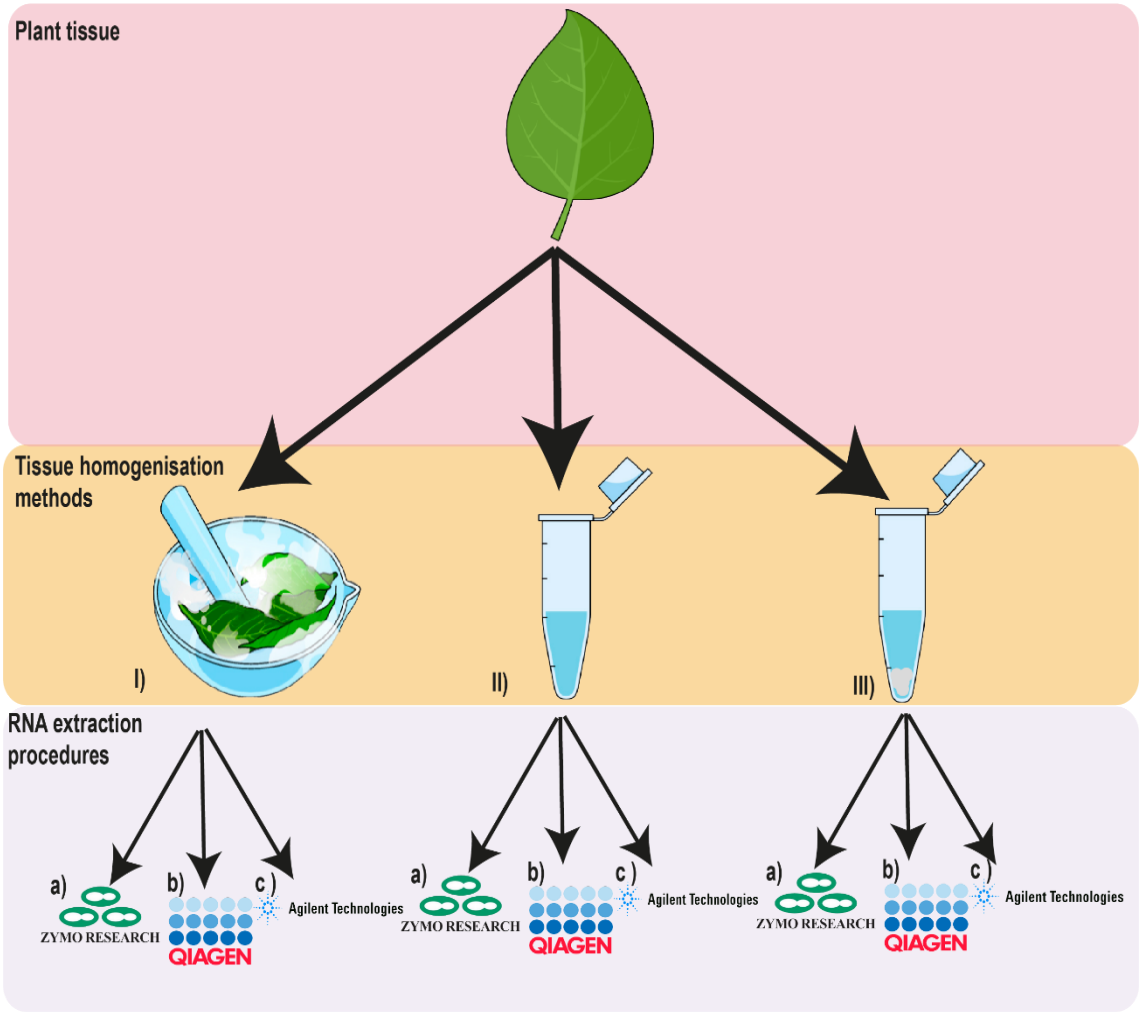

Figure 3.2. Evaluation of different tissue homogenisation methods and RNA extraction procedures. Plant tissues were ground in three different ways: I) mortar with a pestle and glass powder, II) in an Eppendorf tube without glass powder, III) in an Eppendorf tube with glass powder. RNA extraction procedures implemented in three different kits were evaluated for each homogenate: (a) Direct-zolTM RNA MiniPrep Plus, (b) Rneasy Mini Kit (c) Agilent Plant RNA Isolation Mini Kit.

Generalized linear mixed models (GLMM) were used to analyse the concentration and absorbance variation (response variables), according to the following factors: tissue homogenisation methods and RNA extraction procedures (considered as fixed factors) and plant species (considered as random factor) and their interaction. For each response variable a separate model was carried out. To assess the variance explained of each model performed with each response variable, we calculated R^2 values for each model using the methods of Nakagawa & Schielzeth (2013). We calculated the marginal

R^2 which describes the variance explained by the fixed factors, and the conditional R^2 which is concerned with the variance explained by both the fixed and random factors (Nakagawa & Schielzeth 2013). The difference between each fixed factor classes was analysed by a Least Significant Difference (LSD) **(see section 3.5.4.)**.

Once a tissue homogenisation method and RNA extraction procedure was selected, total RNA extraction was carried out from 0.3 g of plant tissues. However, the quality of the RNA extracts still was unsatisfactory for some species and/or individual samples, for which an extraction procedure using of the Cetyltrimethylammonium (CTAB)–Polyvinyl pyrrolidone (PVP) method (Chang *et al.,* 1993) was assayed, which still proved to be inefficient with a subset of samples. Finally, the Spectrum ™ Plant Total RNA Kit was ultimately used (Sigma-Aldrich, St. Louis, MO, USA), which was not available while assaying the previously described kits. The Spectrum ™ Plant Total RNA Kit uses an extraction solution with β-mercaptoethanol and two types of membranes. The second membrane differs from the membranes provided by other kit types in that it removes almost all secondary metabolites before the RNA binds to the binding substrate or is even washed away. Additionally, Spectrum ™ Plant Total RNA Kit uses a lysis solution with β-mercaptoethanol, that inactivates the ribonucleases and removes them during the washing steps.

The obtained RNA extracts were used for nucleotide sequence determination by high throughput sequencing (HTS) and Sanger method.

3.3.2. Rationalisation of plant species for RNA extraction and HTS analysis

The tissue homogenisation method and the commercial kit which together yielded total RNA extracts with better concentration, purity, and integrity **(see section 3.3.1.)**, were selected to perform the RNA extraction from plant tissue of the field samples for subsequent sequencing.

The inclusion of rare plant species, defined here as those for which less than 5 individuals were collected in at least three of the four habitats, in analyses

of host-parasite interactions may bias the estimation of metrics such as detection of infection and estimation of prevalence (Blüthgen *et al.,* 2008; Blüthgen, 2010; Vázquez *et al.,* 2005). For example, when a plant species is only represented by a single observation (singletons) it is inevitably assigned a single link and considered a specialist, inevitably inflating the estimation of the prevalence. This bias makes the removal of rare plant species a common practice in infection network studies (Blüthgen *et al.,* 2008; Blüthgen, 2010; Vázquez *et al.,* 2005). Therefore, RNA was not extracted from rare plant species, which were not included in sequence analyses. In addition, the omitting of rare plant species contributed to the practicality of the study. The possible effect of removing rare plant species on the categorisation of the different habitat was assessed comparing the whole plant species and the reduced plant species sets using DCA and Kendall's tau coefficient (τ). To perform the Kendall's tau coefficient the R package *wPerm* (Weiss, 2015) *was* used.

Additionally, to rationalise HTS efforts we focused on the frequent plant species collected between the summer of 2015 and spring of 2016 that occurred in more than one habitat, or, if occurring in only one habitat were abundant and characteristic of a specific plant community (e.g., *Quercus ilex* L. in Oakwood). The plant communities of the HTS plant species selected were assessed by a DCA analysis.

3.3.3. Library preparation and high throughput sequencing

For HTS, RNA extracts form individual samples were pooled in equimolar amounts to obtain single HTS libraries. HTS libraries involved extracts from samples of the same plant species, collection and site, and included a maximum of seven individual RNA extracts. Previous simulation analyses in our group had proven that pooling RNA extract of up to seven individuals was appropriate to detect with sufficient resolution by HTS a low titre viral RNA with a proportion of contaminating nucleic acids (Dubay, 2017). A table providing the reference codes of the read libraries, plant species and family, the habitat and site from which they came, and the number of RNA extracts that comprising each HTS library, is listed

in **Annex, Table A3**. The concentration, purity, and integrity of all pooled extracts was checked before being sent for HTS, to compare with the pre-sequencing quality control measures undertaken by the HTS supplier.

The HTS of RNA extracts was outsourced to the Centre de Regulació Genòmica (CRG, Barcelona, Spain; http://www.crg.eu/). Sequencing library preparation was performed using Truseq RNAseq total kit with an additional ribosomal RNA depletion step using the RiboZero kit. Pair-end sequencing of either 125 or 150 base pair reads (2x125; 2x150 bp) was performed on either the HiSeq 2500, 3000 or 4000 Illumina platforms. Electropherograms of each library, RNA concentration, and RNA integrity numbers (RIN) were provided by the supplier prior to sequencing. Only those samples that met the supplier's quality criteria were selected for sequencing. Shotgun sequence libraries based on DNA reference sequences were simulated with the Grinder software (Angly *et al.,* 2012) to estimate the number of reads required to detect viruses that are part of complex communities present in environmental samples. It was estimated that 8.0×10^6 reads were appropriate to detect low titre viruses in sequencing libraries with a proportion of contaminating nucleic acids (Dubay, 2017*).*

3.3.4. Nucleotide sequence determination by Sanger method

The nucleotide sequence of the coat protein (CP) gene of *Watermelon mosaic virus* (WMV) was determined using Sanger's method. All RNA extracts that were positive for WMV by a reverse transcription-polymerase chain (RT-PCR) test **(see section 3.4.5.)** were considered candidates for sequencing. However, to avoid an unbalance number of WMV-CP sequences for some host plant species or collections, the CP gene was sequenced in a maximum of five PCR products of a specific host plant species in a determined collection.

The DNA amplicons were purified using a kit designed and manufactured by the sequencing company, based on magnetic particles, and the nucleotide sequence was determined using the ABI 3730XL DNA Analyzer sequencer (Applied Biosystems, USA), with the forward and reverse primers **(see section 3.4.5.)**, using the sequencing service of the company *Stabvida* (Caparica,

Portugal). The chromatograms obtained from each sequence were read using the MEGA X programme (Kumar *et al*., 2018). Consensus sequences were obtained from the forward and reverse primers. The nucleotide sequences of the WMV from the CP region are available in the NCBI database. The accession number of each sequence is listed in **Annex, Table A6**.

3.4. Sequence detection and validation

3.4.1. Trimming and storing of HTS libraries

The HTS libraries were deposited in a temporary database and then downloaded to a local server for downstream analyses. All reads provided by the supplier have a *Phred* score (metric indicating the probability that a given base will be incorrectly called by the sequencer (Ewing & Green, 1998)) greater than 30. Any adapter sequence contamination was trimmed from the read data using *cutadapt* v1.8.3 (Martin, 2011). The quality of the trimmed reads was assessed using *FastQC*. All read data files were backed-up, and copies of the original FASTQ format converted into FASTA format for use in downstream analyses.

3.4.2. Detection of viral sequences in the HTS libraries

Viral sequences were detected in the HTS read libraries using a Basic Local Assignment Search Tool (BLAST) (Altschul *et al*., 1990) query against a reference database of single-stranded RNA (ssRNA) plant virus sequences. The plant virus sequence references used in local BLAST queries were sourced from NCBI (https://www.ncbi.nlm.nih.gov/genomes/), accessed in December 2018. This database just retains one sequence reference per virus species registered in NCBI.

The large heterogeneity of the HTS libraries expected from different hosts and habitats might affect the detection of virus sequences. Also, BLAST query analysis relies on reference-based alignment, and a significant proportion of virus sequences from the HTS libraries are expected to be divergent to the virus sequence references. To retain homologous sequences that are variants of our

reference sequence and at the same time show a sufficiently high stringency to discriminate matches with query sequences from other genes (Sandhya *et al.,* 2003), the stringency of the BLAST query was modulated with a moderate expected value (*E.value*) threshold of 10^{-4}. However, even if part or all of the query is homologous with the reference, it does not guarantee that it is a true positive. Thus, to reduce the frequency of false positive virus sequence detection results obtained by BLAST queries, a series of validation steps were conducted.

3.4.3. Validation of virus sequence detections by HTS

The operational taxonomic units (OTUs) defined by the identical or variant sequences identified using BLAST queries (Rohwer *et al.,* 2018), were validated *in silico* to reduce the frequency of false positives. The presence of homoplasious or orthologous associations between the query sequences and BLAST database references can overestimate the number of detections and result in the detection of false positives due to sequence similarities between different viruses (Fitch *et al.,* 2000). The frequency of such artefacts was reduced by removing OTUs that did not have both paired-end reads matching the same reference, or that matched more than one virus species sequence from the same library read (Fan *et al.,* 2015). In addition, a minimum *query length* of 100 nucleotides (adaptor contamination produced a number of reads that were shorter than 150 nt) with a percentage of identity greater or equal to 90%, and only those queries that completely covered the reference sequence (i.e., *query coverage* equal to 100%) were retained for downstream analyses (Wall *et al.,* 2003).

Kernel density estimation (KDE) was used to assess changes in sequence similarity between BLAST queries and references using the BLAST query flags Bit-score (*S`*), before and after *in silico* validation of the virus sequence detection. The *S`* shows the quality of each match as defined by the size of the space (measured in bits) one would need to search to find as strong a match by chance (Shah *et al.,* 2018). It is expected that the application of the validation criteria will increase the mean of *S`*. KDE is a non-parametric estimator more accurate than a histogram that allows the visualisation of data that have a large amplitude, as well as smoothing the data (Ramachandran & Perkins 2013). Therefore, KDE

allows a proper visualisation of the BLAST *S`* query flags for all reads of all libraries for the evaluation of the validation steps. To perform this analysis, the R package *ggplot2* (Wickham, 2011) was used.

Additionally, for WMV, a validation step involved detection by RT-PCR using specific primers **(see section 3.4.5.)**. The results obtained by both methods were compared.

3.4.4. Evaluation of the quality of HTS libraries

The variation of quality between plant RNA extracts may negatively affect the library preparation and HTS performance, consequently decreasing HTS virus sequence detection.

To test whether the quality of the RNA extracts might bias the sensitivity / accuracy of virus sequence detection, messenger RNAs (mRNAs) known to be stable of conserved endogenous plant genes, were used as quality controls of HTS libraries. These mRNAs are expected to be present in the tissue sample and RNA extracts, and therefore subject to the same sources of variability in detection than the viral RNAs (Expósito-Rodríguez *et al.,* 2008; Huggett *et al.,* 2005). The mRNAs of the following four genes were selected for this analysis: *DnaJ-like protein, Elongation factor 1-α, Glyceraldehyde-3-phosphate dehydrogenase,* and *Ribosomal protein L2,* which have been reported previously as conserved genes with stable mRNAs (Dean *et al*., 2002*,* Expósito-Rodríguez *et al.,* 2008, Løvdal *et al.,* 2009). As a further internal control, the mRNA of a fifth gene known to be not conserved across plant species was included in the analysis: *Topoisomerase II associated protein*, a gene specific to *Arabidopsis thaliana* (Yu *et al.,* 2016). Details of each of the plant genes used to develop the controls, and their NCBI accession numbers, are given in the **Table 3.3**.

The sequences of these five plant genes were downloaded from the NCBI browser (https://www.ncbi.nlm.nih.gov/genomes/) and used to perform a BLAST query against the FASTA read files of each library. Therefore, the BLAST query was performed under a minimum *E.value* threshold of 10^{-4}, as for virus sequence detection **(see section 3.4.2.)**.

Table 3.3. Name, nucleotide number, accession number and use that is given to each plant gen for the evaluation of the quality of each library and the plant specificity of genes.

Plant gen	Nucleotides	Accession number[1]	Plant gen used as:	Plant species
DnaJ-like protein	1,566	AF124139	Positive control	*Solanum lycopersicum*
Elongation factor 1-α	4,565	X53043	Positive control	*Solanum lycopersicum*
Glyceraldehyde-3-phosphate dehydrogenase	1,285	U97257	Positive control	*Solanum lycopersicum*
Ribosomal protein L2	816	X64562	Positive control	*Solanum lycopersicum*
Topoisomerase II associated protein	3,041	AK318693	Negative control	*Arabidopsis thaliana*

[1] Accession number used to download the Genbank sequences from the NCBI.

3.4.5. Virus detection in individual RNA extracts by RT-PCR

RT-PCR was used to detect WMV in RNA extracts from samples of HTS libraries that were positive for WMV or in extracts not included in HTS libraries. The test was designed to amplify the whole CP gene of WMV (843 nucleotides).

The complementary DNA (cDNA) was synthesized in a reaction mixture containing 1 µg of total RNA, 10 mM deoxynucleotide triphosphates (dNTPs), 0.5 µg Oligo(dT)$_{18}$ primer (Life Technologies, Carlsbad, CA, USA), and the remaining water to volume make up to 10 µl. This mixture of 10 µl was incubated at 65°C for 5 minutes and 200 units (U) of the reverse transcriptase (RT) of Moloney murine leukemia virus (M-MLV) (Promega Corporation, Madison, WI , USA), 2 µl of reverse transcription buffer (250mM Tris-HCl (pH 8.3 at 25°C), 375mM KCl, 15mM $MgCl_2$, 50mM DTT), 0.2 µl ribonuclease inhibitor (40 U/ml) and the remaining water to volume make up to 20 µl were added. The reaction was carried at 42°C for 60 minutes, and 65ºC for 20 minutes. The resulting cDNA was used as a template for the PCR.

To reduce the fraction of false negative reactions that could be due to the variation in the quality of RNA extracts, two sets of primers were used for the PCR reaction **(Table 3.4.)**. Nucleotide sequences of 100 WMV accessions were downloaded from the NCBI (https://www.ncbi.nlm.nih.gov/) and aligned using the MUSCLE tool v3.8.31 (Edgar *et al.,* 2004) of MEGA version 6 (Tamura *et al.,* 2013). On this alignment, F2 and R4 primers were designed to amplify a region of 1032 nucleotides between position 8,776 (in the cistron for the nuclear inclusion b, NIb) and 9808 (in the 3' non-translated region), positions numbered according to WMV NCBI accession NC_00626). The amplicon fully covers all 843 nucleotides of the CP cistron. The WMV 826-CPF and WMV 826-CPR primer set were designed to amplify a region of 831 nucleotides between the CP cistron from positions 8975 to 9806 (positions numbered as in accession NC_00626). The OligoAnalyzer® tool (http://eu.idtdna.com/ calc/analyzer) was used to select the appropriate melting temperature and to identify potential for the formation of self-dimers and hairpins. The PCR was performed in a reaction containing 1 µg of cDNA, 1X Taq polymerase buffer with $MgCl_2$, primer mix (0.2 µM), dNTPs (2.5 mM) and 0.05 µl of enzyme Taq polymerase (5U / µl) (Roche Diagnostics) to make up a total volume of 20 µl. The following PCR conditions were used: 94°C for 2 minutes, 35 cycles of: 94°C, 30 seconds, 55°C, 30 seconds and 72°C, 35 seconds. Finally, an extension of 72ºC for 10 minutes was carried out.

Table 3.4. Sequence of the specific primers used to amplify the CP gene of WMV.

Primer	Primer sequences (5'-3')	Position[1]
F2	GTGGCTTCTAAGCAAAGAYG	8757-8776
R4	GRCTAGGTAAACTGGTCACAG	9788-9808
WMV 826-CPF	TNGARAATTTGGATGYAGG	8757-8975
WMV 826-CPR	CGACCCGAAATGCTAACTG	9806-9824

[1]The position in nucleotides according to the sequence of WMV (accession no. NC_006262).

As positive control for the RT-PCR, RNA purified from WMV isolate m-116, kindly provided by Professor Miguel A. Aranda (Centro de Edafología y Biología Aplicada del Segura, CEBAS-CSIC), was used. Young zucchini (*Cucurbita pepo* L.*)* leaves were mechanically inoculated with crude sap from the infected lyophilized leaf tissue in 1 mM sodium phosphate buffer pH 7.2, with 0.2% Na-diethyldithiocarbamate (DIECA). The inoculated plants were maintained in a greenhouse for 21 days (24-27ºC, and 16 hours of light), after which they were harvested. Total RNA of the infected zucchini plants was extracted using Agilent Plant RNA Isolation Mini Kit, and the viral particles purified according to the protocol described in (Purcifull *et al.,* 1979*)*. The viral RNA was extracted from the purified viral particles using a 1% SDS treatment and phenol-chloroform extraction method with final ethanol precipitation, according to the protocol described in (Purcifull *et al.,* 1979). Viral RNA was then resuspended in DEPC (Diethyl pyrocarbonate)-treated sterile water. Quantification of the purified RNA concentration (ng/µl) and purity (with the absorbance ratios 260/280 and 260/230) was performed using OD measurements and then stored at 80ºC.

In addition, negative controls comprising either the reagent mixture without the RNA template, or water only, were included in the RT-PCR.

The PCR amplicons were resolved by electrophoresis in 2 % agarose gels in 1X TAE buffer (40mM Tris, 20mM acetic acid, and 1 mM EDTA) (Sambrook *et al.,* 1989) run for 90 minutes at 90V, and were visualised under UV light after ethidium bromide (EtBr) staining.

3.5. Computational and statistical analyses

3.5.1. Host range, incidence, and population genetics analyses

It is expected that the RT-PCR could produce false negative results, associated with the virus titre in the total RNA extract. Therefore, our approach combined data of HTS and RT-PCR to estimate WMV host range (defined as the number of host species used by a pathogen (McLeish *et al.,* 2019)). First of all, we used HTS data to detect a wider range of the host species of WMV. Thereafter

we used RT-PCR to validate HTS results in addition to target the WMV genetic regions of interest (McLeish *et al.,* 2020b). WMV host range was determined as the number of plant species with samples that were positive for WMV by RT-PCR.

The number of RT-PCR positive individual plant samples allowed to calculate the incidence, defined as the percentage of infected plant individuals over the total plant individuals (Campbell & Neher 1994)). The incidence was measured as $\frac{\text{number of WMV infected plant individuals}}{\text{total number of analysed plant individuals}}$. The differences of incidence among habitats were assessed using permutation *F*-tests and *t*-tests. The Benjamini & Hochberg *P*-value adjustment method for multiple comparisons and 1,000 permutations were used in all tests. The permutation tests are non-parametric and do not assume normally distributed errors. These tests were conducted with the R package *RVAideMemoire* (Hervé, 2020). It is expected that the difference in the host relative abundance present in each habitat may affect virus incidence. The possible association between relative host abundance and virus incidence was evaluated for each habitat using a correlation test. Non-parametric correlation analyses were performed, as appropriate for small sample sizes and valid for all types of distributions. Test was conducted with the R package *stats* (R Core Team, 2020).

The nucleotide sequences of the WMV-CP cistron were used in population genetic analyses. Sequences were aligned using MUSCLE v3.8.31 (Edgar, 2004) and alignments were adjusted manually. The association with either habitat or host plant species analysed in this thesis were used to define the WMV populations and thus conducting the different population genetic analysis.

Population genetic diversity of WMV was calculated in terms of nucleotide diversity (π) and haplotype diversity (H_d) (Tajima, 1983). Genetic diversity estimates were calculated with *Arlequin* v.3.11 programme (Excoffier *et al.,* 2005) and their standard errors based on 1000 bootstrap replicates. A permutation procedure was carried out to test whether the observed nucleotide/haplotype diversity differences between the populations are greater than expected by

chance, as implemented with the R package *genetic_diversity_diffs* (Alexander, 2017).

Genetic diversity of WMV-CP may be modulated by selection associated to habitat or host. Selection pressure was analysed by the mean pairwise non-synonymous (d_N) to synonymous (d_S) site ratios (ω) using the MegaX programme (Kumar *et al.,* 2018) using Nei-Gojobori method (Nei *et al.,* 1986).

Differences in population genetic structure of WMV according to habitat or host of origin were assessed by as analysis of molecular variance (AMOVA) implemented in *GENALEX V6.1* (Peakall & Smouse, 2006). AMOVA provides estimates of variance components and *F*-statistic analogs, such as *Phi-statistics* (Φ_{ST}) which reflect the correlation of haplotypic diversity at different levels of hierarchical subdivision (Excoffier *et al.,* 1992). In addition, the pairwise Φ_{ST} was conducted to test population differentiation between habitats. The statistical significance of this analysis was obtained by performing a bootstrap of 1000 replicates for the different groups of habitats and hosts.

The association between the phylogenetic relationships of WMV populations with habitat and host species was analysed using different association indices obtained by Bayesian tip association significance test (*BaTS*) (Parker *et al*., 2008). A random sample of 1,000 post-burnin trees from the phylogenetic analysis **(see section 3.5.2.)** was used to account for phylogenetic uncertainty expected at the population level. In the *BaTS* approach, the data are repeatedly simulated under the null hypothesis that characters at the tips are randomly distributed across the phylogeny. We designated either a habitat or a host species trait for each taxon (tip) of the posterior set of trees (PSTs) in two separate analyses. The analysis generates the association index statistics (AI) and parsimony statistics (PS) which measure the degree to which sequences share previously assigned trait (habitat and host species). The inverse value of PS relates to the strength of association of the trait in the terminal branches of the phylogeny. According to the value, it implies a gain or loss of the trait. The AI value also represents the phylogeny-trait association; however, this value takes

into account any imbalance in the phylogeny topology. The MC value assess association of the viral sequences for each habitat.

3.5.2. Phylogenetic analyses

The presence of recombination events could interfere with phylogenetic inference. Hence, recombination breakpoints, which are associated with recombination events (Martin *et al.,* 2015), were analysed using the RDP4 package (http://web.cbio.uct.ac.za/~darren/rdp.html) that implements different methods for detecting recombinants based on different assumptions (Posada, 2002). The methods used to detect recombination events were RDP (Martin & Rybicki, 2000), GENECONV (Padidam *et al.,* 1999), Bootscan/Recscan (Martin *et al.,* 2005) , MaxChi (Maynard Smith, 1992), Chimaera (Posada & Crandall, 2001), SiScan (Gibbs *et al.,* 2000) and 3SEQ (Lam *et al.,* 2018). The default setting and a Bonferroni corrected *P-value* cut-off of 0.01 were used (Martin et al., 2015). Only recombination breakpoints identified by at least four of the methods mentioned above with an associated $P<10^{-3}$ were considered significant.

A phylogenetic inference based on the nucleotide sequence of the CP cistron was conducted. Selection of the best-fit substitution model was done using the Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) implemented with the *modelTest* function (Posada, 2008) of the R package *phangorn* (Schliep, 2011). In addition of the CP gene sequences determined in this work, those from data bases exemplifying the "classic" and "emerging" genetic groups of WMV proposed by Desbiez *et al.* (2009) were included in the phylogenetic inference. The *Soybean mosaic virus* sequence, closely related to WMV (Desbiez *et al.,* 2007) was used to root the phylogenetic tree. The accession numbers of the reference sequences included in the phylogenetic analysis are shown in **Table 3.5.**

Table 3.5. Accession number, specimen code and genetic group to which the added external sequences belong.

Accesion number[1]	Genetic group [2]
EU660578	Classic isolate-A
EU660589	Classic isolate-B
EU660587	G2 isolate
EU660581	Emerging isolate 1
EU660583	Emerging isolate 2
EU660586	Emerging isolate 3
EU660585	Emerging isolate 4
D00507	*Soybean mosaic virus*

[1]Accession number used to download the Genbank sequences from the NCBI.
[2]Genetic group established according to Desbiez *et al.,* 2009.

A Bayesian maximum credibility consensus phylogeny was inferred using MrBayes 3.2.7a (Huelsenbeck *et al.,* 2001). Four Monte Carlo Markov Chains (MCMC) were run for 10 million generations and sampled every 1,000 generations. Posterior probabilities and mean branch lengths were derived from 7,500 post-burnin trees. Convergence and posterior parameter distributions were assessed using the MCMC Tracer Analysis Tool v1.7.1 (Rambaut *et al.,* 2014). The phylogenetic trees were visualized using the *FIGTREE* v1.4.4 programme (Rambaut & Drummond, 2012).

To assess the unresolved phylogenetic relationships expected from low-level genetic divergences and contextualize genetic structuring of haplotypes among the habitats and host species a haplotype network was inferred. The inference and visualisation of the genetic relationships among intra specific sequences was performed following a network representation conducted by the *HaploNet* function from the R package *pegas* (Paradis, 2010). The function implements a maximum parsimony approach, which constructs the haplotype network using the infinite sites model (Templeton *et al.,* 1992).

3.5.3. Virus richness relationships with habitat and host

To evaluate the relationship between the virus OTUs detected and habitat across host species relative abundance, two complementary methods were conducted: RLQ (Dolédec *et al.,* 1996) and a fourth-corner analysis (Legendre *et*

al., 1997). The RQL ordination method relates three data matrices. The first (R), relates each site to a particular habitat in a binary matrix (site x habitat). The second matrix, (L) shows host plant species abundances per site (species x site). The third matrix, (Q), shows the virus OTUs detected per host plant species (virus taxa x plant species) (Dolédec *et al.,* 1996)**.**

The fourth-corner analysis is a two-step procedure (Dray & Legendre, 2008), which tests the relationship between virus OTUs and habitat in explaining host species relative abundance using a randomisation test (with 9999 random permutations). The null hypothesis is that the R and Q tables are independent of each other, and the alternative hypothesis is that both habitat and virus richness influence host species abundance (Dolédec *et al.,* 1996). This multiple testing may increase the incidents of false significant associations (i.e., Type I errors). To correct multiple tests, procedures were conducted using a false discovery rate method for adjusting *P*-values (FDR, Benjamini & Hochberg, 1995). To perform this analysis, the R package *ade4* (Dray *et al.,* 2007) was used.

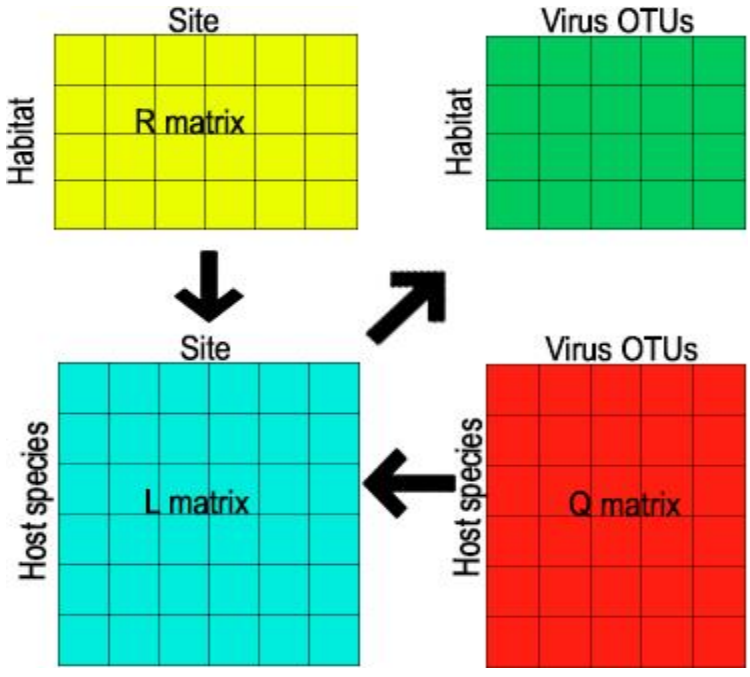

Figure 3.3. Graphical representation of the fourth-corner method (Dray & Legendre, 2008). This method tests the relationship between virus OTUs (Q matrix) and habitat (R matrix) mediated by host species relative abundance (L matrix).

3.5.4. Statistical significance tests

Shapiro-Wilk tests was performed (Sokal & Rohlf, 1995) to test the normality of each variable. Homocedasticity was tested using the Levene test (Sokal & Rohlf, 1995). When distributions were normal and homoscedastic, a parametric test was performed (one-way ANOVA test). When one of these two assumptions was not fulfilled, a nonparametric test was conducted (Kruskal-Wallis test, Z-test). Post hoc tests were used to determine whether variables were significantly different between groups within each class. The Tukey Honest Significant Difference test was performed using the R packages *car, dplyr*, *stats* (R Core Team, 2020).

GLMM were used to analyse the difference in the concentration and absorbance results (response variables), according to each factor involved in the RNA extraction (tissue homogenisation method, procedures implemented in different kits and plant species) and their interaction. The “tissue homogenisation methods” and the “extraction procedures implemented in different kits” were considered as fixed effects, while “plant species” was considered as a random effect: Conc. ~ Homogenisation * Kit + (Species); where Conc. is the concentration results, Homogenisation is the tissue homogenisation method used for the extraction and Kit is the extraction procedures implemented in different kits used for the extraction. For each response variable a separate model was conducted: Abs.1 ~ Homogenisation * Kit + (Species) and Abs.2 ~ Homogenisation * Kit + (Species), where Abs.1 is the absorbance 260/280 and Abs.2 is the absorbance 280/230. Variance components were determined using GLMMs by the Restricted Maximum Likelihood (REML) method (Lynch & Walsh 1998). The most suitable model for the data distribution of each response variable was analysed using the R package *rriskdistributions* (Belgorodski, 2017), selecting it according with the lowest AIC and BIC value. The GLMM was conducted under the most suitable model for the distributions to the response variable and performed using the R package *lme4* (Bates *et al.,* 2007). To determine whether the values of the variables analysed were significantly different between classes within each fixed factor, a Least Significant Difference

(LSD) analysis was performed, with *P*-value adjusted for multiple comparisons (Sokal & Rohlf, 1995) using the R package *lsmeans* (Lenth, 2016).

4. RESULTS

4.1. Sampling collection and sample effort assessment

Plant samples were collected over two years (2015-2016 and 2016-2017) in four different habitats. The four habitats were nominated *a priori* to represent dominant plant communities of the agricultural ecosystem of the centre of Spain. These habitats differed in the degree of human intervention, from lower to higher: Oakwood (Oa), Wasteland (Wl), Edge (Ed) and Crop (Cr). The Oakwood represents the evergreen oak forest communities that are the region's primary habitat for native species. Wasteland represents abandoned deforested scrubland areas that are structurally distinct from oak forest and have been subjected to different disturbance regimes. Edge consists of communities on the narrow borders of the crops, which are composed of relatively permanent plant assemblages where perennial species persist for years. Finally, Crop consists of annual seasonal monocultures interspersed by weed species, which are rotated or left fallow between seasons.

In the two-year sampling period, a total of 78 collections were made at 23 different sites, 4 of Oakwood, 4 of Wasteland, 4 of Edge, and 11 of Crop **(see section 3.1.1.)**. A total of 6709 individual plant samples were collected, which comprised 272 different plant species **(Table 4.1)**. Crop had the lowest observed plant species richness, followed by Edge, Oakwood, and Wasteland.

For evaluation of sample completeness and sampling effort, rarefaction analyses were conducted**.** Most collections showed a near asymptotic relationship between the sample size and the expected plant species richness **(Figure 4.1)**, which indicated that our sampling effort was near to that necessary to produce the estimated asymptotic species richness. Exceptions were one collection from Crop and three from Edge (CrZ2V, Ed2V, Ed2P, Ed3P), which indicated that the sampling effort was insufficient for each of these collections. Additionally, the coverage-based asymptotic rarefaction and extrapolation analyses of plant species richness **(Table 4.1)** used to standardise the comparison of samples of uneven size, indicated Wasteland had the highest extrapolated number of species, followed by Oakwood, Edge, and Crop.

Table 4.1. Number of individual plant samples (*n*), collected during years 2015-2017, observed plant species richness (Observed) and plant species richness estimated by Chao1 estimator (Estimator) with their standard error (S.E.).

Habitats	*n*	Observed	Estimator	S.E.
Crop (Cr)	1191	63	116.67	29.61
Edge (Ed)	1080	104	184.21	30.09
Wasteland (Wl)	2249	124	212.82	32.09
Oakwood (Oa)	2189	117	187.90	26.23
Total	6709	272	402.21	34.58

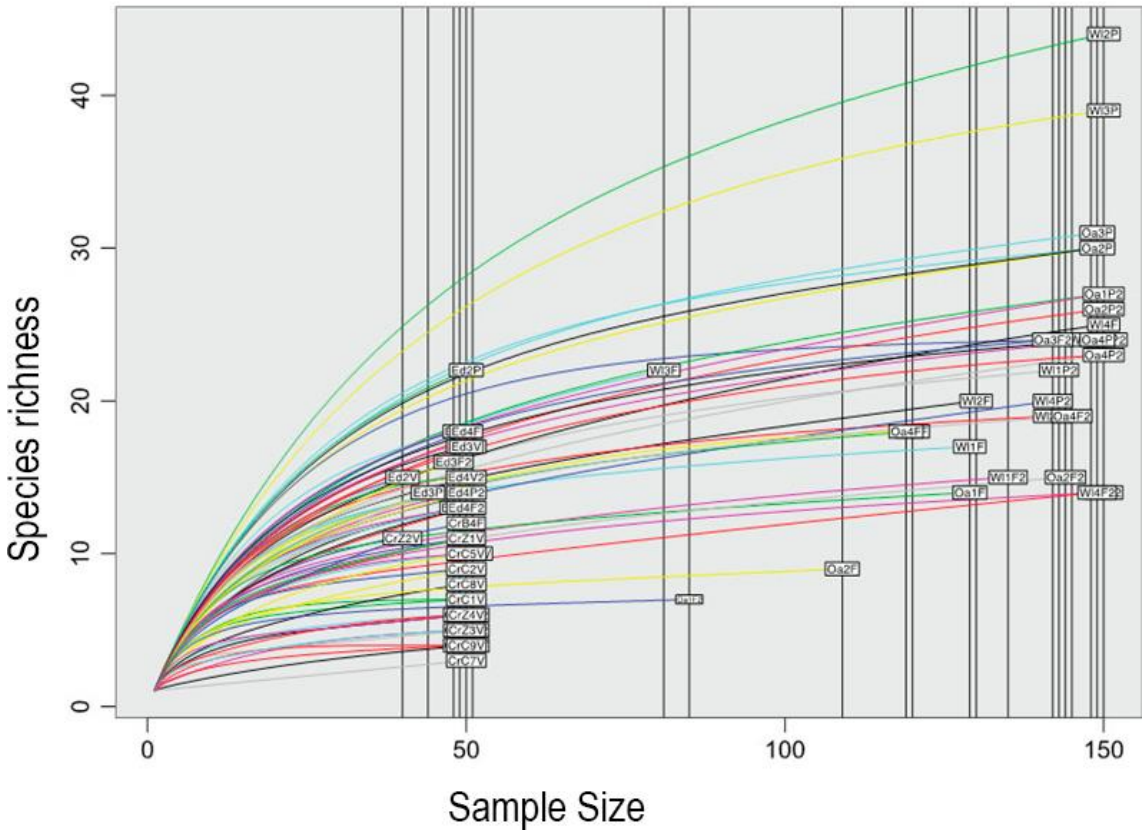

Figure 4.1. Plant species richness rarefaction curves for each collection (*n* = 78). The vertical lines indicate the respective sample sizes. The first two characters of the site codes indicate the habitat: (Cr) Crop of (B) cauliflower, (Z) maize, (C) melon, (H) barley; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.

4.2. Characterisation of plant communities

4.2.1. Differentiation of plant communities

To evaluate the relative homogeneity of the collections among each of the habitats and justify our *a priori* categorisation of each habitat, a detrended correspondence analysis (DCA) **(see section 3.2.1.)** of a collection-by-plant species abundance matrix (78 collections x 272 species) was performed. Plant communities of each collection clustered in discrete groups according to the each *a priori* habitat category **(Figure 4.2)**.

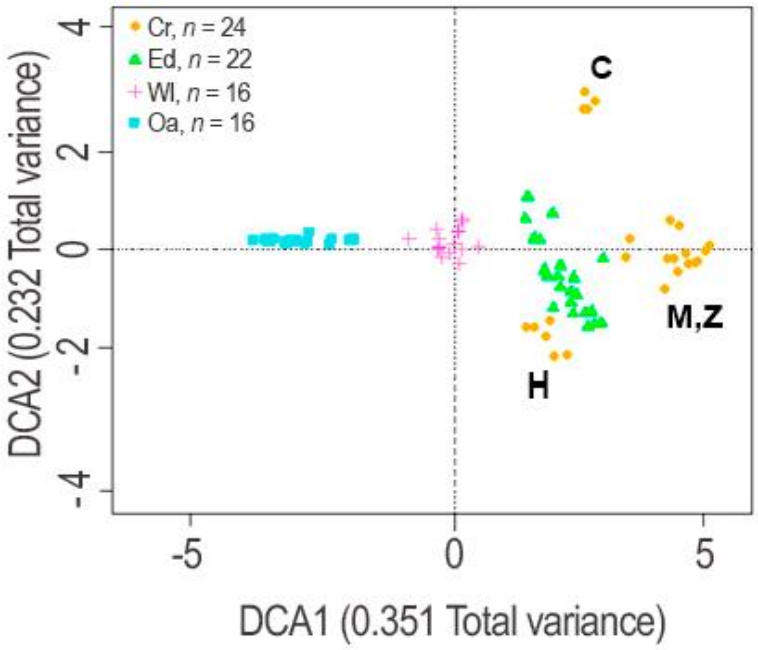

Figure 4.2. Detrended correspondence analysis (DCA) of a collection-by-plant species abundance matrix of the four habitats. Abbreviations, (C) Cauliflower, (Z) Maize, (M) Melon, (H) Barley; (Cr) Crop; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.

However, collections from Crop clustered according to the crop species that was grown at a given site and time. The DCA distinguished plant communities of Crop from each other and from plant communities of the other habitats. The influence of the Wasteland and Oakwood collections, and the high relative abundance of the crop species on the clustering of plant communities from Crop, were assessed in subsequent analyses. When the collections of Wasteland and Oakwood were omitted from the DCA, distinctions between the

plant communities associated with each crop species (i.e., C, M, H, & Z) became better resolved in that collections from M and Z formed two different clusters and not just one **(Figure 4.3.)**. When the dominant crop species were omitted from the DCA, clusters of the collections from Crop converged with those of Edge **(Figure 4.4.)**. The convergence of these clusters indicated the relative similarity in the plant assemblages represented in the collections from Crop and Edge, and the dissimilarity with the plant assemblages of both Wasteland and Oakwood. Therefore, Oakwood had the most differentiated plant community, with reduced dissimilarities between the communities of Crop and Edge when the abundant crop species were removed. The communities of Wasteland were also differentiated but were nearer to Edge and Crop than to Oakwood.

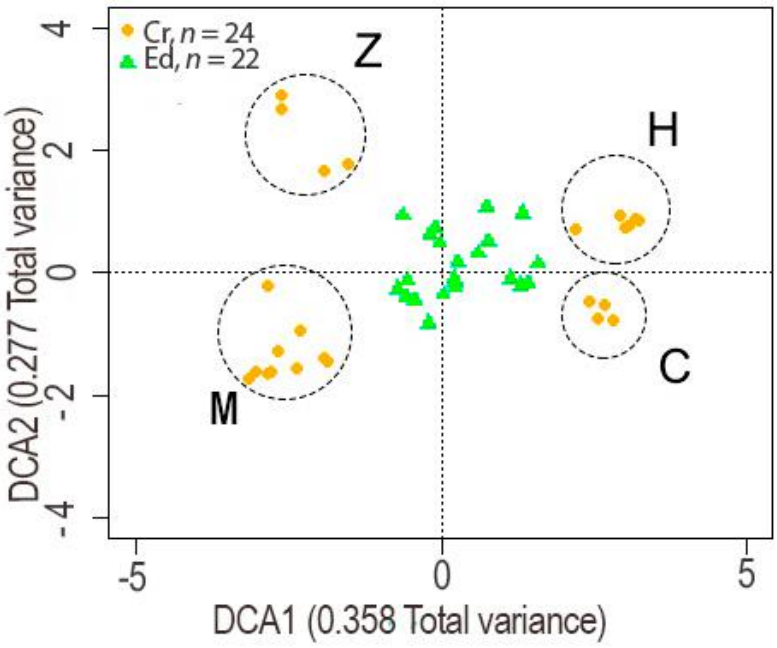

Figure 4.3. Detrended correspondence analysis (DCA) of a collection-by-plant species abundance matrix of the Edge and Crop habitats. Abbreviations, (C) Cauliflower, (Z) Maize, (M) Melon, (H) Barley; (Cr) Crop and (Ed) Edge.

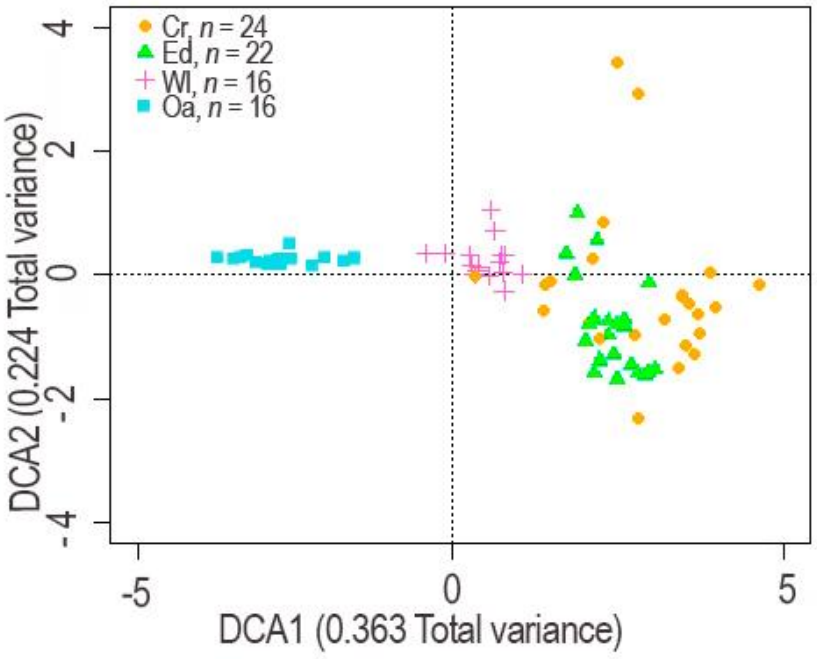

Figure 4.4. Detrended correspondence analysis (DCA) of a collection-by-plant species abundance matrix of the four habitats, with the abundant crop species omitted from the analysis. Abbreviations, (Cr) Crop; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.

4.2.2. Comparison of plant species diversity among habitats

The comparison of the habitats in terms of plant species diversity was conducted by using two different diversity estimators: the Tsallis entropy estimate of diversity (S_q), and the asymptotic estimator of Shannon diversity (D_{AE}) **(see section 3.2.2.)**. We used two measures because the choice of diversity estimator may affect the interpretation of the results. (McLeish *et al.,* 2017).

As multiple collections were conducted at each site, and multiple sites of each habitat were sampled, species relative abundance estimates could be aggregated according to site or habitat. However, data aggregation may influence estimates of diversity and any differences between the habitats. Therefore, to explore the effect of data aggregation on diversity estimates, diversity was estimated without data aggregation at the level of collection (n = 78), and with aggregation at the level of site (n = 23).

Differences in the mean values of S_q and D_{AE} among the habitats were apparent. The distinctions between Edge and Oakwood were most pronounced when the D_{AE} estimator was used and when the abundance data were aggregated **(Table 4.2** and **Figure 4.5)**. However, when the S_q diversity estimator

was calculated without data aggregation, the relationship between Edge and Oakwood was reversed. The distinctions among the habitats were sensitive to the choice of diversity estimator, which did not distinguish similarities and differences in species assemblages among the collections and habitats as were evident in the DCA.

There were significant differences in the diversity estimates of the habitats, both without data aggregation (Kruskal-Wallis rank sum tests: S_q $\chi^2_{(3)}$ = 54.832, P < 0.001; D_{AE} $\chi^2_{(3)}$ = 43.558, P < 0.001) **(see section 3.5.4.)** and when the data were aggregated by site (one-way ANOVA: $F_{(3,19)}$ = 102.00, P < 0.001). Comparisons of means by Tukey Honest Significant Difference (HSD) tests, adjusted for multiple comparisons, indicated all pair-wise habitat contrasts were significantly different (P-adjusted < 0.005) except that between Edge and Oakwood (P-adjusted = 0.999) when the S_q estimator was used. The distinction between these two habitats was greater (D_{AE}: P-adjusted = 0.052) when the D_{AE} estimator was given. Therefore, although data aggregation resulted in a relative increase in the plant species diversity of Edge, in all analyses, Crop was the habitat with the lowest diversity and Wasteland that with highest diversity.

Table 4.2. Plant species diversity estimated as Tsallis entropy (S_q) or the Shannon diversity asymptotic (D_{AE}) estimator, without data aggregation (n = 78) or with data aggregated by site (n = 23).

Aggregation	Habitat	n	S_q μ (S.E.)[1]	D_{AE} μ (S.E.)
None	Crop (Cr)	24	1.481 (0.155)	5.578 (0.598)
	Edge (Ed)	22	3.085 (0.114)	13.648 (0.849)
	Wasteland (Wl)	16	4.545 (0.297)	18.601 (2.292)
	Oakwood (Oa)	16	4.236 (0.306)	16.502 (1.735)
Site	Crop (Cr)	11	1.946 (0.268)	7.891 (1.401)
	Edge (Ed)	4	6.551 (0.406)	33.531 (2.384)
	Wasteland (Wl)	4	8.653 (0.158)	40.404 (2.188)
	Oakwood (Oa)	4	6.470 (0.206)	24.861 (1.502)

[1]Mean (μ) ± standard error (S.E.) of diversity estimators of each habitat.

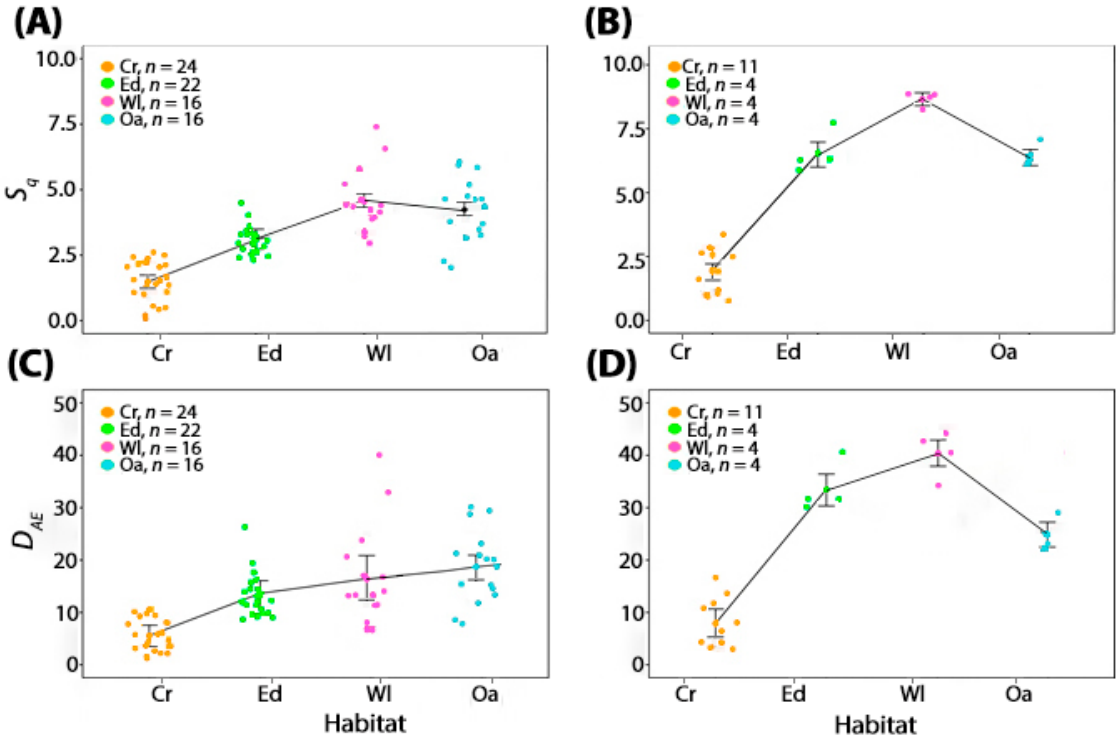

Figure 4.5. Values of Tsallis diversity (S_q) and the Shannon diversity asymptotic estimator (D_{AE}) of each habitat. Estimates calculated from samples of each collection (n = 78; A, C) and then aggregated by site (n = 23; B, D). Means and standard errors are indicated for each habitat. Abbreviations; Crop (Cr), Edge (Ed), Wasteland (Wl), and Oakwood (Oa).

Analyses were also conducted to detect the presence of temporal signals in the diversity of plant communities **(Figure 4.6)**. Kruskal-Wallis rank sum tests indicated significant differences in the diversity of plant communities between seasons (S_q $\chi^2_{(2)}$ = 18.61, P < 0.001; D_{AE} $\chi^2_{(2)}$ = 17.23, P < 0.002). Communities of Edge, Oakwood, and Wasteland experienced a decrease in diversity from Spring through the other seasons, compared to Crop where diversity increased marginally between spring and autumn.

Therefore, land-use practices in Crop produced a trend in seasonal diversity variation that was opposite to the other habitats.

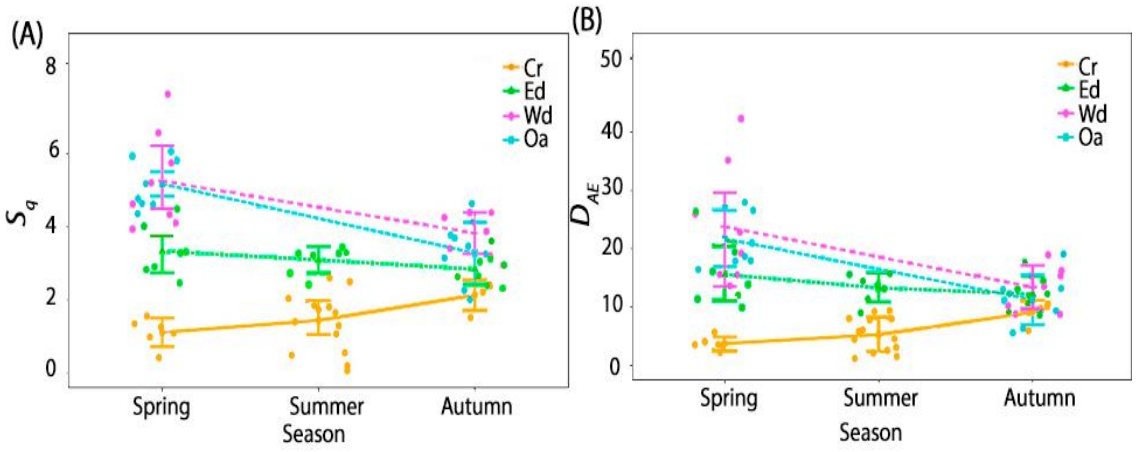

Figure 4.6. Seasonal variation of Tsallis diversity (S_q) (A) and the Shannon diversity asymptotic estimator (D_{AE}) (B) among collections (n = 78) from each habitat. Means and standard errors are indicated for each habitat. The line style and the dots and lines are coloured according to each habitat.

4.2.3. Rationalisation of plant species selection for HTS

The inclusion of rare plant species in the analysis of plant-virus infection networks at the landscape level can bias results **(see section 3.3.2.)**. However, when rare plant species are omitted, the categorisation of habitats may also be affected. The effect of omitting rare plant species on the differentiation of the habitats was evaluated by comparing DCAs that included all plant species and that which omitted rare species **(Figure 4.7)**. The DCAs indicated that when rare plant species were omitted, the clusters of collections were largely preserved with the exception of collections of Maize. Collections of Maize became differentiated from those of Melon, while the cluster of collections from Barley became merged with those of Edge.

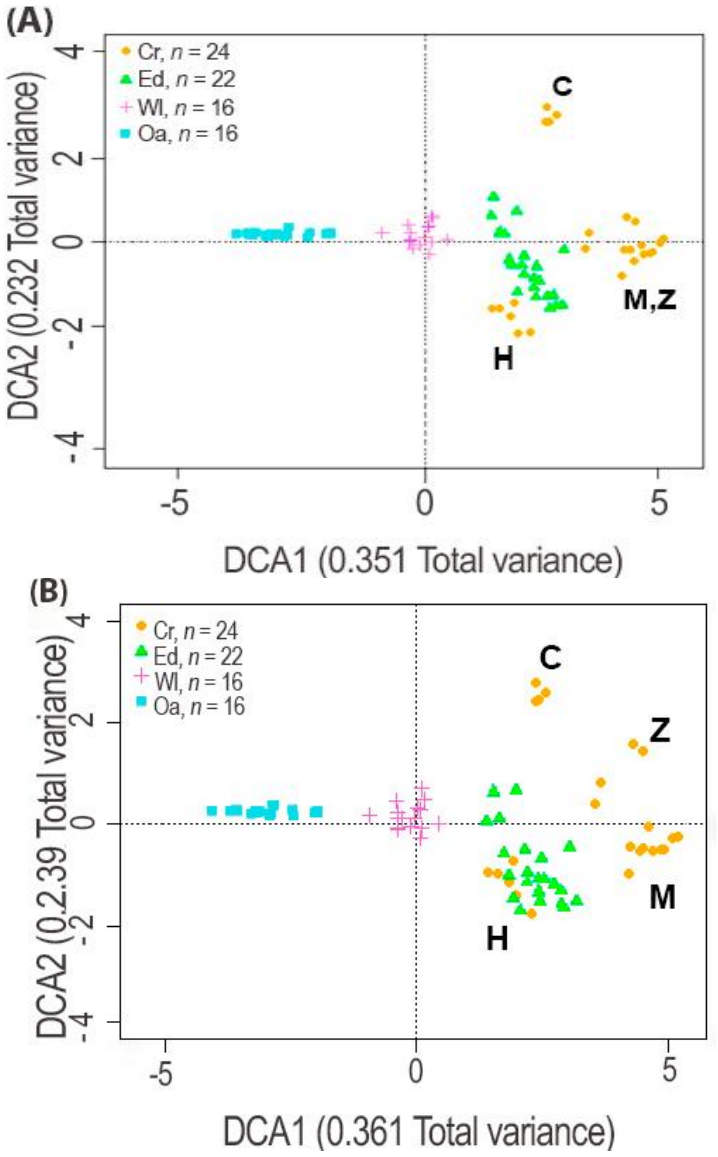

Figure 4.7. Detrended correspondence analysis (DCA) of collection-by-plant species abundance matrix of the four habitats considering all plant species (A) or only frequent plant species (B). Abbreviations, (C) Cauliflower, (Z) Maize, (M) Melon, (H) Barley; (Cr) Crop; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.

Host species observed in more than one habitat are "key" in plant virus infection networks, as they may be central to among-habitat transmission dynamics and be potential reservoirs for emerging viruses. To analyse if the omission of rare species preserved species shared among habitats, Kendall's tau rank correlation coefficient (τ) was used to test the correlation between the set memberships of species shared among habitats and unique before and after the rare plant species were omitted. Species richness of each habitat before and after rare species omission was positively correlated (Kendall rank correlation, z =

3.91, $P < 0.001$, $\tau = 0.776$). These results showed that the distribution of plant species richness among the habitats was similar before and after rare plant species were omitted **(Table 4.3)**. Therefore, there was a low contribution of rare plant species to the categorisation of the habitats and plant species richness overall, the relative proportions of which remained well preserved after rare species had been omitted.

Wasteland and Oakwood had the largest reduction in plant species richness, followed by Edge and Crop. However, the frequencies of the rare plant species omitted in each habitat were not related to the frequencies of individual plant samples of rare species omitted in each habitat ($\chi^2_{(3, 4)} = 0.003$, $P = 0.999$). Additionally, there was no relationship between the frequencies of individual plant samples omitted and habitats ($\chi^2_{(3, 4)} = 0.006$, $P = 0.999$). Oakwood and Edge had the largest reduction in the number of individual plant samples, followed by Crop and Wasteland.

Table 4.3. Comparison of the plant species richness and number of individual plant samples in each habitat before and after rare plant species were omitted.

Habitat	All species		Frequent species	
	S	***n***	***S***	***n***
Crop (Cr)	63	1191	48 (23.8%)[1]	1157 (2.9%)[2]
Edge (Ed)	104	1080	77 (26.0%)	1036 (4.1%)
Wasteland (Wl)	124	2249	89 (28.2%)	2187(2.8%)
Oakwood (Oa)	117	2189	78 (33.3%)	2091(4.5%)
Total	272	6709	172 (36.8%)	6471(3.6%)

[1] Percentage of the plant species richness after omitting rare plant species.
[2] Percentage of the individual plant samples after omitting rare plant species.

The removal of rare plant species resulted in a decrease of 36.8%, in the number of plant species, and a decrease of 3.6% in the number of plant samples (238 samples removed). This data underscores that although 100 species were removed, the reduction in the number of individual plant samples was low. Therefore, the exclusion of these rare plant species in analyses of host-parasite interactions implies a reduction in the possible bias in the detection of infection and estimation of prevalence.

To further rationalise the protocol, we limited the HTS analyses to species that were frequent between the summer of 2015 and spring of 2016, which occurred in more than one habitat, or were highly characteristic of each plant community. A DCA was conducted to assess whether the distinctions among the plant communities were preserved in the subset of plant species selected for HTS. The distinctions among plant communities of the habitats were preserved in the subset of species selected for HTS and showed negligible differences with the DCA that comprised all species observed **(Figure 4.8)**. Although not substantial, the subset of collections from Edge and Crop converged, as the increase in shared species between them would be expected to affect the ecological distance estimates between collections. By omitting from the ordination information provided by particular plant species, there was less information to distinguish the habitats. In addition, the subset of species selected from Oakwood for HTS, which had a large proportion of rare species omitted, also remained distinct, but with some convergence with Wasteland that indicated similarities between their plant communities. In addition, the analysis corresponded to the first year of sampling, and this reduction in species occurrence and abundance from those sampled between 2015 and 2017, was expected to reduce the discriminatory power of the ordination. This was a desirable result because species selected for HTS were shared predominantly between habitats, thus decreasing the number of species specific to each habitat, and those which contributed to their differentiation.

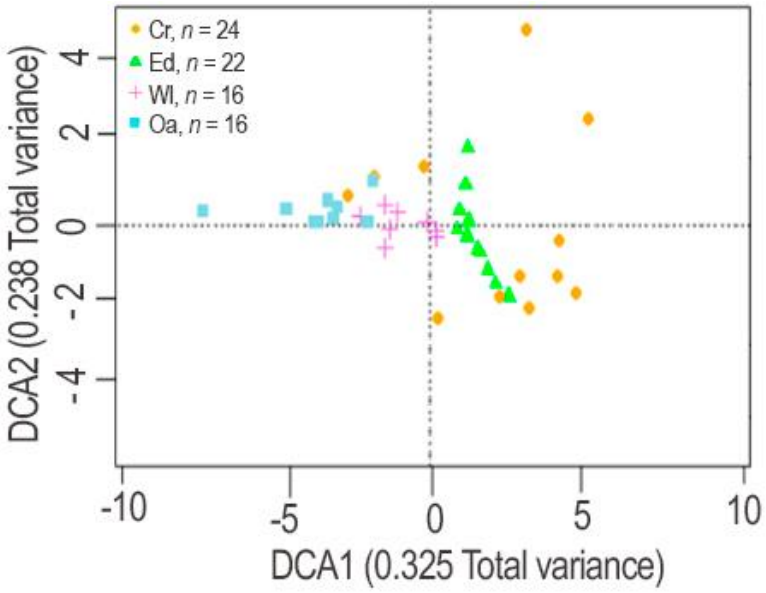

Figure 4.8. Detrended correspondence analysis (DCA) of a collection-by-plant species abundance matrix of the four habitats considering the plant species selected for HTS. Abbreviations, (Cr) Crop; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.

The distribution of plant species selected for HTS across habitats was: 25 of Crop, 52 of Edge, 51 of Oakwood, and 52 of Wasteland **(Table 4.4)**. The percentages of plant species selected for HTS to the species sampled during the first year were: 64.10% in Crop, 91.07% in Edge, 78.46% in Oakwood and 73.24% in Wasteland **(Table 4.4)**.

Table 4.4. Plant species richness in each habitat after different sets of species were omitted.

Habitats	S^1	S_1^2	S_r^3	S_{r1}^4	S_{HTS}^5	S_{HTS}/S_{r1}^6
Crop (Cr)	63	51	48	39	25	64.10 %
Edge (Ed)	104	73	77	56	52	91.07 %
Wasteland (Wl)	124	96	89	71	52	73.24 %
Oakwood (Oa)	117	86	78	65	51	78.46 %
Total	272	211	172	144	118	81.94 %

[1]Total plant species richness.
[2]Species richness of the first year of collection (2015-2016).
[3]Frequent plant species (with more than five individuals in at least one habitat).
[4]Frequent plant species for the first year of collection.
[5]Plant species selected for HTS.
[6]Percentage of plant species selected for HTS to frequent plant species of the first year of collection.

4.3. Sample preparation and obtention of nucleotide sequences

4.3.1. Optimisation of RNA preparations

Due to the large variety of plant species in our collections and the different phenological states of specimens across seasons, it was necessary to evaluate which factors in the RNA extraction procedure (plant species, tissue homogenisation, extraction) had the greatest effect on the quality of the RNA extract. For this purpose, an experiment was conducted in which we assessed three methods of tissue homogenisation, and three RNA extraction procedures implemented in commercial kits **(see section 3.3.1.)**. The resulting nine procedures were assayed on five randomly selected samples from each of six species (*Bromus rubens* L., *Carduus bourgeanus* L., *Cucumis melo* L., *Papaver rhoeas* L., *Quercus ilex* L., *Stipa parviflora* L.). Final extracts were re-suspended in 50 µl of nuclease-free water. These six species were also randomly selected from those that yielded low RNA concentration (ng/µl), poor absorbance ratios at 260/230 and 260/280 nm, and had low integrity determined using agarose gel electrophoresis analyses **(see section 3.3.1.)**. All tissues were homogenised with a mortar and pestle using glass powder and extracted with the Direct-zol™ RNA MiniPrep Plus kit.

Two hundred and seventy RNA extracts were thus obtained (5 individuals x 6 species x 3 tissue homogenisation methods x 3 RNA extraction procedures). A table with concentration and absorbance ratio values obtained for each extract is shown in **Annex, Table A1.** The mean values of the concentration and absorbance ratios obtained with each of the 9 procedures are shown in **Table 4.5**.

Table 4.5. Mean and standard error of RNA concentration (ng/µl) and absorbance ratios for different plant species, purified by different tissue homogenisation methods and RNA extraction procedures as implemented in different kits.

Species	Species	Homogenisation I[1]			Homogenisation II[1]			Homogenisation III[1]		
		Kit A[2]	Kit B[2]	Kit C[2]	Kit A[2]	Kit B[2]	Kit C[2]	Kit A[2]	Kit B[2]	Kit C[2]
B. rubens	Concentration	52.1 ± 8.2	55.1 ± 8.2	143.9 ± 28.5	37.3 ± 10.4	15.8 ± 5.1	14.5 ± 2.3	22.5 ± 11.5	8.7 ± 1.8	49.8 ± 10.1
	260/230 Absorbance	1.6 ± 0.0	2.0 ± 0.1	2.1 ± 0.0	1.5 ± 0.1	1.8 ± 0.1	2.1 ± 0.2	1.3 ± 0.1	1.5 ± 0.0	2.0 ± 0.1
	260/280 Absorbance	1.1 ± 0.1	1.4 ± 0.3	1.9 ± 0.2	1.2 ± 0.2	1.0 ± 0.2	1.5 ± 0.4	0.9 ± 0.1	0.9 ± 0.3	1.8 ± 0.3
C. bourgeanus	Concentration	203.2 ± 30.6	254.3 ± 91.7	318.3 ± 95.6	241.1 ± 119.3	157.4 ± 76.8	103 ± 36.9	115.3 ± 28.1	119.8 ± 44.9	129.0 ± 33.0
	260/230 Absorbance	1.8 ± 0.1	2.0 ± 0.0	2.1 ± 0.0	1.8 ± 0.1	2.0 ± 0.1	2.2 ± 0.0	1.8 ± 0.1	2.0 ± 0.1	2.1 ± 0.0
	260/280 Absorbance	1.5 ± 0.2	1.2 ± 0.3	1.9 ± 0.2	1.5 ± 0.2	1.0 ± 0.4	1.6 ± 0.2	1.6 ± 0.3	0.9 ± 0.3	1.8 ± 0.2
C. melo	Concentration	519.6 ± 112.4	266.6 ± 37.8	490.3 ± 58.0	776.8 ± 102.5	194.7 ± 31.1	216.4 ± 21.8	676.9 ± 152.4	199.5 ± 71.3	345.8 ± 47.3
	260/230 Absorbance	2.0 ± 0.0	2.1 ± 0.0	2.1 ± 0.0	2.1 ± 0.0	2.1 ± 0.0	2.1 ± 0.0	2.0 ± 0.0	2.1 ± 0.0	2.1 ± 0.0
	260/280 Absorbance	2.2 ± 0.1	1.8 ± 0.1	2.2 ± 0.0	2.2 ± 0.0	1.7 ± 0.2	1.9 ± 0.1	2.2 ± 0.0	1.4 ± 0.4	2.0 ± 0.1
P. rhoeas	Concentration	224.1 ± 39.2	229.4 ± 97.3	234.9 ± 106.0	115.0 ± 65.7	67.4 ± 35.6	65.9 ± 26.9	97.1 ± 30.4	27.3 ± 12.3	94.1 ± 72.4
	260/230 Absorbance	2.0 ± 0.0	2.1 ± 0.0	2.1 ± 0.0	1.9 ± 0.0	2.1 ± 0.1	2.0 ± 0.1	1.7 ± 0.1	2.0 ± 0.0	2.2 ± 0.1
	260/280 Absorbance	1.7 ± 0.1	1.6 ± 0.3	1.7 ± 0.2	1.5 ± 0.2	0.9 ± 0.4	1.2 ± 0.3	1.2 ± 0.2	0.7 ± 0.2	1.1 ± 0.3
Q. ilex	Concentration	112.3 ± 57.6	20.5 ± 4.2	271.5 ± 89.5	45.4 ± 17.9	19.1 ± 7.8	534.9 ± 383.4	38.9 ± 11.4	16.2 ± 6.5	158.9 ± 30.7
	260/230 Absorbance	1.4 ± 0.0	1.7 ± 0.1	1.6 ± 0.0	1.3 ± 0.1	1.6 ± 0.1	1.6 ± 0.0	1.4 ± 0.0	1.4 ± 0.1	1.6 ± 0.1
	260/280 Absorbance	0.6 ± 0.1	0.9 ± 0.1	1.1 ± 0.4	0.6 ± 0.0	0.8 ± 0.1	0.8 ± 0.1	0.7 ± 0.1	0.9 ± 0.2	0.9 ± 0.1
S. parviflora	Concentration	161.1 ± 37.1	189.9 ± 69.5	120.9 ± 61.1	88.1 ± 27.6	14.7 ± 3.0	48.9 ± 24.3	93.9 ± 23.2	19.3 ± 11.8	32.9 ± 11.7
	260/230 Absorbance	2.0 ± 0.1	2.1 ± 0.0	2.2 ± 0.1	2.0 ± 0.1	2.1 ± 0.1	3.3 ± 1.3	1.8 ± 0.1	1.9 ± 0.2	2.1 ± 0.1
	260/280 Absorbance	1.8 ± 0.2	1.8 ± 0.2	1.7 ± 0.4	1.8 ± 0.2	0.7 ± 0.2	1.6 ± 0.4	1.4 ± 0.2	1.1 ± 0.4	1.2 ± 0.4

[1]Tissue homogenisation methods: I; mortar with pestle and glass powder, II; Eppendorf tube without glass powder, III: Eppendorf tube with glass powder.
[2]Kit: A; Direct-zol™ RNA MiniPrep Plus, B; Rneasy Mini Kit. Kit, C; Agilent Plant RNA Isolation Mini Kit.

Generalised linear mixed models (GLMM) were conducted in which the tissue homogenisation and RNA extraction procedures were considered as fixed factors, and the plant species as a random factor, in a full model considering all interactions between factors. Either RNA concentration or absorbance ratio was the response variable **(Table 4.6)**. For each response variable a separate model was implemented **(see section 3.5.4.)**. The concentration response variable was modelled using a gamma distribution and log-link function. The absorbance ratio response variables were modelled using a normal distribution and identity-link function.

GLMM analyses indicated that the RNA extraction procedures had an effect on each of the response variables (concentration: Wald $\chi^2_{(2,3)} = 40.025$, $P < 0.001$; 260/230 absorbance ratio: Wald $\chi^2_{(2,3)} = 32.79$, $P < 0.001$; 260/280 absorbance ratio: Wald $\chi^2_{(2,3)} = 26.095$, $P < 0.001$). The tissue homogenisation method had an effect on concentration and 260/230 absorbance ratio ($\chi^2_{(2,3)} = 16.762$, $P < 0.001$). The interaction of both factors only had a significant effect on the RNA concentration (Wald $\chi^2_{(2,3)} = 9.991$, $P = 0.041$). The variance explained was assessed by R^2 values for each model implemented using the methods of Nakagawa & Schielzeth (2013). The marginal R^2 value, that considered both fixed factors, showed a higher value for concentration response variable (R^2: 11.9%) than for absorbance ratios at 260/280 nm and 260/230 nm (R^2: 10.9%, R^2: 10.6%, respectively). The conditional R^2 value, that considered both fixed factors and random factor, showed a higher value for concentration response variable (R^2: 40%) than for absorbance ratios at 260/280 nm and 260/230 nm (R^2: 29.6, R^2: 38.7%, respectively).

To test the significant difference between adjacent mean values within each fixed factor, LSD analyses were employed **(Annex, Table A2)**. LSD test showed that for the RNA concentration mean values obtained with the tissue homogenisation method mortar and pestle with glass powder, and Direct-zol™ RNA MiniPrep Plus Kit and Agilent Plant RNA Isolation Mini Kit were significantly higher ($F_{(2,3)} \geq 0.641$, $P < 0.001$ and $F_{(2,3)} \geq 0.664$, $P < 0.001$, respectively) than with the other tissue homogenisation methods and RNA extaction procedures. The mean values of the 260/230 absorbance ratio obtained with the tissue homogenisation methods did not differ significantly ($F_{(2,3)} \geq 0.031$, $P \leq 0.875$).

While the mean values obtained with Agilent Plant RNA Isolation Mini Kit were closer to the optimum value 2.0-2.2, differing significantly to the other kits ($F_{(2,3)} \geq 0.173$, $P \leq 0.018$). Finally, the 260/280 absorbance ratio mean values obtained with the tissue homogenisation mortar with pestle and glass powder and the kits Direct-zol™ RNA MiniPrep Plus Kit and Agilent Plant RNA Isolation Mini Kit were lower than the optimum value 2-2.2, differing significantly to the tissue homogenisation methods ($F_{(2,3)} \geq 0.239$, $P \leq 0.008$) and RNA extaction procedures ($F_{(2,3)} \geq 0.270$, $P < 0.002$).

Since absorbance is an ordinal scale, but one with a specific value of reference that is not determined by magnitude, one value may be significantly higher but with a higher standard error value than the rest. For the set of values obtained with the different tissue homogenisation methods and RNA extraction procedures, the range of variation of the variances for 260/230 and 260/280 absorbance ratios was equal ($F_{(8,261)} = 1.304$, $P = 0.242$; $F_{(8,261)} = 0.463$, $P = 0.881$, respectively).

Table 4.6. Generalised linear mixed model of concentration and absorbance ratio responses with fixed factors for tissue homogenisation method and RNA extraction procedures, and a random factor for plant species. The significant (Type II Wald χ^2) relationships are indicated in bold text.

Response variable	Fixed effects	Estimate	χ^2	Pr (>χ^2)	AIC
Concentration[1]	Homogenisation	-0.78	47.97	**<0.001**	3,144.2
	Kit	0.29	40.03	**<0.001**	
	Homogenisation * Kit	2.34	9.99	**0.041**	
260/230 absorbance[1]	Homogenisation	-0.17	5.20	0.074	332.4
	Kit	0.45	32.79	**<0.001**	
	Homogenisation * Kit	0.28	3.49	0.479	
260/280 absorbance[1]	Homogenisation	-0.16	16.76	**<0.001**	453.6
	Kit	0.22	26.10	**<0.001**	
	Homogenisation * Kit	-1.13	5.10	0.277	

[1]Model equations used: Concentation ~ F1*-0.78 + F2*0.29 + F1*F2*2.34+ 5.17; 260/230 absorbance ~ F1* -0.17 + F2*0.45 + F1*F2*0.28 + 1.79; 260/280 absorbance ~ F1* -0.16 + F2*0.22 + F1*F2*-1.13 + 1.48; where F1 is the first fixed factor (Tissue homogenisation method) and F2 is the second fixed factor (Kit).

The integrity of the RNA extracts of two species randomly selected from the six used in the optimisation assay was assessed with agarose gel electrophoresis **(Figure 4.9)**. Well-defined bands that corresponded to 28S and 18S ribosomal RNA (rRNA) were present after using the different tissue homogenisation methods and RNA extraction procedures and indicated that all of the yielded RNA had good integrity, except for that from the homogenisation of *C. borgeanus* in Eppendorf tubes without glass powder and the Agilent Plant RNA Isolation Mini Kit.

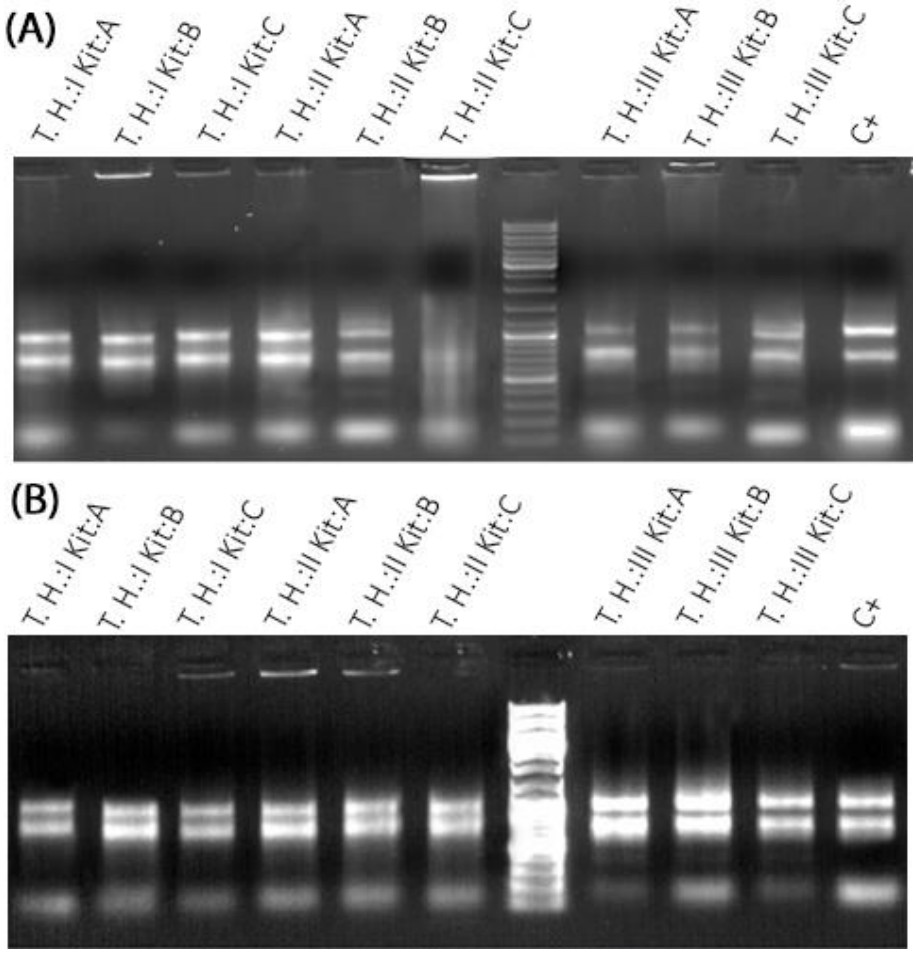

Figure 4.9. Agarose gel electrophoresis or RNA extracted by different procedures from different plant species. Tissue was homogenised (T.H.) by: ((I) mortar with glass powder, (II) Eppendorf without glass powder, (III) Eppendorf with glass powder). RNA was extracted (Kit) by ((A) Direct-zol™ RNA MiniPrep Plus, (B) Rneasy Mini Kit, (C) Agilent Plant RNA Isolation Mini Kit). Plant species were *C. bourgeanus* (A) and *C. melo* (B) shown as examples. The results are supported with a positive control of *Nicotiana benthamina* L. RNA extract extracted by Direct-zol™ RNA MiniPrep Plus and mortar with glass powder.

4.4. Detection and validation of viral OTUs

4.4.1. Quality control of plant RNA extracts for detection of viral sequences

For HTS, RNA extracts of individual plant samples from the same plant species and collection (i.e., same time, same site) were pooled to obtain a single HTS library **(see section 3.3.3.)**. However, in some cases, samples from the same plant species for a given collection were sparsely present. To avoid pooling a small number of individual RNA extracts, in some cases individual RNA extracts were pooled by site. After sequencing, adapter sequence contamination was trimmed from the HTS libraries **(see section 3.4.1.)**. Reference codes of the 306 HTS libraries obtained, plant species and family, the habitat and site and collection from which the plant samples came, and the number of RNA extracts that had been pooled are shown in **Annex, Table A3.**

The most efficient protocol selected for RNA extraction was not always sufficient to produce the desired RNA quality. This can negatively affect library preparation, HTS performance, and the sensitivity/accuracy of virus sequence detection and potentially result in a biased estimation of plant-virus interactions. To test whether the quality of the RNA extracts might bias the performance of the HTS run, endogenous messenger RNA (mRNA) expected to be subject to the same sources of variability as viral RNAs, were used as quality controls. Four mRNAs that are known to have stable levels of expression were chosen as quality controls (Dean *et al*., 2002; Expósito-Rodríguez *et al.,* 2008; Løvdal *et al.,* 2009) **(Table 4.7.) (see section 3.4.4.)**. BLAST query analyses showed that these mRNAs (*DnaJ-like protein, Elongation factor 1-α, Glyceraldehyde-3-phosphate dehydrogenase,* and *Ribosomal protein L2)* were detected in a large fraction of libraries **(Table 4.7.)**. In addition, the *Topoisomerase II associated protein* gene, which is exclusive to *Arabidopsis thaliana* L. (Catalá *et al.,* 2019), was included as a negative control. Reads of the mRNA of *Topoisomerase II associated protein*, were detected in about 6% of the HTS libraries, and indicated that the presence of reads of the other mRNAs were not artefacts. The reads of the mRNA of *Topoisomerasae II associated protein* in the 19 HTS libraries corresponded to the genomes of plant families that comprised *Brassicaceae, Apiaceae,*

Asteraceae, Geraniaceae, Lamiaceae, Poaceae, Polygonaceae, Rhamnaceae, and Rubiaceae. As mentioned in **section 3.4.2.**, the BLAST query does not guarantee true positive detections, and may result in false positive detections of both the mRNA of *Topoisomerase II associated protein*, and that of the mRNA of other genes. However, this analysis showed that the detection of virus sequences should be sufficiently robust to variation in RNA extract quality as to allow comparative analyses among HTS libraries.

Table 4.7. Number and proportion of HTS libraries with reads derived from mRNAs from plant genes.

Plant gen	Libraries detected
DnaJ-like protein	249/306 (81.37%)
Elongation factor 1-α	306/306 (100.00%)
Glyceraldehyde-3-phosphate dehydrogenase	305/306 (99.67%)
Ribosomal protein L2	284/306 (92.81%)
Topoisomerase II associated protein	19/306 (6.21%)

4.4.2. Validation of viral OTUs

Virus detection was conducted using BLAST queries of the HTS libraries against a local database of single-stranded RNA (ssRNA) virus reference genomes **(see section 3.4.2.)**. We limited the database to ssRNA, because they are the major class of plant viruses (King *et al.,* 2011). Reads that matched with a reference were provisionally assigned to the virus species of the corresponding target genome in the local database and considered as operational taxonomic units (OTUs). A series of *in silico* validation steps were then conducted to reduce the frequency of false positive detections that potentially resulted from the BLAST queries **(see section 3.4.3.)**. These steps retained OTUs of a given library that: 1) matched both paired-end reads with the same reference, and that did not match the same reference more than once; 2) had a percentage identity with the reference equal to or greater than 90%; 3) had a query length greater or equal to 100 nucleotides; and 4) had a query coverage equal to 100%.

Kernel density estimations (KDE) of the BLAST query bit-scores (*S'*) were used to assess the change in sequence similarity between BLAST queries and

references before and after the application of the *in silico* validation criteria. The KDE bandwidth was adjusted to remove multimodal distributions. Before applying the validation criteria, the *S'* mean values were 185 and 187 for each read orientation (range, 45.4 to 279.0) **(Figure 4.10.A)**. Once the validation criteria were applied, the mean values of *S'* increased to 243 and 241 (range 124.0 – 279.0) **(Figure 4.10.B)**.

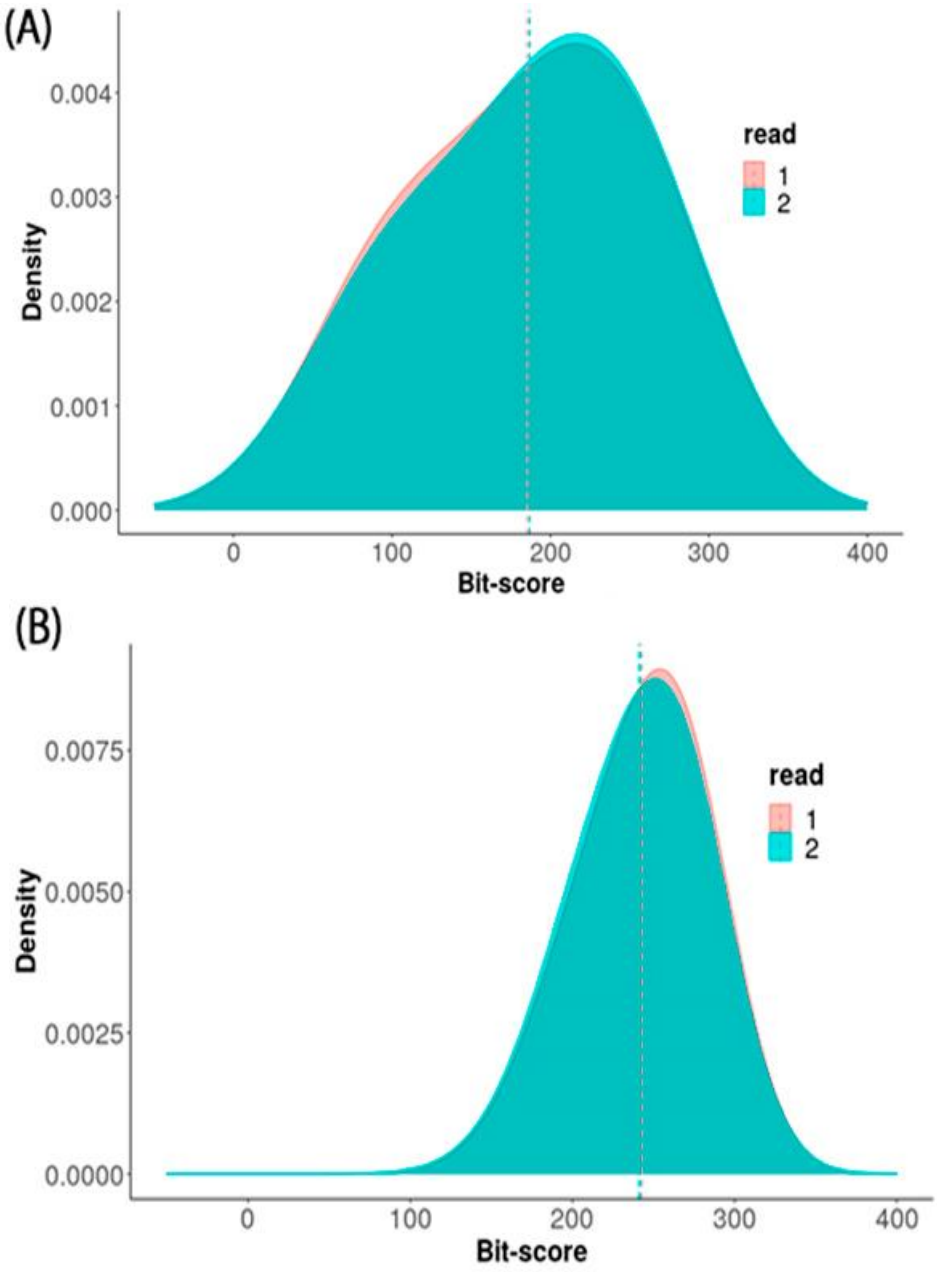

Figure 4.10. A kernel density estimation (KDE) of *the* BLAST bit-score parameter (*S'*) before (A) and after (B) applying validation criteria. Dashed lines show the mean bit-score value for both read orientations.

The improvement in the quality of each match is shown by the positive shift in *S'* distributions after applying the validation criteria and was partly an effect of removing reads that had sequence similarity with the reference lower than 90%.

4.4.3. Viral OTUs detected in HTS libraries

After applying the validation criteria to the BLAST query results performed against a local database of ssRNA virus reference genomes, virus OTUs were detected in 287 of 306 total HTS libraries. We detected 92 OTUs, defined by similarity to known virus species. From here on, OTUs will be identified according to the name of the corresponding virus species. Of the 92 OTUs, 90 corresponded to plant-infecting viruses and two to fungal viruses *(Botrytis ourmia-like virus, Botrytis virus F).* These fungal viruses were not included in the following analyses **(Table 4.8)**.

Table 4.8. Habitat occurrence of virus OTUs.

Species	Genus	Family	Detections[1]	Cr[2]	Ed[2]	Wl[2]	Oa[2]
Alfalfa mosaic virus	*Alfamovirus*	*Bromoviridae*	3	0	0	1	1
Pelargonium zonate spot virus	*Anulavirus*	*Bromoviridae*	143	1	1	1	1
Brome mosaic virus	*Bromovirus*	*Bromoviridae*	3	0	1	1	0
Cucumber mosaic virus	*Cucumovirus*	*Bromoviridae*	188	1	1	1	1
Gayfeather mild mottle virus	*Cucumovirus*	*Bromoviridae*	33	1	1	0	1
Tomato aspermy virus	*Cucumovirus*	*Bromoviridae*	59	1	1	1	1
Parietaria mottle virus	*Ilarvirus*	*Bromoviridae*	18	0	1	0	0
Prunus necrotic ringspot virus	*Ilarvirus*	*Bromoviridae*	3	0	0	0	1
Barley yellow dwarf virus	*Luteovirus*	*Luteoviridae*	10	1	1	1	0
Barley virus G	*Polerovirus*	*Luteoviridae*	1	1	0	0	0
Beet chlorosis virus	*Polerovirus*	*Luteoviridae*	35	1	1	0	1
Beet mild yellowing virus	*Polerovirus*	*Luteoviridae*	38	1	1	0	1
Beet western yellows virus	*Polerovirus*	*Luteoviridae*	45	1	1	0	1
Brassica yellows virus	*Polerovirus*	*Luteoviridae*	37	1	1	1	1
Carrot red leaf virus	*Polerovirus*	*Luteoviridae*	2	0	1	0	0
Cereal yellow dwarf virus	*Polerovirus*	*Luteoviridae*	3	1	1	0	0
Chickpea chlorotic stunt virus	*Polerovirus*	*Luteoviridae*	3	0	1	0	0
Cotton leafroll dwarf virus	*Polerovirus*	*Luteoviridae*	2	0	1	0	0
Cowpea polerovirus 1	*Polerovirus*	*Luteoviridae*	2	0	1	0	0
Cucurbit aphid-borne yellows virus	*Polerovirus*	*Luteoviridae*	8	0	1	1	0
Ixeridium yellow mottle virus 1	*Polerovirus*	*Luteoviridae*	3	0	1	1	0
Ixeridium yellow mottle virus 2	*Polerovirus*	*Luteoviridae*	2	0	1	1	0
Melon aphid-borne yellows virus	*Polerovirus*	*Luteoviridae*	9	0	1	0	0
Pepo aphid-borne yellows virus	*Polerovirus*	*Luteoviridae*	6	0	1	0	0
Potato leafroll virus	*Polerovirus*	*Luteoviridae*	21	0	1	0	0

[1]Detections represent the number of libraries, out of the total of 306 libraries, in which the OTUs were detected.
[2]Presence (1) or absence (0) of viral OTUs in each habitat. Abbreviations, (Cr) Crop; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.

Table 4.8. Continued.

Species	Genus	Family	Detections[1]	Cr[2]	Ed[2]	Wl[2]	Oa[2]
Suakwa aphid-borne yellows virus	*Polerovirus*	*Luteoviridae*	4	0	1	0	0
Sugarcane yellow leaf virus	*Polerovirus*	*Luteoviridae*	1	0	1	0	0
Tobacco virus 2	*Polerovirus*	*Luteoviridae*	21	0	1	0	0
Turnip yellows virus	*Polerovirus*	*Luteoviridae*	51	1	1	1	1
Wheat leaf yellowing-associated virus	*Polerovirus*	*Luteoviridae*	1	0	1	0	0
Wheat yellow dwarf virus	*Polerovirus*	*Luteoviridae*	3	1	1	0	0
Calla lily latent virus	*Potyvirus*	*Potyviridae*	6	1	0	0	0
Iranian johnsongrass mosaic virus	*Potyvirus*	*Potyviridae*	1	0	1	0	0
Maize dwarf mosaic virus	*Potyvirus*	*Potyviridae*	31	0	1	0	1
Plum pox virus	*Potyvirus*	*Potyviridae*	10	1	1	1	1
Potato virus Y	*Potyvirus*	*Potyviridae*	1	0	1	0	0
Sorghum mosaic virus	*Potyvirus*	*Potyviridae*	3	0	1	0	0
Soybean mosaic virus	*Potyvirus*	*Potyviridae*	9	1	1	0	0
Turnip mosaic virus	*Potyvirus*	*Potyviridae*	18	1	1	1	1
Watermelon mosaic virus	*Potyvirus*	*Potyviridae*	66	1	1	1	1
Zucchini yellow mosaic virus	*Potyvirus*	*Potyviridae*	2	1	1	0	0
Wheat streak mosaic virus	*Tritimovirus*	*Potyviridae*	1	1	0	0	0
Barley yellow striate mosaic virus	*Cytorhabdovirus*	*Rhabdoviridae*	1	0	1	0	0
Eggplant mottled dwarf virus	*Nucleorhabdovirus*	*Rhabdoviridae*	3	0	0	1	1
Lettuce big-vein associated virus	*Varicosavirus*	*Rhabdoviridae*	1	0	1	0	0
Lettuce yellow mottle virus	*Varicosavirus*	*Rhabdoviridae*	6	0	1	0	0
Rubus chlorotic mottle virus	*Sobemovirus*	*Solemoviridae*	103	1	1	1	1
Turnip rosette virus	*Sobemovirus*	*Solemoviridae*	1	0	0	1	0

[1]Detections represent the number of libraries, out of the total of 306 libraries, in which the OTUs were detected.
[2]Presence (1) or absence (0) of viral OTUs in each habitat. Abbreviations, (Cr) Crop; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.

Table 4.8. Continued.

Species	Genus	Family	Detections[1]	Cr[2]	Ed[2]	Wl[2]	Oa[2]
Olive latent virus 1	*Alphanecrovirus*	*Tombusviridae*	1	0	1	0	0
Olive mild mosaic virus	*Alphanecrovirus*	*Tombusviridae*	1	0	1	0	0
Potato necrosis virus	*Alphanecrovirus*	*Tombusviridae*	1	0	1	0	0
Cucumber leaf spot virus	*Aureusvirus*	*Tombusviridae*	3	1	1	0	0
Pothos latent virus	*Aureusvirus*	*Tombusviridae*	3	1	1	0	0
Yam spherical virus	*Aureusvirus*	*Tombusviridae*	2	0	1	0	0
Cardamine chlorotic fleck virus	*Betacarmovirus*	*Tombusviridae*	3	0	1	1	0
Turnip crinkle virus	*Carmovirus*	*Tombusviridae*	30	0	1	1	1
Beet black scorch virus	*Necrovirus*	*Tombusviridae*	1	0	1	0	0
Tobacco necrosis virus A	*Necrovirus*	*Tombusviridae*	2	0	1	0	0
Tobacco necrosis virus D	*Necrovirus*	*Tombusviridae*	1	0	1	0	0
Bermuda grass latent virus	*Panicovirus*	*Tombusviridae*	1	0	0	0	1
Artichoke mottled crinkle virus	*Tombusvirus*	*Tombusviridae*	2	0	1	1	0
Carnation Italian ringspot virus	*Tombusvirus*	*Tombusviridae*	2	0	1	1	0
Cucumber necrosis virus	*Tombusvirus*	*Tombusviridae*	2	0	1	1	0
Cymbidium ringspot virus	*Tombusvirus*	*Tombusviridae*	1	0	0	1	0
Eggplant mottled crinkle virus	*Tombusvirus*	*Tombusviridae*	2	0	1	1	0
Grapevine Algerian latent virus	*Tombusvirus*	*Tombusviridae*	2	0	1	1	0
Havel river virus	*Tombusvirus*	*Tombusviridae*	1	0	0	1	0
Moroccan pepper virus	*Tombusvirus*	*Tombusviridae*	2	0	1	1	0
Pelargonium leaf curl virus	*Tombusvirus*	*Tombusviridae*	3	0	1	1	0
Pelargonium necrotic spot virus	*Tombusvirus*	*Tombusviridae*	3	0	1	1	0
Petunia asteroid mosaic virus	*Tombusvirus*	*Tombusviridae*	1	0	0	1	0

[1]Detections represent the number of libraries, out of the total of 306 libraries, in which the OTUs were detected.
[2]Presence (1) or absence (0) of viral OTUs in each habitat. Abbreviations, (Cr) Crop; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.

Table 4.8. Continued.

Species	Genus	Family	Detections[1]	Cr[2]	Ed[2]	Wl[2]	Oa[2]
Tomato bushy stunt virus	*Tombusvirus*	*Tombusviridae*	3	0	1	1	0
Dulcamara mottle virus	*Tymovirus*	*Tymoviridae*	1	0	0	0	1
Turnip yellow mosaic virus	*Tymovirus*	*Tymoviridae*	2	0	1	1	0
Bell pepper mottle virus	*Tobamovirus*	*Virgaviridae*	2	0	1	0	0
Cucumber green mottle mosaic virus	*Tobamovirus*	*Virgaviridae*	2	0	1	0	1
Obuda pepper virus	*Tobamovirus*	*Virgaviridae*	2	0	0	0	1
Paprika mild mottle virus	*Tobamovirus*	*Virgaviridae*	2	0	0	0	1
Pepper mild mottle virus	*Tobamovirus*	*Virgaviridae*	54	1	1	1	1
Rehmannia mosaic virus	*Tobamovirus*	*Virgaviridae*	20	1	1	1	1
Ribgrass mosaic virus	*Tobamovirus*	*Virgaviridae*	5	0	0	1	1
Tobacco mild green mosaic virus	*Tobamovirus*	*Virgaviridae*	110	1	1	1	1
Tobacco mosaic virus	*Tobamovirus*	*Virgaviridae*	81	1	1	1	1
Tomato brown rugose fruit virus	*Tobamovirus*	*Virgaviridae*	14	1	1	1	1
Tomato mosaic virus	*Tobamovirus*	*Virgaviridae*	22	1	1	1	1
Tomato mottle mosaic virus	*Tobamovirus*	*Virgaviridae*	9	1	1	1	1
Tropical soda apple mosaic virus	*Tobamovirus*	*Virgaviridae*	2	0	1	0	0
Turnip vein-clearing virus	*Tobamovirus*	*Virgaviridae*	6	0	0	1	1
Wasabi mottle virus	*Tobamovirus*	*Virgaviridae*	9	0	1	1	1
Youcai mosaic virus	*Tobamovirus*	*Virgaviridae*	28	1	1	1	1

[1]Detections represent the number of libraries, out of the total of 306 libraries, in which the OTUs were detected.
[2]Presence (1) or absence (0) of viral OTUs in each habitat. Abbreviations, (Cr) Crop; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.

The most common OTUs detected were: *Cucumber mosaic virus* (CMV) (in 61% of libraries), *Pelargonium zonate spot virus* (PZSV) (47%), *Tobacco mild green mosaic virus* (TMGMV) (36%), *Rubus chlorotic mottle virus* (RuCMV) (34%), *Tobacco mosaic virus* (TMV) (27%), *Watermelon mosaic virus* (WMV) (22%), *Tomato aspermy virus* (TAV) (19%), *Pepper mild mottle virus* (PMMoV) (18%), *Turnip yellows virus* (TuYV) (17%) and *Brassica yellows virus* (BrYV) (12%).

Seventeen out of 90 OTUs (19%) were detected in all habitats **(Annex, Table A4.1** and **Figure 4.11)**, seven OTUs (8%) in three habitats, 27 (30%) in two habitats, and 39 (43%) in a single habitat. Thus, a large fraction of OTUs showed habitat specificity. Of the 39 OTUs detected in a single habitat, the largest fraction occurred in Edge (69%), followed by Oakwood (13%), Wasteland (10%) and Crop (8%). Additionally, Edge was the habitat that shared a higher number of OTUs with other habitats **(Figure 4.11)**. The 90 OTUs detected belonged to 22 virus genera **(Table 4.8** and **Annex, Table 4.6)**, most of them with positive sense ssRNA genomes, except *Cytorhabdovirus*, *Nucleorhabdovirus* and *Varicosavirus* that have negative sense ssRNA genomes. Of the 22 virus genera, *Cucumovirus* were detected in 63% of libraries, *Tobamovirus* in 55%, *Anulavirus* in 47%, *Potyvirus* in 35%, *Sobemovirus* in 34%, *Polerovirus* in 30%, *Carmovirus* in 10%, *Ilarvirus* in 7%, *Luteovirus* in 3%, and *Varicosavirus* in 2% of HTS libraries. Six of 22 genera (27%) were detected in all habitats, three (14%) in three habitats, seven (32%) in two habitats, and 6 (27%) in a single habitat.

The 90 OTUs belonged to 8 virus families **(Table 4.8** and **Annex, Table 4.7)** all with positive sense ssRNA genomes, except *Rhabdoviridae*. The families detected in the libraries were *Bromoviridae* (in 73% of libraries)**,** *Virgaviridae* (55%), *Potyviridae* (35%), *Solemoviridae* (34%), *Luteoviridae* (31%), *Tombusviridae* (13%), *Rhabdoviridae* (3%), and *Tymoviridae* (1%). Six of eight virus families (63%) were detected in all habitats, while eight were present in three habitats.

A large fraction of OTUs were detected in Edge (74/90) and Wasteland (43/90). While Oakwood (34/90) and Crop (31/90) were the habitats with fewer

OTUs **(Annex, Table A4.2-5)**. Similarly, the largest fraction of genera (18/22) was detected in Edge **(Annex, Table A4.7-10** and **Annex, Table A4.12-15)**.

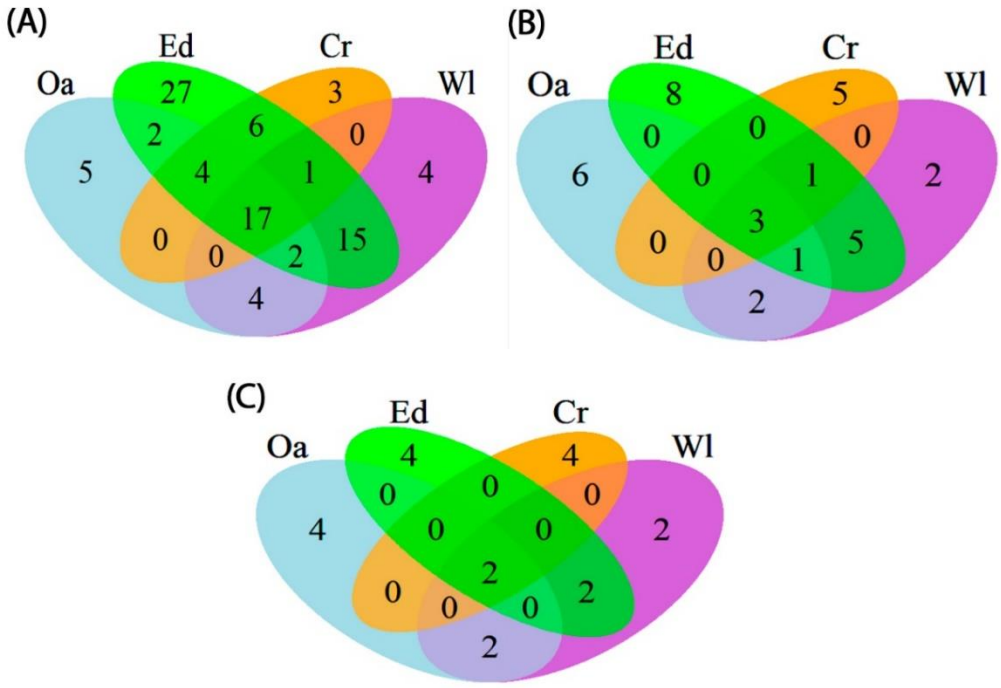

Figure 4.11. Venn diagrams showing the number of OTUs (A), virus genera (B) and families (C) detected in the different habitats. Abbreviations, (Cr) Crop; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.

Table 4.8. Habitat distribution of the number of virus OTUs, genera and families.

Taxonomic level	Crop	Edge	Wasteland	Oakwood	Total
Species	31	74	43	34	90
Genus	9	18	14	12	22
Family	6	8	8	8	8

4.4.4. Viral OTUs distributions over hosts and host range

The 306 HTS libraries in which validated OTUs were detected came from 106 plant species (host species) of 118 total plant species sequenced. These 106 plant species correspond to 32 families **(Table 4.9)**. The remaining 12 plant

species in which no virus OTUs were detected come from Wasteland and Oakwood habitats.

The top ten host species with highest OTU richness were: *Convolvulus arvensis* L. (*n* = 30 OTUs), *Picris echioides* L. (*n* = 28), *C. bourgeanus* (*n* = 25), *Rubia peregrina* L. (*n* = 24), *Diplotaxis erucoides* DC. (*n* = 24), *Lactuca serriola* L. (*n* = 23), *P. rhoeas* (*n* = 21), *Amaranthus sp* L. (*n* = 21), *Conyza canadensis* (L.) Cronquist (*n* = 20) and *Chenopodium album* L. (*n* = 20). These host species were from families: *Asteraceae* (*n* = 101), *Brassicaceae* (*n* = 53), *Rubiaceae* (*n* = 46), *Convolvulaceae* (*n* = 41), *Amaranthaceae* (*n* = 30) and *Papaveraceae* (*n* = 28).

The OTU richness distribution might be influenced by the taxa relative abundance or the number of plant species that correspond to each family. Kendall's tau rank correlation coefficient (τ) showed significant correlations between OTU richness distribution and taxa relative abundance (plant species: $z = 4.043$, $P < 0.001$, $\tau = 0.261$; plant family: $z = 4.601$, $P < 0.001$, $\tau = 0.581$) and the number of plant species that correspond to each family ($z = 3.673$, $P < 0.001$, $\tau = 0.491$). Thus, these results provide evidence of hosts preferences, which were not driven by the abundance of plant taxa but correlated with taxa diversity.

In Crop, OTUs were detected in 23 host species of 14 families. *C. melo* (*n* = 12), *C. bourgeanus* (*n* = 10), *C. arvensis* (*n* = 7), *Sisymbrium runcinatum* Lag. (*n* = 7), *D. erucoides* (*n* = 6), *R. peregrina* (*n* = 6), Desconocida 4 (*n* = 5), *Hirschfeldia incana* L. (*n* = 5), *Hordeum vulgare* L. (*n* = 5), and *Portulaca oleraceae* L. (*n* = 5) were the top ten host species with the highest OTU richness. These host species were from the families *Asteraceae* (*n* = 15), *Brassicaceae* (*n* = 12) *Cucurbitaceae* (*n* = 12), *Poaceae* (*n* = 10), *Convolvulaceae* (*n* = 7), *Portulacaceae* (*n* = 6), *Rubiaceae* (*n* = 6), Desconocida 4 (*n* = 5).

In Edge, OTUs were detected in 52 host species of 18 families. *C. arvensis* (*n* = 29), *P. echoides* (*n* = 26), *R. peregrina* (*n* = 24), *D. erucoides* (*n* = 23), *L. serriola* (*n* = 23), *Amaranthus sp.* Cronquist (*n* = 19), and *Brachypodium retusum* PERS. *(Beauv.)* (*n* = 18) were the top ten host species with the highest OTU richness. These host species were from the families *Asteraceae* (*n* = 48),

Poaceae (*n* = 35), *Brassicaceae* (*n* = 29), *Convolvulaceae* (*n* = 29), and *Amaranthaceae* (*n* = 26), *Rubiaceae* (*n* = 25).

In Wasteland, OTUs were detected in 41 host species of 16 families. *Centranthus calcitrapae* (L.) Dufr*.,* (*n* = 13), *Taraxacum officinale* L. (*n* = 9), *Verbascum sinuatum* L. (*n* = 9), *Asparagus acutifolius* L. (*n* = 8), *Bromus sp* L. (*n* = 8), *Bromus madritensis* L. (*n* = 8), *Galium verum* L. (*n* = 6), *Leontodon sp* L. (*n* = 6), *Avena sterilis* L. (*n* = 5), and *C. arvensis* (*n* = 5) were the top ten host species with the highest OTU richness. These host species were from the families *Asteraceae* (*n* = 20), *Poaceae* (*n* = 14), *Caprifoliaceae* (*n* = 13), *Scrophulariaceae* (*n* = 9), *Asparagaceae* (*n* = 8), *Rubiaceae* (*n* = 6), and *Convolvulaceae* (*n* = 5).

In Oakwood, OTUs were detected in 45 host species of 19 families. *Quercus coccifera* L. (*n* = 13), *Leontodon sp.* (*n* = 12), *Marrubium vulgare* L. (*n* = 9), *Teucrium capitatum* L. (*n* = 9), *A. acutifolius* (*n* = 7), *Dactylis glomerata* L. (*n* = 7), *G. verum* (*n* = 7), *Jasminum fruticans* L. (*n* = 7), *Staehelina dubia* L. (*n* = 7), and *Aristolochia pistolochia* L. (*n* = 6) were the top ten host species with the highest OTU richness. These host species were from the families *Asteraceae* (*n* = 18), *Lamiaceae* (*n* = 14), *Fagaceae* (*n* = 13), *Poaceae* (*n* = 11), *Asparagaceae* (*n* = 9), *Rubiaceae* (*n* = 9), *Oleaceae* (*n* = 7), and *Aristolochiaceae* (*n* = 6).

Table 4.9. Virus OTUs detection in plant hosts according to habitat.

Plant family[1]	Plant species	Habitat association	Number of virus OTUs[2]	Cr[3]	Ed[3]	Wl[3]	Oa[3]
Amaranthaceae	*Amaranthus sp* L.	Cr/Ed	21	3	20	-	-
Amaranthaceae	*Bassia scoparia* (L.) Voss	Ed	3	-	3	-	-
Amaranthaceae	*C. album*	Cr/Ed	20	1	20	-	-
Amaryllidaceae	*Allium sativum*	Ed	3	-	3	-	-
Apiaceae	*Anthriscus caucalis* Bieb	Ed	14	-	14	-	-
Apiaceae	*Bupleurum rigidum* L.	Oa	-	-	-	-	-
Apiaceae	*Daucus sp* L.	Ed	13	-	13	-	-
Apiaceae	*Eryngium campestre* L.	Wl/Oa	6	-	-	2	5
Apiaceae	*Foeniculum vulgare* Mill.	Wl	-	-	-	-	-
Apiaceae	*Scandix pecten-veneris* L.	Ed	9	-	9	-	-
Apiaceae	*Thapsia villosa* L.	Wl/Oa	6	-	-	5	2
Apiaceae	*Torilis nodosa* (L.) Gaertn.	Ed	5	-	5	-	-
Aristolochiaceae	*Aristolochia pistolochia* L.	Ed/Oa	12	-	7	-	6
Asparagaceae	*Aphyllanthes monspeliensis* L.	Oa	2	-	-	-	2
Asparagaceae	*Asparagus acutifolius* L.	Wl/Oa	11	-	-	8	7
Asteraceae	*Anacyclus clavatus* Desf. Pers.	Cr/Ed/Wl	8	1	6	3	-
Asteraceae	*Andryala arenaria* (DC.) Boiss*. & Reuter*	Oa	3	-	-	-	3
Asteraceae	*Artemisia vulgaris* L.	Wl/Oa	6	-	-	4	2
Asteraceae	*Asteriscus aquaticus* (L.) LESS	Wl/Oa	4	-	-	3	2
Asteraceae	*C. bourgeanus*	Cr/Ed/Wl/Oa	25	10	17	-	6

[1]The plant species and families were identified following the taxonomy of the botanical keys Flora Iberica (Castroviejo, - 1986-2012) and Flora Europea (Tutin *et al.,* 2010), respectively.
[2]Number of OTUs detected for each host species.
[3]Number of virus OTUs detected in each habitat. Abbreviations, (Cr) Crop; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.
[4] The hyphen ("-"), indicates an OTU was not detected.

Table 4.9. Continued

Plant family[1]	Plant species	Habitat association	Number of virus OTUs[2]	Cr[3]	Ed[3]	Wl[3]	Oa[3]
Asteraceae	*Centaurea melitensis* L.	Wl/Oa	6	-	-	1	6
Asteraceae	*Cirsium arvense* L.	Ed	8	-	8	-	-
Asteraceae	*C. bonariensis*	Ed	19	-	19	-	-
Asteraceae	*C. canadensis*	Ed	20	-	20	-	-
Asteraceae	*Dittrichia viscosa* (L.) Greuter	Wl	1	-	-	1	-
Asteraceae	*Helichrysum stoechas* (L.) Moench	Oa	5	-	-	-	5
Asteraceae	*Hieracium pilosella* L.	Wl	-	-	-	-	-
Asteraceae	*Klasea pinnatifida* (Cav.) Cass.	Cr/Oa	1	-	-	-	1
Asteraceae	*L. serriola*	Ed	23	-	23	-	-
Asteraceae	*Leontodon sp*	Ed/Wl/Oa	17	-	6	6	12
Asteraceae	*P. echioides*	Cr/Ed/Wl	28	4	26	3	-
Asteraceae	*Senecio jacobaea* L.	Ed	3	-	3	-	-
Asteraceae	*Silybum marianum* (L.) Gaertner	Ed/Wl	18	-	18	-	-
Asteraceae	*Sonchus oleraceus* L.	Ed	8	-	8	-	-
Asteraceae	*S. dubia*	Oa	7	-	-	-	7
Asteraceae	*T. officinale*	Wl	9	-	-	9	-
Asteraceae	*Tragopogon sp* L.	Wl	4	-	-	4	-
Asteraceae	*Xanthium strumarium* L.	Cr	1	1	-	-	-
Boraginaceae	*Anchusa undulata* L.	Ed/Wl	12	-	8	4	-

[1]The plant species and families were identified following the taxonomy of the botanical keys Flora Iberica (Castroviejo, - 1986-2012) and Flora Europea (Tutin *et al.,* 2010), respectively.
[2]Number of OTUs detected for each host species.
[3]Number of virus OTUs detected in each habitat. Abbreviations, (Cr) Crop; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.
[4] The hyphen ("-"), indicates an OTU was not detected.

Table 4.9. Continued

Plant family[1]	Plant species	Habitat association	Number of virus OTUs[2]	Cr[3]	Ed[3]	Wl[3]	Oa[3]
Boraginaceae	*Borago officinalis* L.	Wl	-	-	-	-	-
Boraginaceae	*Echium vulgare* L.	Ed/Wl	12	-	8	5	-
Boraginaceae	*Lithodora fruticosa* (L.) Griseb.	Oa	-	-	-	-	-
Boraginaceae	*Lithospermum arvense* L.	Cr/Ed	6	1	5	-	-
Brassicaceae	Desconocida 5	Cr/Wl	4	1	-	3	-
Brassicaceae	*Descurainia sophia* (L.) Webb ex Prantl	Ed/Wl	6	-	3	4	-
Brassicaceae	*D. erucoides*	Cr/Ed/Wl	24	6	23	3	-
Brassicaceae	*Diplotaxis sp* DC.	Ed	8	-	8	-	-
Brassicaceae	*Diplotaxis virgata* (Cav.) DC.	Oa	3	-	-	-	3
Brassicaceae	*Eruca vesicaria* (L.) Cav.	Ed	7	-	7	-	-
Brassicaceae	*H. incana*	Cr/Ed	10	5	9	-	-
Brassicaceae	*Lepidium draba* (L.) Desv.	Ed	16	-	16	-	-
Brassicaceae	*S. runcinatum*	Cr/Wl	8	7	-	1	-
Caprifoliaceae	*C. calcitrapae*	Wl/Oa	18	-	-	13	5
Cistaceae	*Helianthemum cinereum* (Cav.) Pers.	Oa	-	-	-	-	-
Convolvulaceae	*C. arvensis*	Cr/Ed/Wl	30	7	29	5	-
Cucurbitaceae	*C. melo*	Cr	12	12	-	-	-
Cyperaceae	*Cyperus longus* L.	Cr	2	2	-	-	-
Fabaceae	*Astragalus incanus* L.	Oa	5	-	-	-	5

[1]The plant species and families were identified following the taxonomy of the botanical keys Flora Iberica (Castroviejo, - 1986-2012) and Flora Europea (Tutin *et al.,* 2010), respectively.
[2]Number of OTUs detected for each host species.
[3]Number of virus OTUs detected in each habitat. Abbreviations, (Cr) Crop; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.
[4] The hyphen ("-"), indicates an OTU was not detected.

Table 4.9. Continued

Plant family[1]	Plant species	Habitat association	Number of virus OTUs[2]	Cr[3]	Ed[3]	Wl[3]	Oa[3]
Fabaceae	*Astragalus sesameus* L.	Wl	-	-	-	-	-
Fabaceae	*Lotus corniculatus* L.	Oa	6	-	-	-	6
Fabaceae	*Medicago orbicularis* (L.) Bartal.	Wl	2	-	-	2	-
Fabaceae	*Medicago sp* L.	Wl/Oa	2	-	-	1	2
Fabaceae	*Melilotus sp* L.	Oa	1	-	-	-	1
Fabaceae	*Retama sphaerocarpa* L.	Wl	-	-	-	-	-
Fabaceae	*Trifolium campestre* Schreb. in Sturm	Ed	3	-	3	-	-
Fabaceae	*Trifolium tomentosum* L.	Wl	1	-	-	1	-
Fabaceae	*Vicia sp* L.	Ed/Wl	16	-	16	1	-
Fagaceae	*Q. coccifera*	Oa	13	-	-	-	13
Fagaceae	*Q. ilex*	Oa	5	-	-	-	5
Geraniaceae	*Erodium cicutarium* (L.) L'Hérit. ex Alton	Cr/Ed/Wl/Oa	11	1	9	1	3
Geraniaceae	*Geranium sp* L.	Ed/Wl/Oa	6	-	4	1	3
Hypericaceae	*Hypericum pubescens* Boiss.	Oa	2	-	-	-	2
Labiatae	*Origanum vulgare* L.	Oa	-	-	-	-	-
Lamiaceae	*Lavandula latifolia* Medik.	Oa	-	-	-	-	-
Lamiaceae	*M. vulgare*	Oa	9	-	-	-	9
Lamiaceae	*Phlomis lychnitis* L.	Wl/Oa	5	-	-	3	2
Lamiaceae	*Phlomis purpurea* L.	Wl	-	-	-	-	-

[1]The plant species and families were identified following the taxonomy of the botanical keys Flora Iberica (Castroviejo, - 1986-2012) and Flora Europea (Tutin *et al.,* 2010), respectively.
[2]Number of OTUs detected for each host species.
[3]Number of virus OTUs detected in each habitat. Abbreviations, (Cr) Crop; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.
[4] The hyphen ("-"), indicates an OTU was not detected.

Table 4.9. Continued

Plant family[1]	Plant species	Habitat association	Number of virus OTUs[2]	Cr[3]	Ed[3]	Wl[3]	Oa[3]
Lamiaceae	*Salvia verbenaca* L.	Wl	2	-	-	2	-
Lamiaceae	*Teucrium botrys* L.	Oa	5	-	-	-	5
Lamiaceae	*T. capitatum*	Wl/Oa	9	-	-	-	9
Lamiaceae	*Teucrium chamaedrys* L.	Wl/Oa	5	-	-	3	2
Lamiaceae	*Teucrium pseudochamaepitys* L.	Wl/Oa	3	-	-	1	3
Lamiaceae	*Thymus vulgaris* L.	Oa	5	-	-	-	5
Liliaceae	*Asphodelus aestivus* BROT	Oa	2	-	-	-	2
Malvaceae	*Malva sylvestris* L.	Ed	13	-	13	-	-
Oleaceae	*J. fruticans*	Oa	7	-	-	-	7
Orobanchaceae	*Odontites luteus* (L.) Clairv.	Oa	5	-	-	-	5
Papaveraceae	*Fumaria parviflora* Lam.	Cr/Ed	10	2	8	-	-
Papaveraceae	*P. rhoeas*	Ed/Wl/Oa	21	-	17	4	5
Plantaginaceae	*Plantago sp* L.	Wl	2	-	-	2	-
Poaceae	*Avena sterilis* L.	Wl	5	-	-	5	-
Poaceae	*Avenula bromoides* ((Gouan) H.Scholz)	Oa	2	-	-	-	2
Poaceae	*Brachypodium phoenicoides* Roemer & Schultes	Ed	8	-	8	-	-
Poaceae	*B. retusum*	Ed	18	-	18	-	-
Poaceae	*Bromus sp*	Cr/Ed/Wl/Oa	17	3	11	8	4
Poaceae	*Cynodon dactylon* (L.) Pers.	Ed	16	-	16	-	-

[1]The plant species and families were identified following the taxonomy of the botanical keys Flora Iberica (Castroviejo, - 1986-2012) and Flora Europea (Tutin *et al.,* 2010), respectively.
[2]Number of OTUs detected for each host species.
[3]Number of virus OTUs detected in each habitat. Abbreviations, (Cr) Crop; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.
[4] The hyphen ("-"), indicates an OTU was not detected.

Table 4.9. Continued

Plant family[1]	Plant species	Habitat association	Number of virus OTUs[2]	Cr[3]	Ed[3]	Wl[3]	Oa[3]
Poaceae	*D. glomerata*	Wl/Oa	8	-	-	3	7
Poaceae	*Festuca sp* L.	Wl	3	-	-	3	-
Poaceae	*B. madritensis*	Wl/Oa	10	-	-	8	3
Poaceae	*H. vulgare*	Cr	5	5	-	-	-
Poaceae	*Lolium perenne* L.	Ed	17	-	17	-	-
Poaceae	*Milium vernale* Bieb.	Ed	16	-	16	-	-
Poaceae	*Phalaris minor* Retz.	Ed	6	-	6	-	-
Poaceae	*Grass*	Ed	6	-	6	-	-
Poaceae	*S. parviflora*	Wl/Oa	6	-	-	-	6
Poaceae	*Zea mays* L.	Cr	3	3	-	-	-
Polygonaceae	*Rumex pulcher* L.	Ed/Wl	11	-	11	-	-
Portulacaceae	*P. oleraceae*	Cr	5	5	-	-	-
Resedaceae	*Reseda lutea* L.	Wl/Oa	5	-	-	1	4
Rhamnaceae	*Rhamnus lycioides* L.	Oa	-	-	-	-	-
Rosaceae	*Crataegus monogyna* Jacq.	Oa	1	-	-	-	1
Rosaceae	*Potentilla sp* L.	Ed	5	-	5	-	-
Rubiaceae	*G. verum*	Ed/Wl/Oa	15	-	9	6	7
Rubiaceae	*R. peregrina*	Cr/Ed/Wl/Oa	24	6	24	1	2
Scrophulariaceae	*V. sinuatum*	Wl	9	-	-	9	-

[1]The plant species and families were identified following the taxonomy of the botanical keys Flora Iberica (Castroviejo, - 1986-2012) and Flora Europea (Tutin *et al.,* 2010), respectively.
[2]Number of OTUs detected for each host species.
[3]Number of virus OTUs detected in each habitat. Abbreviations, (Cr) Crop; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.
[4] The hyphen ("-"), indicates an OTU was not detected.

Table 4.9. Continued

Plant family[1]	Plant species	Habitat association	Number of virus OTUs[2]	Cr[3]	Ed[3]	Wl[3]	Oa[3]
Solanaceae	*Datura stramonium* L.	Cr/Ed	11	2	11	-	-
Solanaceae	*Solanum nigrum* L.	Cr/Ed	9	-	9	-	-
Desconocida 4	Desconocida 4	Cr/Ed	8	5	5	-	-

[1]The plant species and families were identified following the taxonomy of the botanical keys Flora Iberica (Castroviejo, - 1986-2012) and Flora Europea (Tutin *et al.,* 2010), respectively.
[2]Number of OTUs detected for each host species.
[3]Number of virus OTUs detected in each habitat. Abbreviations, (Cr) Crop; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.
[4] The hyphen ("-"), indicates an OTU was not detected.

The variation in the composition of each host community may correspond to variation in available resources, ecological interactions, and affect the host range of viruses. Therefore, the host range of the virus OTUs was analysed at the landscape **(Figure 4.12.A)** and habitat levels **(Figure 4.12.B)**.

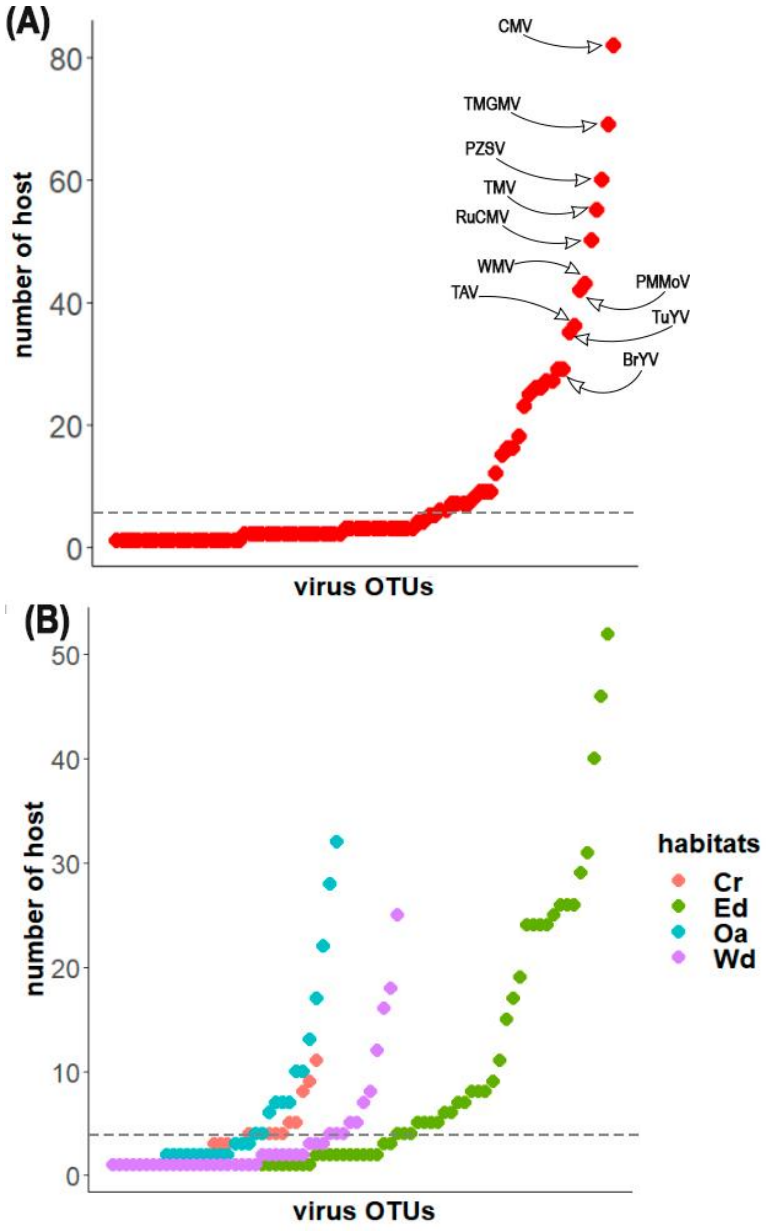

Figure 4.12. Host range of the virus OTUs detected over the landscape (A) and for each habitat (B). The x axis represents virus OTUs, one dot for each, and the y axis represents the number of plant host species that in which the virus OTU was detected. Abbreviations, (Cr) Crop; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.

For the total set of virus OTUs detected in the landscape it was shown most (56/90 (62.2%)) had a narrow (below 5) host range, while a smaller set

(17/90 (18.9%)) had host ranges of 20 or more species. *Cucumber mosaic virus* (CMV) was the virus OTU with the largest host range (82 host species). The virus host range showed a significant variation between habitats (Mann-Whitney U test: $U = 158$, $P \leq 0.001$).

Additionally, the number of virus OTUs that infect each plant species was also evaluated over the landscape **(Figure 4.13.A)** and in each habitat **(Figure 4.13.B)**.

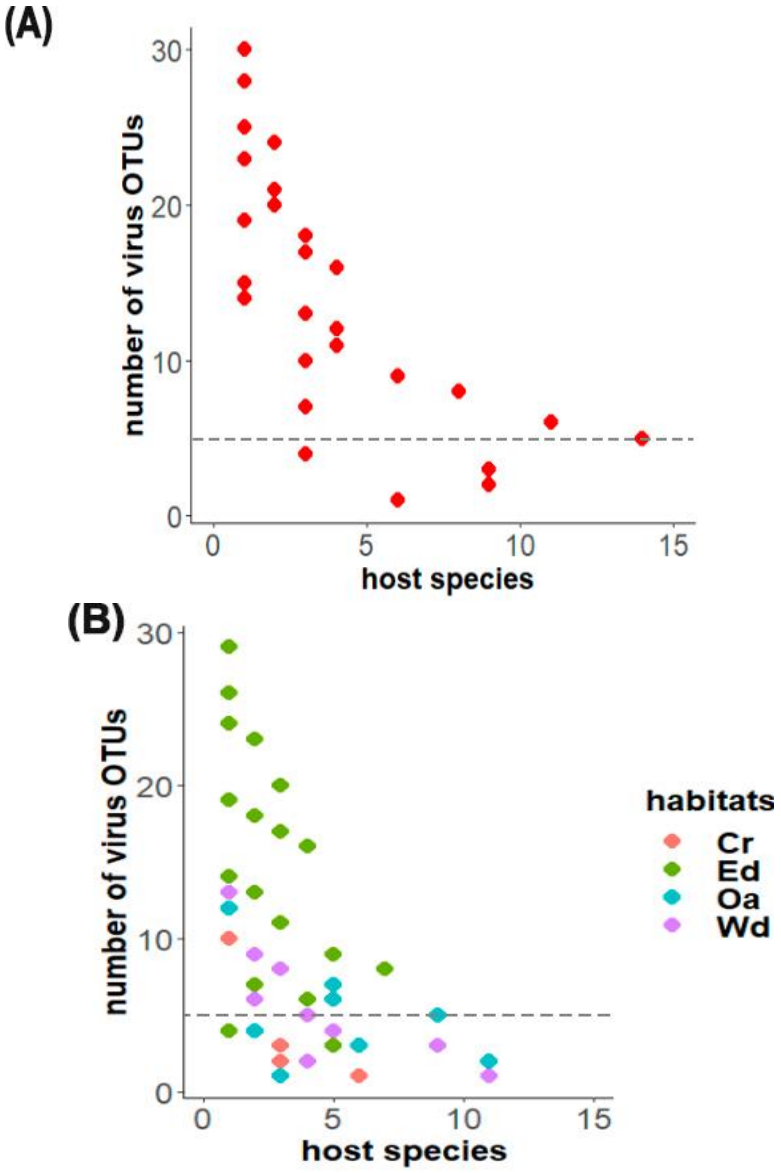

Figure 4.13. Number of virus OTUs that infect each host species over the landscape (A) and in each habitat (B). Abbreviations, (Cr) Crop; (Ed) Edge; (Wl) Wasteland; and (Oa) Oakwood.

These results showed that a large fraction of the plant species (65 host species) was infected by more than five virus OTUS. While there was a subset of less permissive plant species (41 host species) that were infected by less than five virus OTUs. Plants growing in Edge were infected, (on the average) by more virus OTUs (11.7%) than those in other habitats (Crop: 4.0%; Wasteland: 3.7%; Oakwood: 4.5%), although this difference was no significant (Mann-Whitney U test: $U = 938$, $P = 0.168$).

After describing the associations of OTUs with host species of the four habitats, we measured the link between virus OTUs and the habitats when these communities are described by host relative abundance (Dray & Legendre, 2008). The extension of the fourth-corner method proposed by Legendre *et al.* (1997) that was used to measure the link between virus species and the habitats, depends on the associations among three matrices. The fourth-corner is solved by linking three matrices: the classification of each site into a particular habitat (R); host species abundances per site (L); and virus OTUs presence on each host plant species (Q) **(see section 3.5.3.)**. The significance of the relationship between virus OTUs (Q) and habitat (R), linked by host species abundance (L), was tested with a two-step procedure, using 9999 random permutations, and adjusting *P*-values for multiple testing (Benjamini & Hochberg, 1995).

The fourth-corner method showed that 11 OTUs in Edge had a significant relationship with high host species relative abundances at sites of Edge ($P < 0.036$). Only one OTU had a significant relationship with low host species relative abundance in Oakwood ($P = 0.011$) **(Figure 4.14)**. Therefore, these findings indicated that the distribution of viruses among the habitats in the ecosystem, as associated with host species abundance, were most strong within Edge.

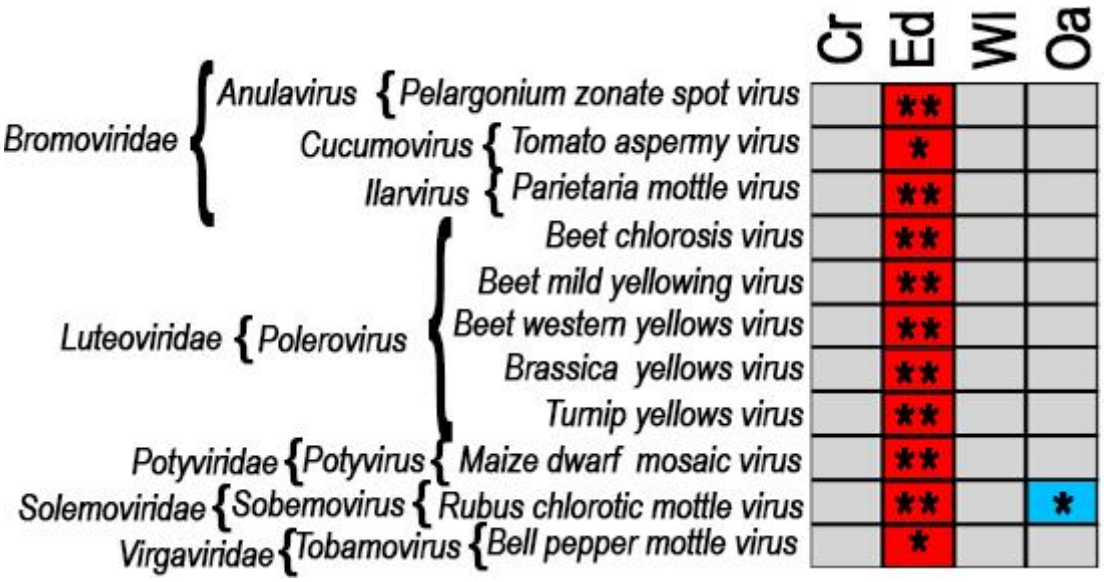

Figure 4.14. Relationship between the virus OTUs and habitats. Only significant relationships are shown. The blue tiles correspond to significant and negative relationships while the red tiles correspond to significant and positive relationships (*$P < 0.05$, **$P < 0.01$).

4.5. A plant virus in a heterogeneous environment

To assess the contribution of the diversity of plant communities of agricultural ecosystems in the structuration of the incidence, genetic diversity, and evolution of viruses, the host range of WMV was investigated. The host range of WMV was determined from HTS data as well as detection by RT-PCR. From validated BLAST queries of reads against a local database of ssRNA plant virus genomic references, 66 of 306 HTS libraries were positive for *Watermelon mosaic virus* (WMV). The BLAST queries revealed a host range of 43 species across the four habitats. In contrast to the HTS estimates of host range, detection by RT-PCR revealed 116 WMV positive samples in 17 host species. In addition, RT-PCR analyses of plant species that were not analysed by HTS increased the host range by 7, to a total of 24 species. By targeting the CP cistron in RNA extracts from individual plant samples, RT-PCR was used to estimate WMV host range, as well as incidence over hosts and habitats, and to characterise the genetic structure **(see section 3.5.1.)** of its populations.

4.5.1. WMV host range and incidence

The validation of host range by RT-PCR showed a narrower host range than by HTS. Except for Crop, PCR-validated host range estimates were narrower across the habitats compared to those of the HTS **(Table 4.10)**. Detection by RT-PCR may be subject to poor sensitivity in the detection of low-titre infections, genetic variants that were not detected due to primer non-specificity, the high variation in the quality of the RNA preparations or the size of the amplicon.

Table 4.10. Host range of WMV estimated from HTS and RT-PCR data.

Habitat	Host species[1]	WMV positive by HTS	WMV positive by RT-PCR
Cr	39	9	10
Ed	56	19	7
Wl	71	4	3
Oa	65	17	7
Total	144	43	24

[1]Abundant host plants species collected in the first year of sampling (2015-2016).

Each habitat supported different assemblages of host species of WMV **(Table 4.11)**, with 5 of the 24 occurring in three habitats, 8 in two habitats, and 11 being specific to one habitat.

The incidence of WMV estimated from all host species was 14.00% (116 positive samples out of 831 analysed). Incidence was highest in Crop at 48.00% (73/152), followed by Oakwood at 14.90% (28/188), Edge at 4.90% (12/246), and lowest in Wasteland at 1.22% (3/245). However, when calculated with the inclusion of all plant species in the habitat, regardless of whether they were hosts of WMV or not, incidence decreased to 4.20% (116/2759). Incidence remained highest in Crop (38.22%, 73/191), followed by Oakwood and Edge (2.75%, 28/1019 and 2.37%, 12/507, respectively) and was again lowest in Wasteland (0.29%, 3/1042).

Table 4.11. Incidence (%) of WMV across its host range.

Host taxon	Crop	Edge	Wasteland	Oakwood
Amaranthus sp.	53.9% (13)[1]	0% (19)	-	-
A. undulata	-	0% (2)	3.5% (29)	-
B. retusum	-	5% (20)	% (-)	-
Bromus sp.	-	0% (8)	3.1% (32)	0% (12)
B. madritensis	-	-	4.2% (24)	0% (2)
C. bourgeanus	-	5.9% (17)	0% (53)	0% (8)
C. album	44.4% (18)	6.7% (15)	-	-
C. tinctoria	33.3% (3)	-	-	-
C. salviifolius	-	-	-	7.1% (28)
C. arvensis	38.9% (18)	0% (60)	0% (21)	-
C. canadensis	-	20% (5)	-	-
C. melo	63.2% (57)	-	-	-
C. dactylon	-	6.7% (15)	-	-
D. erucoides	33.3% (3)	7.5% (40)	0% (27)	-
H. stoechas	-	-	-	80% (5)
L. arvense	22.7% (22)	0% (7)	-	-
P.s minor	66.7% (6)	-	-	-
P. echioides	100% (1)	14.3% (28)	0% (11)	-
Q. coccifera	-	-	-	12.7% (63)
R. lycioides	-	-	-	2.5% (40)
S. nigrum	27.3% (11)	0% (4)	-	-
S. dubia	-	-	-	28.6% (21)
T. officinale	-	-	0% (48)	50% (2)
T. botrys	-	-	-	85.7% (7)
Total[2]	48% (152)	4.9% (246)	1.2% (245)	14.9% (188)
Grand total[3]	38.2% (191)	2.4% (507)	0.3% (1042)	2.8% (1019)

[1]Data indicate incidence as percentage (number of analysed samples). A dash ("-") indicates that there is no record of the host in that habitat.
[2]WMV total incidence using host species from each habitat.
[3]WMV total incidence using host and non-host species from each habitat.

To test the differences of WMV incidence among habitats, non-parametric permutation *F*-tests and *t*-tests were performed to compare the observed statistic with the sampling (null) distribution using 1,000 iterations **(see section 3.5.1.)**. The *F*-test indicated that Crop had significantly different variance in incidence compared to Edge (*F*-test, $P = 0.006$) and Wasteland (*F*-test, $P = 0.006$), but not Oakwood (*F*-test, $P = 0.24$) **(Table 4.12)**, and Edge had significantly different variance in incidence compared to Oakwood (*F*-test, $P = 0.036$). Additionally, the two-sided *t*-test showed that the incidence was significantly higher in Crop compared to Edge (*t*-test, $P = 0.006$) and Wasteland (*t*-test, $P = 0.006$). The differences in incidence between Crop and Oakwood were not significant (*t*-test, $P = 0.116$) due to over-dispersion of incidence values in these two habitats.

Table 4.12. Permutation *F*-tests and *t*-tests of differences in incidence between the four habitats.

	Crop	Edge	Wasteland	Oakwood
Crop	-	**0.006**	**0.006**	0.24
Edge	**0.006**[1]	-	0.055	**0.036**
Wasteland	**0.006**	0.116	-	**0.039**
Oakwood	0.116	0.054	0.054	-

[1]*P*-values of *F*-tests are above the diagonal and *t*-tests below. Significant differences indicated with bold text. Benjamini & Hochberg adjusted *P*-values are shown.

The correlations between the incidence of WMV and host species relative abundance in each habitat were evaluated. The incidence of WMV was correlated with host relative abundance in Crop ($R^2 = 0.91$, $F_{(1,8)} = 97.02$, $P < 0.001$), but not in the other habitats. Thus, these results provide evidence of hosts preferences either in Edge, Wasteland, or Oakwood, which were not driven by the abundance of any given plant species.

The incidence data shown in **Table 4.13** corresponds to infection counts across seasons. Incidence was found to be highest in summer (24.92%, 81/325) and lowest in autumn (0.98%,11/1122).

Table 4.13. Incidence (%) of WMV across each season.

Season	Total individuals
Summer	24.9% (325)[1]
Spring	1.83% (1312)
Autumn	0.98% (1122)

[1]Data indicate incidence as percentage (number of analysed samples).

Season differentially affected the incidence of WMV according to habitat **(Table 4.14)**. The frequencies of WMV incidence for each habitat in spring were no correlated with the frequencies of WMV incidence for each habitat in autumn ($\chi^2_{(2, 4)} = 0.001$, $P = 0.999$). Therefore, season and habitat are no dependent.

Table 4.14. Incidence (%) of WMV across each habitat for each season.

Habitat/season	Crop	Edge	Wasteland	Oakwood
Spring	-	1.13% (177)[1]	0.35% (572)	3.55% (563)
Autumn	-	1.02% (196)	0.21% (470)	1.75% (456)
Summer	38.22% (191)	5.97% (134)	-	-

[1]Data indicate incidence as percentage (number of analysed samples). A dash ("-") indicates that there is no record of incidence in that habitat-season.

4.5.2. Analysis of WMV genetic diversity

The PCR amplicons of the WMV-CP were sequenced **(see section 3.3.4.)** and used to assess the distribution of WMV haplotypes among the habitats and its host species. The number of samples of each host species of WMV was unbalanced. To better balance the representation of each host species in the sequence data, a maximum of five PCR products of a specific host plant species of a collection was sent for sequencing. The number of samples of each host sent for sequencing are listed in the **Annex, Table A5**. Consensus sequences of 108 WMV-CP genes were obtained.

One hundred of the 108 consensus sequences covered all 843 nucleotides of the CP cistron. The remaining sequences covered 831 nucleotides of the CP

cistron **(see section 3.4.5.)**. The accession number of each sequence, details of host species, and location of each collection are given in **Annex, Table A6**.

The alignment of all consensus sequences was used to determine the nucleotide (π) and haplotype (H_d) diversity of WMV populations defined by either habitat or host plant species affiliation. To minimise spurious estimates of genetic diversity due to small sample sizes (Goodall-Copestake *et al.,* 2012) obtained for some populations, sequences from Wasteland (n = 3) were clustered with those from Oakwood (n = 25) and categorised as Wild habitat (Wd) (n = 28). Together with sequences from Edge (n = 12) and Crop (n = 68), the analysis constituted three habitats. Similarly, sequences of the 24 host species of WMV were divided into those from *C. melo* (n = 32) and those from Wild hosts species (n = 76). The variation in haplotype (H_d range; 0.740 to 0.773) and nucleotide (π range; 0.026 to 0.036) diversity was negligible between each habitat. Wild hosts supported higher haplotype (H_d = 0.850) and nucleotide (π = 0.218) diversity than *C. melo* **(Table 4.15)**. The permutation procedure showed that the pairwise comparison of the nucleotide and haplotype diversity between populations at both habitat (H_d: P > 0.604 and π: P > 0.343) and host species levels were not significant (H_d: P > 0.427 and π: P > 0.301).

Table 4.15. Haplotype diversity (H_d), nucleotide diversity (π) values of WMV populations at habitat and host level.

Level	Group	n[1]	H_d ± S.E.[2]	π ± S.E.[2]
Habitat	Crop (Cr)	68	0.772 ± 0.046	0.036 ± 0.018
	Edge (Ed)	12	0.773 ± 0.128	0.026 ± 0.014
	Wild habitat (Wd)	28	0.740 ± 0.084	0.035 ± 0.017
Host	*C. melo*	32	0.802 ± 0.049	0.191 ± 0.097
	Wild hosts	76	0.850 ± 0.031	0.218 ± 0.108

[1]n: total sample size (number of sequences).
[2]S.E.: standard error.

Selection on the CP was analysed by the estimation of ω **(see section 3.5.1.)**, which indicated purifying selection (d_N / d_S < 1) **(Table 4.16)**. Evidence of purifying selection and conservation of the CP was consistent with other studies.

Table 4.16. d_N/d_S ratio (ω) of WMV populations at habitat and host level.

Level	Group	$\omega \pm$ SE
Habitat	Cr	0.087 ± 0.004
	Ed	0.137 ± 0.036
	Wd	0.162 ± 0.010
Host	*C. melo*	0.067 ± 0.009
	Wild hosts	0.101 ± 0.003

4.5.3. Analysis of WMV population structure

Analysis of molecular variance (AMOVA) was used to test hypotheses of population structure of WMV due to either habitat or host affiliation. The *Phi*-statistic (Φ_{ST}) showed significant differentiation (Φ_{ST} = 0.224, P < 0.001) among the three populations defined by the habitats, which explained 22% of the total genetic variance. There was also a significant difference (Φ_{ST} = 0.139, P < 0.001) between the population of *C. melo* and the population of Wild host species, which explained 14% of the total genetic variance **(Annex, Table A7)**. Population differentiation was relatively strong between Wild and Crop (Φ_{ST} = 0.260, P < 0.001) and between Wild and Edge (Φ_{ST} = 0.350, P < 0.001) habitats compared to between Crop and Edge (Φ_{ST} = 0.050, P = 0.093) **(Table 4.17)**.

The AMOVA indicated that although habitat (22% of variance) and host (14% of variance) had a role in the genetic differentiation of WMV, within population differentiation was higher than between them (Habitat: 78% of variance; Host: 86% of variance).

Table 4.17. Degree of population differentiation (lower diagonal) and associated significance indication (upper diagonal).

Φ_{ST}	Cr (68)	Ed (12)	Wd (28)
Cr (68)[1]	--	0.093	0.000
Ed (12)	0.050	--	0.000
Wd (28)	0.260	0.350	--

[1]The number of sequences for each habitat is shown in brackets.
[2]Abbreviations are: (Cr) Crop; (Ed) Edge; (Wd) Wild habitats.

4.5.4. Phylogenetic analysis of WMV

The presence of recombinant sequences produces cyclical relationships among lineages rather than bifurcating relationships, and affects the accuracy of phylogenetic inference (Posada & Crandall, 2002). Therefore, we used breakpoint detection at the WMV CP, implemented in the RDP4 package **(see section 3.5.2.)**. No recombination breakpoints were detected in the 108 sequences alignment of the WMV-CP.

To evaluate the evolutionary relationships among habitat and host species associations of WMV haplotypes, a Bayesian maximum credibility consensus phylogenetic tree of the WMV CP cistron was inferred with MrBayes 3.2.7a **(Figure 4.15)**, using a general time reversible DNA substitution model (GTR) with gamma distributed (+G) rates and a proportion of invariant sites (+I) as the best-fit substitution model **(see section 3.5.2.)**. It was determined that the convergence and posterior parameter distributions of the chains reached stationarity.

The Bayesian maximum clade credibility consensus tree revealed four major clades of WMV haplotypes **(Figure 4.15)**. The clades A and B were well-supported, each with a posterior probability (*PP*) of 0.99, and each included lineages recovered from Crop, Edge, and Oakwood. Clade A also included lineages from Wasteland. The lineages of clade A compromised sequences associated with 14 host species, clade B with 9 species, and had 6 species shared between them. Clade C was well-supported (*PP* = 1.00) and included lineages associated with all four habitats. Lineages of clade C were associated with 9 host species, 4 species of which were shared with clades A and B. The other well-supported (*PP* = 1.00) clade D included lineages associated with only Oakwood, and compromised 5 host species that were shared among the other three major clades.

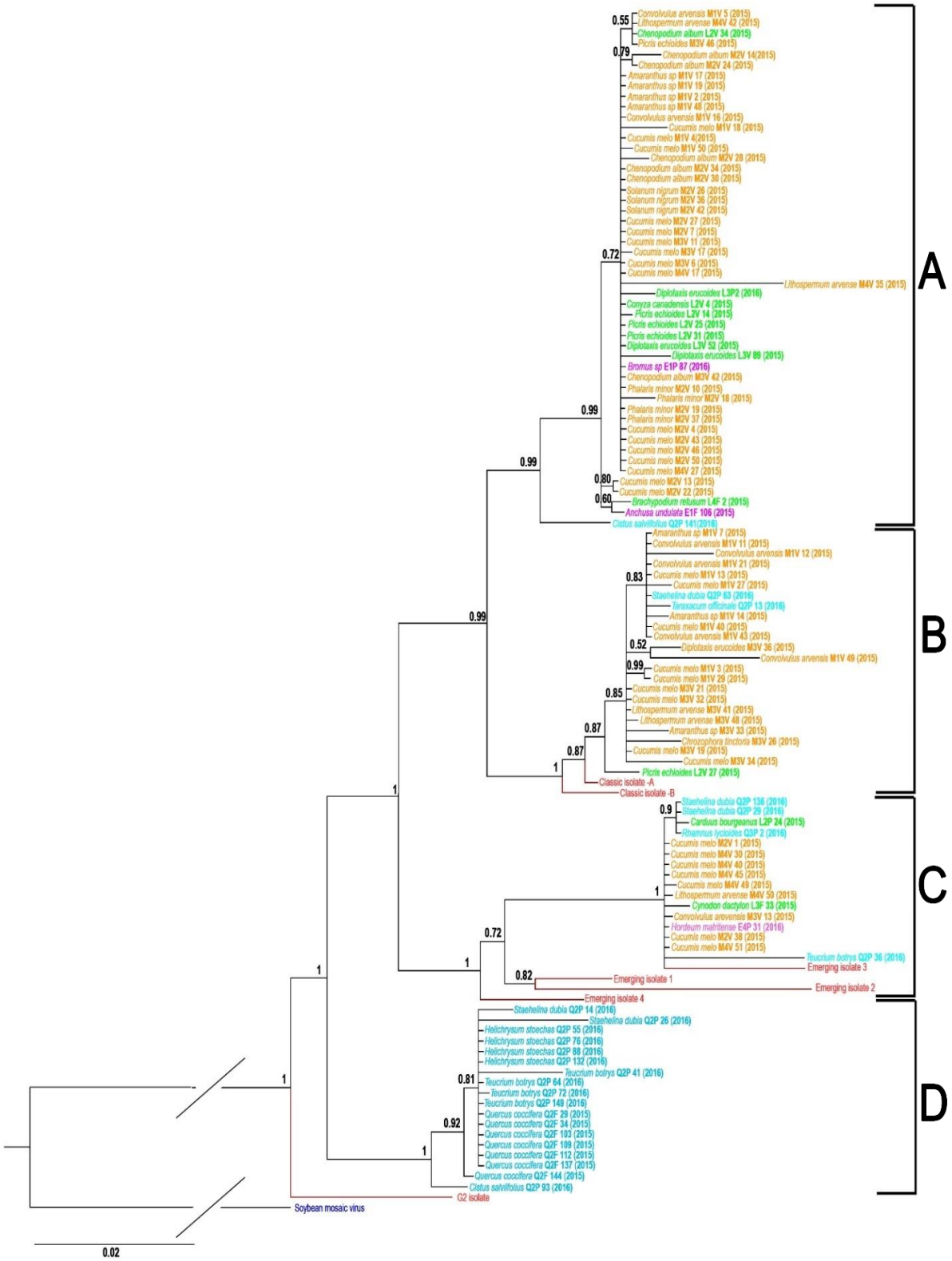

Figure 4.15. Maximum credibility Bayesian consensus phylogeny based on a 831 nucleotide region of the CP gene of WMV. The name of the host of each sequence is shown and coloured according to habitat; Crop (orange), Edge (green), Oakwood (blue), Wasteland (purple). The four major clades were designated A to D. Sequences representatives of the genetic groups proposed by *Desbiez et al.* (2009) are coloured in red **(see section 3.5.2)**. Outgroup: *Soybean mosaic virus* (SMV: Genbank Accession D00507).

Sequences from other work were introduced to the phylogenetic reconstruction and intended as references to genetic groups proposed by Desbiez *et al*. (2009). The “emerging” reference isolate clustered with lineages of clade C of the Bayesian consensus tree. Reference sequences that belonged to the “classic” strains A and B, clustered with lineages of clade B of the consensus tree.

The Bayesian consensus tree showed strong backbone support for most deep stem clades (*PP* > 0.98). Nearly all poorly resolved relationships (*PP* < 0.80) were inferred within the major clades and may be a consequence of saturation of substitutions (too much homoplasy) (Carbone *et al.,* 2004). The poorly resolved relationships of the consensus tree were consistent with the reticulated relationships shown in the haplotype network **(Figure 4.16)**. The haplotype network formed four major clusters of haplotypes. The clusters A, B and C included haplotypes with mixed host/habitat associations. By contrast, cluster D included only haplotypes from Oakwood, but with mixed host associations.

To test the influence of habitat and host species in the diversification of haplotypes of WMV a quantitative analysis was carried out using Bayesian Tip-association Significance tests (*BaTs*) **(see section 3.5.1.)**. The AI and PS statistics for Habitat and Host tip associations were both significant (*P* < 0.001) **(Table 4.18)**. The inverse of the PS statistic is proportional to the strength of the tip associations. The AI statistic measures the imbalance of internal nodes of a phylogeny, and the lower the value, the stronger the association is with the phylogeny. Both statistics indicated that Habitat associations (PS: 1/23.8 = 0.042; AI: 2.75) were stronger than Host associations (PS: 1/29.8 = 0.034; AI: 3.27).

Similarly, the monophyletic clade (MC) size index indicated that here was a stronger association with habitat. Higher values of the MC size are positively correlated with the strength of tip associations. There were significant associations of WMV haplotypes from wild habitat (17.968; *P* < 0.001) and from Crop (6.996; *P* = 0.026), but not from Edge (1.277; *P* = 1.000). Additionally, the MC index showed significant host species associations of WMV haplotypes and wild species (5.237; *P* = 0.010) and with *C. melo* host (1.097; *P* = 0.010).

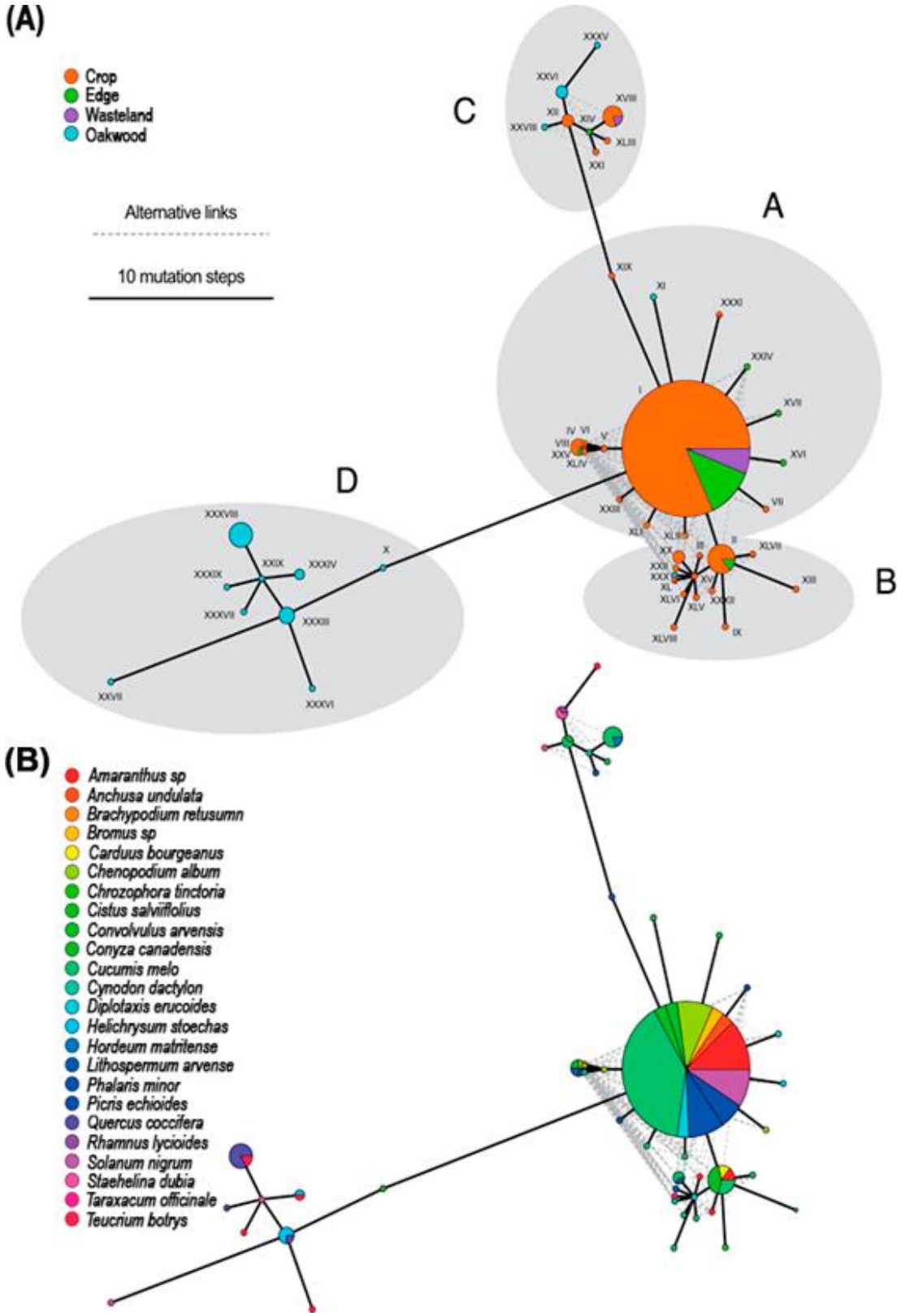

Figure 4.16. Maximum parsimony haplotype network constructed using the infinite sites model. Network (A) depicts the frequency of haplotypes associated with the four habitats, while network (B) shows host species associations. The groups (A to D) that we nominated are indicated with grey ellipses in network A only, and are to assist cross-comparisons.

Table 4.18. Bayesian Tip-association Significance test for habitat and host factors.

Factors	Statistic	Average observed CI (95%)[4]	Null average CI (95%)	*P*-value[5]
Habitat	AI[1]	2.75 (1.84 - 3.69)	5.85 (5.29 - 6.35)	<0.001***
	PS[2]	23.79 (22.00 - 25.00)	43.68 (41.55 - 45.34)	<0.001***
	MC[3] Cr	7.00 (4.00 - 11.00)	4.49 (3.62 - 5.82)	0.026*
	MC[3] Ed	1.28 (1.00 - 2.00)	1.26 (1.07 - 1.67)	1.000
	MC[3] Wd	17.97 (18.00 - 18.00)	2.02 (1.72 - 2.37)	<0.001***
Host	AI[1]	3.27 (2.24 - 4.31)	4.92 (4.49 - 5.37)	<0.001***
	PS[2]	29.81 (27.00 - 33.00)	35.84 (33.92 - 37.32)	<0.001***
	MC[3] *C.melo*	2.00 (1.00 - 3.00)	1.10 (1.00 - 1.23)	0.010**
	MC[3] Wild hosts	17.98 (18.00 - 18.00)	5.24 (4.14 - 7.20)	0.010**

[1]AI, association index.
[2]PS, parsimony score.
[3]Mc monophyletic clade size.
[4]CI, confidence interval.
[5]*P*-value. Significant values are marked with a * (*P<0.05, ** P <0.01, *** P <0.001).

5. DISCUSSION

The socio-economic relevance of emergent viruses has prompted a considerable research effort aimed at understanding the process of emergence (Elena *et al.,* 2014). Virus emergence depends on a range of factors intrinsic to the virus, such as genetic traits that determine its fitness in different hosts (i.e., evolutionary factors), or extrinsic to the virus, related to its ecology and epidemiology that ultimately result in the virus encountering and infecting a new host population (i.e., ecological factors) (Jones, 2009; McLeish *et al.,* 2019). Therefore, the study of host range evolution is fundamental to understanding the process of emergence, and as such, has received much research attention in the recent past (Bedhomme *et al.,* 2012; 2015; Elena, 2017; Fraile & García-Arenal 2010; McLeish *et al.,* 2019; Seabloom *et al.,* 2015). However, until recently, the study of host range evolution has focused on the evolutionary factors, with little knowledge of the effect of ecological factors on host range evolution (Rossinck & García-Arenal, 2015).

Previous work has highlighted the importance of knowing the host range of viruses to understand how host and virus ecology influences the epidemiology of the virus, and thus the risk of infection (Borer *et al.,* 2010; Malmstrom *et al.,* 2005; McLeish *et al.,* 2017; Pagán *et al.,* 2012). Nevertheless, it is unusual to find literature from studies that have characterised the host range of a virus in a given ecosystem (McLeish *et al.,* 2017). Understanding how the emergence process operates at the landscape scale presents practical and analytical challenges, such as handling multivariate datasets with dissimilar probability distributions (Warton et al., 2015), taxonomic anomalies (Blüthgen *et al.,* 2008; Novotny *et al.,* 2002), and spatial or temporal scale dependencies (Nash *et al.,* 2014; O'Dwyer & Cornell, 2018; Peterson *et al.,*1998). Nevertheless, new approaches, such as High Throughput Sequencing (HTS) and new tools for big-data management have opened the possibility to obtain rapidly data on multi-host multi-virus communities at the landscape scale. The approach promises to facilitate a better understanding of the influence of plant-virus interactions on the risk of infection and provide information on the processes leading to virus emergence.

Host range is related to the specificity of host-pathogen interactions. Ecological factors that affect the occurrence of plant viruses in different host species and host communities may modulate the specificity of plant-virus

interactions in the ecosystems (Jones 2009). The interactions among species in a community, or in a given environment, determine what organisms are present in it, that is, the community structure (Lavorel & Garnier, 2002; Van der Heijden *et al*., 1998). Morphological, physiological, and behavioural traits that suit a species to a particular habitat or resource and constrain its capacity to utilise other environments (i.e., ecological specialisation), is an important cause of biological diversity of communities (Via & Hawthorne, 2002). Resource use, particularly the host species used by a virus, is typically generalised in terms of the number of hosts used to fulfil their reproductive needs (McLeish *et al.,* 2018), and the specificity in the use of hosts defined by host range. Alternatively, the range or variety of environmental conditions a species can exist or grow in may be captured in what is termed an ecological niche (Sexton *et al.,* 2017). Niche theory establishes the relationship between a species' resource exploitation strategy and the breadth of resources available to it (Hutchinson, 1957). A major development in niche theory was the concept of the realised niche, which refers to the subset of available resources that a species is able to use, and the fundamental niche (Futuyma & Moreno, 1988; Soberón & Peterson, 2005). The fundamental niche stresses the potential of a species to utilise an environment, and for viruses this typically includes susceptible host tissue and host genotype (Al-Naimi *et al.,* 2005). The realised niche of a virus may also include ecological interactions, such as predation, mutualism, herbivory, and parasitism, and the hosts that are infected in a particular environment, the realised host range. Additionally, other factors such as co-infection, that is, the coexistence of different parasites or strains in the same host, influence the niche of the virus (May & Nowak, 1994), and contributes to host resource use and the realised host range.

Identifying the ecological determinants that affect plant-virus interaction and transmission dynamics at the landscape scale is a goal that far exceeds the scope of a doctoral thesis. Nevertheless, this thesis aims at fulfilling a series of objectives that will set the bases for reaching that ambitious goal.

5.1. Characterisation of sampled habitats in an agricultural ecosystem in central Spain

As a first step in the objectives of this thesis, field work was performed between 2015 and 2017 in four key habitats (Oakwood, Wasteland, Edge, and Crop, in increasing level of human intervention) of an agricultural landscape in central Spain. Land use practices cause changes in the distribution and abundance of host species. These practices are common and diverse in the agricultural region of the Tajo basin known as Vega del Tajo-Tajuña, making it a suitable area to study the ecological and genetic factors in virus emergence. In this ecosystem, the *a priori* habitat categories differ in the types of vegetation cover and plant species diversity. The variation in the configuration and area of cover types associated with plant communities of each habitat required that the sampling effort would be different for each of them; greater in wild habitats than in anthropic habitats. Sampling biodiversity is a necessary step in quantifying and comparing communities, but prone to sample bias. A sufficient sampling effort is necessary to obtain an accurate representation of species richness and their relative abundances in the habitats sampled (Cardoso *et al.,* 2004; Chao *et al.,* 2009; Nielsen *et al.,* 2007). To achieve a sampling effort appropriate to capture the variability between habitats, replication (multiple sites of the same habitat) and resampling (multiple collections at each site) strategies were implemented. Hence, twenty-three different sites, 4 of Oakwood, 4 of Wasteland, 4 of Edge, and 11 of Crop, were chosen so that site and habitat were not spatially correlated, and a total of 78 collections of plant samples were made at these sites following procedures to minimise bias in samples and estimates of plant community diversity. The rarefaction analysis of plant collections and the estimation of habitat species richness by Chao 1 estimator **(Figure 4.1** and **Table 4.1)** indicated that the sample size of the collections was sufficient and supported the categorisation of the *a priori* habitats.

We used an ordination approach to evaluate the relative homogeneity of the collections among each of the habitats, and the validity of our *a priori* habitat categorisation. The DCA showed homogeneity of collections from each of the habitats and supported our *a priori* categorisation of these vegetation cover types **(Figure 4.2)**. Furthermore, the DCA indicated that Oakwood was the most

ecologically differentiated community, Crop and Edge communities were less differentiated, and Wasteland was also a differentiated community closer to Edge and Crop than to Oakwood. Thus, these analyses supported again our *a priori* categorisation of the habitats.

It is often assumed that agricultural landscapes have lower diversity than 'wild' communities (Alexander *et al.,* 2014) due to ecosystem simplification (Roossinck & García-Arenal, 2015). The hypothesis that ecosystem simplification and loss of biodiversity in agricultural ecosystems may favour new encounters in the epidemiological dynamics of plant viruses (Fraile & García-Arenal, 2016; Islam *et al.,* 2020; Jones *et al.,* 2013; Roossinck & García-Arenal, 2015; Stukenbrock & McDonald, 2008), highlights the importance of plant structure-function relationships in virus diversity. Estimation of community parameters, such as host species diversity, is often the basis for many epidemiological studies on disease dynamics. However, the use of specific diversity estimators, and the aggregation of data for analyses, may have an effect on the interpretation of the results (McLeish *et al.,* 2017; 2021). Both of diversity estimators, the Tsallis entropy (S_q) and the extrapolated Shannon diversity estimator (D_{AE}), indicated similar distinctions among the habitats at the collection level **(Table 4.2** and **Figure 4.5)**. However, after aggregating relative abundance counts by habitat, the distinction between Edge and Oakwood using S_q changed, and inflated the D_{AE} estimate of Edge in particular. Therefore, the distinctions among the habitats were sensitive to the aggregation of abundance counts and the choice of diversity estimator. The effect of aggregation possibly reflects the property of 'doubling' (Jost, 2007), which results in an increase in the diversity estimate due to dissimilarities in species compositions of the communities being grouped. The effect of data aggregation on explanatory models has been demonstrated in the prediction of the spread of *West Nile virus* (WNV), where a consistent pattern of model error was related to the spatial scales of the analyses (Young *et al.,* 2013). The estimates showed that the species diversity of each habitat was different, and although the aggregation of data overestimated the diversity of Edge, Crop was always the habitat with the lowest diversity, and Wasteland the most diverse.

All existing methods for quantifying species richness are underestimates because they are affected by the right-tail bias caused by the relatively high

proportion of rare species typical of any community (Chao *et al.,* 2014). Also, species richness does not differentiate between rare or frequent species (differences in species evenness), nor consider differences in community composition (Ostfeld & LoGiudice, 2003), while the DCA does. To demonstrate structure-function relationships, landscape-scale studies typically characterise compositional differences among communities in terms of species richness and abundance that may not correspond to biological processes (Yuen & Mila, 2015; Viana & Chase, 2019). Furthermore, compositional variation of species traits among communities, rather than species diversity *per se*, is also expected to influence biological processes (Tilman *et al*., 1997), such as disease risk. Here, diversity estimates were used to generalise about differences between habitats of the ecosystem. The DCA provided information that described ecological distances among communities of the habitats, not in a spatially explicit context, but based on species relative abundances.

5.2. Assessment of biases in viral detection

The characterisation of the virome by HTS of RNA extracts of the wide diversity of plant species sampled in the study area, may include potential sources of bias that invalidate conclusions.

In field-based studies, biases may arise from the sampling strategy when the sample does not represent the relative richness and abundance of species. This bias was minimised using a systematic sampling strategy conducted during each collection at each site. In addition, the inclusion of rare species in the sample may lead to a bias in the study of plant-virus interactions due to the large variances associated with small samples. For example, if a plant species is represented by a single observation (singletons), it is unavoidably assigned a single link in an interaction network, and according to its infection status, it will bias the prevalence estimate. This bias makes the elimination of rare species a common practice in studies of ecological interactions (Blüthgen *et al.,* 2008; Blüthgen, 2010; Vázquez *et al.,* 2005), which is the criterion followed here. The omission of rare species implies an artificial reduction of the sample size **(Table 4.3)** that may blur distinctions between plant communities in the different habitats. This effect was evaluated **(Figure 4.7.B)** and shown to be minor. In addition, the

omission of rare species may affect the study by removing host plant species shared by more than one habitat that are potentially "key" elements in plant virus infection networks. Again, this effect was shown to be minor, as the species richness shared between each habitat was highly correlated before and after rare species omission.

Another potential source of bias is the method followed for preparing the HTS libraries. Among the various recent approaches to study the virome (Nerva *et al.,* 2019), we chose to sequence total RNA because it does not bias according to virus taxonomy (as dsRNA and VANA do) but may bias against viruses with low titre or with low prevalence. Since our aim was not virus discovery, we chose the total RNA approach to maximise the ability to detect viruses of all genome types, as a compromise for the potential to miss low titre and low prevalence viruses. Among the many reported protocols for plant RNA extraction (Yockteng *et al.,* 2013) we focused on kit-based methods that would allow rapid processing of a large number of samples. Prior to mass extraction over the entire sample, we analysed factors that would optimise the procedure.

The total RNA approach requires the preparation of extracts of optimal quality and concentration for HTS (Yockteng *et al.,* 2013), which is partly dependent on the plant species and their phenology. For instance, plant-associated factors that may result in poor RNA extracts include the presence of large amounts of secondary metabolites (mostly phenolics and oxidants) in the plant tissues, low concentrations of nucleic acids, and/or large amounts of carbohydrate and other polymers such as lignin (wood) that are difficult to disrupt and remove (Hilario & Mackay, 2007). The large diversity of plant species in our collections and the seasonal variation in the life stage of specimens collected throughout the year, make obtaining a sufficient quantity of RNA with a quality homogeneous across samples, a difficult task. The results of previous assays indicated that RNA quality will inevitably be heterogeneous between samples due to both taxonomic (species concerned) and biological (phenological state of the sample) factors **(Table 4.6).** The quality variation present in RNA extracts could negatively affect library preparation and HTS performance, underestimating the detection of viral sequences in HTS (Bester *et al.,* 2021). However, the potential bias associated with the heterogeneity of RNA extract quality was estimated to

be acceptable, on the basis of the detection of mRNAs of conserved plant genes, studied as internal controls **(Table 4.7)**, which was supported by the low read number of the mRNA of *Topoisomerase II associated protein* gene detected in a few libraries (about 6%).

Last, the pipeline used to detect viral sequences in the HTS libraries may introduce a further source of bias. A first step in most pipelines is a BLAST query analysis that relies on reference-based alignment. However, a significant proportion of virus sequences from the HTS libraries are expected to be divergent from the virus sequence references, which makes it difficult to assign taxonomy accurately. The bias in plant virus species available from databases, is because a significant proportion of them have been isolated from domesticated plant species (Wren *et al.,* 2006). The lack of references available for species not associated with agriculture implies that many deeply divergent viruses, or those that lack common ancestry with known virus families, remain undiscovered (Bejerman *et al.,* 2020). Therefore, a significant proportion of read sequences produced from metagenomic samples do not have references in the currently available sequence databases (Lefebvre *et al.,* 2019; Roossinck, 2012). To retain homologous sequences that are variants of our genomic references, and maintain a sufficiently high stringency to discriminate matches with query sequences from other genes (Sandhya *et al.,* 2003), the BLAST query was implemented with a moderate expected value (*E.value*) threshold, a measure of reliability of the sequence similarity between the query and reference.

However, given that homologous regions among related virus taxa can produce false positives, there is uncertainty in the accuracy of detection. Thus, a series of *in silico* validation steps applied to the BLAST query matches were conducted to reduce the frequency of false positive detections. The change in sequence similarity between BLAST queries and the references was assessed by comparing bit-score (*S'*) density distributions before and after the application of the *in silico* validation criteria. The *S'* mean/median improved after applying the validation steps. The improvement in the quality of each match is shown by the positive shift in *S'* distributions and was partly an effect of removing reads that had sequence similarity with the reference lower than 90% **(Figure 4.10)**. The richness of virus OTUs that were detected decreased (from 150 to 90 virus OTUs)

in each habitat (Crop from 46 to 31; Edge from 106 to 74; Wasteland from 85 to 43; Oakwood from 87 to 34). The largest changes to OTU richness were in Wasteland and Oakwood habitats. This may be explained by the high frequency of “unknown” virus OTUs in the wilder habitats, which had low sequence homology with the currently available references used to construct the local BLAST database (Lefebvre *et al.,* 2019; Roossinck, 2012).

5.3. Characterisation of the virome

The socioeconomic impact of diseases (Aranda & Freitas-Astúa 2017; Nicaise 2014) has biased knowledge of plant viruses towards those pathogenic to crops, at the expense of interactions with wild plants (Malmstrom & Alexander, 2016; McLeish *et al.,* 2020b; Roossinck & García-Arenal, 2015). HTS provides a means of identifying viruses that infect an individual host or a community of hosts with no bias towards those that cause diseases (Claverie *et al.*, 2018; Ma *et al.*, 2019; Massart *et al.*, 2017; Roossinck *et al.*, 2010; Wu *et al.*, 2015). HTS has been readily applied to the identification of viromes and analyses of infection dynamics, ecology and evolution of viruses associated with wild plants, allowing a detailed description of associations in plant communities of a few non-agricultural ecosystems (Bernardo *et al.,* 2018; Melcher & Grover, 2011; Muthukumar *et al.,* 2009; Roossinck *et al.,* 2010). HTS analyses of crop viromes still lag behind those of wild ecosystems (Jones & Naidu, 2019; Roossinck & García-Arenal, 2015), and have focussed on virus discovery and diagnosis (Candresse *et al.,* 2014), the temporal or spatial variation of virus communities of crops (de Souza *et al.,* 2020; Jones, 2014; Xu *et al.,* 2017), the comparison of virus communities of crops and wild relatives (Vélez-Olmedo *et al.,* 2021), or the comparison of viromes of crops and weeds as potential sources of inoculum for crops (Alcalá-Briseño *et al.,* 2019; Ma *et al.,* 2020). To our knowledge, comparative HTS studies of virus communities in wild plant communities and in crop communities are still lacking. In this study we used an HTS-based metagenomic approach to characterise the community of ssRNA viruses that infected plants in wild communities of four key habitats of an agricultural ecosystem in central Spain. At odds with previous studies that focused only on wild plant communities, our study also included crop fields. The four habitats that

were studied represent different degrees of human intervention, which should allow the analysis of the effects of habitat disturbance on the virome. The aim of the study was not virus discovery, but to provide information for understanding multi-host – multi-pathogen interactions in complex communities, an underexplored topic (Johnson *et al.,* 2013; Woolhouse & Gowtage-Sequeria, 2005).

Homologous genomic regions among related taxa can produce false positive results during sequence detection, by including viruses that would otherwise be excluded, when unique regions that differentiate taxa are absent in the read sequence data. The specificity of sequence detection by HTS technology was assessed by a series of *silico* validation steps.

A total of 90 fully validated virus OTUs were detected and represented 22 genera in 8 viral families **(Table 4.8)**. Additionally, two fungal viruses (*Botrytis ourmia-like virus, Botrytis virus F*) with virus-like ssRNA reads were also detected. This can be explained by the presence of endophytic or pathogenic fungi associated with the plants that were sampled. Positive-sense single-stranded RNA ((+)ssRNA) OTUs accounted for 95.6% of the total (86/90), with 4.4% (4/90) comprised of (-)ssRNA OTUs. Therefore, our procedures were able to recover OTUs from positive and negative sense single-stranded RNA genomes in proportions that corresponded closely to what is expected of the frequency of species within these genome classes (King *et al.,* 2011; Koonin *et al.,* 2020), suggesting no gross bias in detection.

Differences between the community composition of each habitat result in a subset of possible biotic interactions at any given time (such as predation, mutualism, herbivory and parasitism, and the hosts that are infected), which are often defined as the 'realised' component of the ecological niche (Wells & Clark, 2019). The constraints posed by each habitat on the host species available to a virus (or vector), is observed as the realised host range, and contributes to structuring plant-virus and virus-virus interactions at higher levels of organisation such as at the ecosystem scale. For example, virus distributions at both local and global scales reveal strong habitat-specificity (Bernardo *et al.,* 2018; Páez-Espino *et al.,* 2016). Hence, the ecological variation among host species that occupy plant communities of different habitats is expected to affect the availability of

hosts, and influence the means by which viruses move among habitats (Borer *et al.,* 2010). The virome varied largely between the plant communities of each habitat **(Table 4.8)**. Edge was the habitat that supported the highest number of virus OTUs detected, followed by Wasteland, Oakwood and Crop, in decreasing numbers of OTUs. Crop was the habitat with the lowest richness of virus OTUs, and supports the hypothesis that agriculture has an impact on the diversity of plant viruses (Bernardo *et al*., 2018). However, the richness of virus OTUs was lower in Oakwood and Wasteland than in Edge, and suggests that habitats with assemblies of largely native vegetation support a lower diversity of viruses. This may be explained by the presence of "unknown" virus OTUs that are more frequent in wild habitats (Bernardo *et al*., 2018), which had lower sequence similarity with the references of the local BLAST database (Lefebvre *et al.,* 2019; Roossinck, 2012) and were potentially omitted by the *in silico* validation steps. This speculation is supported by the relatively large reduction in OTU richness in Oakwood and Wasteland compared to Edge or Crop.

Virus OTUs with (+)ssRNA and (-)ssRNA genomes were detected in 106 of the 118 plant species analysed **(Table 4.9)**, in agreement with the frequent association of symbionts with a majority of hosts in other systems (Fleming-Davies *et al.,* 2015; Roossinck 2011). The plant species for which read libraries were produced, and where no virus OTUs were detected were: *Astragalus sesameus* L.*, Borago officinalis* L.*, Bupleurum rigidum* L.*, Foeniculum vulgare* Mill.*, Helianthemum cinereum* (Cav.) Pers., *Hieracium pilosella* L., *Lavandula latifolia* Medik., *Lithodora fruticosa* (L.) Griseb., *Origanum vulgare* L.*, Phlomis purpurea* L., *Retama sphaerocarpa* L. and *Rhamnus lycioides* L. The 12 plant species for which no OTU viruses were detected are from Wasteland and Oakwood, which again may indicate the removal of divergent virus taxa by the validation steps.

The number of virus OTUs varied largely across host plants and were detected in high numbers from species such as *Convolvulus arvensis* L., to being absent in the 12 species listed above. Among the plant species with higher OTU richness were *C. arvensis, Picris echioides* L.*, Carduus bourgeanus* L.*, Rubia peregrina* L., *Diplotaxis erucoides* DC.*, Lactuca serriola* L., *Papaver rhoeas* L., *Amaranthus sp.* L., *Conyza canadensis* (L.) Cronquist and *Chenopodium album*

L. **(Figure 4.13.A)**. These host species are from the families *Asteraceae*, *Brassicaceae*, *Rubiaceae*, *Convolvulaceae*, *Amaranthaceae* and *Papaveraceae*. These results provide evidence of viruses (or a vector) preferences by a determined hosts species.

The variation in the composition of each host community may correspond to variation in resources available to vectors and viruses, and modulate ecological interactions, and affect the host range of viruses. The differences in virus host range between habitats were significant, being the host ranges in Edge larger than in the other habitats **(Figure 4.12.B)**. This indicates that ecological variation among host species occupying plant communities in different habitats influences the means by which vectors transmit viruses within and between habitats (Borer *et al.,* 2010).

Additionally, the fourth-corner analysis shows that 11 virus OTUs with some of the broadest host ranges, had a positive association with the relative abundance of a particular host species in Edge only. Hence, Edge is conducive to viruses with wide and narrow host ranges **(Figure 4.14)** and suggests that the relative stability of these plant assemblies compared to crop, and their adjacency to expansive monocultures, are ecological features that support a high richness of viruses. Our results reveal the importance of Edge as a reservoir community (Ashford, 2003), which might intensify ecological interactions and co-prevalence in close proximity to crops (McLeish *et al.,* 2019).

5.4. Diversification of WMV at the landscape scale

The evolution and diversification of ssRNA plant viruses are often examined under reductionist conditions that ignore potentially much wider biotic interactions. Plant virus evolutionary studies of heterogeneous communities are best served by high-throughput approaches that efficiently detect host range of viral species at ecosystem levels. To analyse how ecological factors may contribute to structuring plant-virus genetic diversity (McLeish *et al.,* 2017 & 2019), we focused on WMV. A pilot study showed WMV to be widespread across the habitats and possessed sufficient variation to warrant a population study suited to the objectives. WMV is a generalist virus, and so far, reported almost exclusively from melon crops (Juárez *et al.,* 2013; Moreno *et al.,* 2004), and

infrequently from non-cucurbit hosts (Malpica *et al.,* 2006). Much of the previous work on the genetic diversity of WMV has focused on infections of cucurbit crop species (Ali & Natsuaki, 2006; Alonso-Prados *et al.,* 2003; Bertin *et al.,* 2020; Desbiez *et al.,* 2019), which ignores potential variation in diversity that may be present non-crop host species. Species-rich agricultural ecosystems of central Spain (Médail & Diadema, 2009) are occupied by hundreds of non-cucurbit species that represent a plant community of potential hosts for viruses. We combined HTS and the more traditional approach of RT-PCR amplification and Sanger sequencing of genomic regions, to analyse the structuring of genetic diversity of WMV in the agricultural ecosystem, and to test if its diversification is modulated by differences among the habitats.

The host range of WMV was determined using the HTS approach, as well as detection by RT-PCR. The RT-PCR assay showed a host range much narrower (24 host species) than that by HTS (43 host species). The realised host range of WMV determined by RT-PCR was narrower than by HTS in all habitats except Crop **(Table 4.11)**. This discrepancy in the sensitivity of detection may be produced by false positive matches of reads with related viruses, non-detection of low-titre infections, genetic variants not detected with our primers, the large variation in the quality of the RNA preparations, and its effect on long amplicons (Waggott, 1998).

Each habitat supported different assemblages of WMV host species **(Table 4.11)**, with 5 of the 24-host species detected by RT-PCR occurring in three habitats, 8 in two habitats, and 11 being specific to a single habitat **(Table 4.11)**. Additionally, data showed a differentiation in WMV transmission between wild and anthropic habitats, being higher in anthropic habitats **(Table 4.10)**. The incidence of WMV outside Crop was contingent on preferences for host species associated with each habitat. In Edge, Wasteland, and Oakwood, incidence was not conditional on host species relative abundance. A growing number of studies have demonstrated the importance of host species relative abundance on virus incidence (Allan *et al.,* 2009; Dizney & Ruedas, 2009; Keesing *et al.,* 2006, LoGiudice *et al.,* 2008,). Notwithstanding, our results show that there is not a linear relationship between WMV incidence and host species relative abundance.

Distinct plant communities **(Figure 4.2)** that differ in their compositions **(Figure 4.4)** underlie virus responses including their host range. Each community is expected to have unique functional characteristics in respect to other plant communities, each of which may elicit a particular virus response within the ecosystem (McLeish *et al.,* 2019). Theoretically, a generalist virus would optimise its fitness in all its hosts, and in each host its fitness will be lower than that of a specialist virus, but equivalent in a larger number of hosts (Bedhomme *et al.,* 2015; Elena, 2017). Equivalence in fitness across host species has been supported by the observation that plant viruses exhibit specialist-like strategies that resemble facultative generalism among wild host populations (Malpica *et al.,* 2006; McLeish *et al.,* 2017). Under this form of resource use, the host range of WMV may be wide, but can be very narrow on any particular species of a habitat without a relatively high fitness cost. For instance, WMV can use a plant species as a maintenance host (*S. dubia*), or as a principal host, (*C. melo*) **(Table 4.11)**, which is dependent on the habitat and preferences for hosts within it. The evolution of facultative generalist viruses depends on associations within subsets of available hosts that offer a range of fitness options. Constraints on the spatial and temporal variation of maximal fitness options among hosts will be observed as habitat- and/or host-specificity. Our results support the hypothesis that resource heterogeneity can boost coexistence and species richness by offering a bigger variety of niches, thereby restricting interspecific competition (habitat-heterogeneity hypothesis, Hutchinson, 1959; MacArthur & MacArthur, 1961; Rosenzweig, 1995). For example, change in both habitat complexity and the diversity of prey resources have been strongly linked to consumer diversity (Tews *et al.,* 2004). Hence, the resources heterogeneity may influence the virus transmission **(Figure 4.14)**.

The nucleotide sequences of the WMV-CP cistron obtained by Sanger sequencing of the RT-PCR amplicons were used to conduct genetic analyses to assess the effect of habitat and host plant species on the diversification of WMV populations. Selection on the CP was analysed by the estimation of ω, which indicated purifying selection ($d_N / d_S < 1$) **(Table 4.16)**, as inferred by other studies (Desbiez *et al.,* 2008; Moreno *et al.,* 2004). In line with other studies, no recombination breakpoints were detected within the CP cistron of the WMV

genome (Desbiez *et al.,* 2009; Moreno *et al.,* 2004). The Bayesian consensus phylogeny showed strong backbone support for most deep stem clades with a posterior probability (*PP*) > 0.98. Nearly all poorly resolved relationships (*PP* < 0.80) were inferred within the major clades and may be a consequence of saturation of substitutions (too much homoplasy) (Carbone *et al.,* 2004) or being similar variants. The poorly resolved relationships of the consensus tree, caused by phylogenetic ambiguity in evolutionary associations, were consistent with the reticulated relationships shown in the haplotype network **(Figure 4.16)**. The haplotype network **(Figure 4.16)** yielded four major clusters of haplotypes, consistent with the four major clades of the phylogeny **(Figure 4.15)**. Clusters A, B and C included haplotypes with mixed host/habitat associations. By contrast, cluster D included only haplotypes from Oakwood, but with mixed host associations. Constraints on virus diversification that correspond to habitat, in particular Oakwood, are also evident in the phylogeny **(Figure 4.15)**. These results suggest that habitat may have an important effect on the diversification and evolution of WMV. This conclusion is supported by the distribution of incidence previously mentioned **(Table 4.13)**, which tended to be associated with between-habitat differences. Additionally, the *BaTS* analysis **(Table 4.21)** showed that habitat effects on haplotypes were stronger than host species effects, although both factors structure WMV diversity. The AMOVA results **(Annex, Table A7)** indicated that both habitat and host had an important role in differentiation between and within WMV populations. Processes such as vector colonisation (Claflin *et al.,* 2017; Tamburini *et al.,*2016), virus traits (Kneitel & Chase, 2004), plant species traits (Han *et al.,* 2015; Lloyd-Smith *et al.,* 2005), and community assembly (Ostfeld & LoGiudice, 2003) are processes that may explain the stronger effect of the habitat in the structure of WMV diversity. The findings suggest these processes are not mutually exclusive and underlie the evolution of WMV.

6. CONCLUSIONS

In this thesis, we assessed the contribution of the ecological factors that structure ssRNA virus communities and genetic diversity, by characterising the viromes of different plant communities in four key habitats of an agricultural landscape in central Spain, as a first step to understand plant virus emergence. From the results it can be concluded that:

1. The Oakwood plant community is the most ecologically differentiated, followed by the Wasteland community, which in turn is closest to Crop and Edge communities.

2. Species diversity was significantly different between each habitat and although the aggregation of data overestimated the diversity of Edge, Crop was always the habitat with the lowest diversity, and Wasteland the most diverse. The low diversity of crop agrees with hypotheses of ecosystem simplification associated with agriculture.

3. Despite the variation in the quality of RNA extracts, which comes as an unavoidable result of environmental samples, viral detection from HTS libraries is robust to quantifying variation at higher levels of organisation between habitats.

4. The stringency of the BLAST query and the application of the validation criteria reduced the frequency of false-positive OTUs, possibly at the expense of missing true-positives.

5. Viral sequences were detected in 106 of the 118 plant species sent for HTS, indicating a frequent association of symbionts with most hosts in various ecosystems.

6. The virome characterised in the present work varied greatly between the plant communities sampled in each habitat, being Edge the habitat that supported the highest viral OTU richness. The ecological features of Edge were conducive to infections by many viruses with wide and narrow host ranges. Hence, it reveals the importance of Edge as a reservoir community.

7. Virus detection by RT-PCR showed much narrower host ranges than by HTS, which indicates that the HTS approach was less sensitive to plant sample heterogeneity than the RT-PCR approach.

8. Ecosystem simplification structures virus populations and is associated with the high incidences of WMV.

9. Heterogeneous environments affect host use by viruses (vectors), which resembled host specialisation within a given habitat.

10. WMV transmission and diversification are more strongly influenced by habitat effects than by host effects. However, both effects were not mutually exclusive. Thus, the WMV pilot study shows that both genetic and ecological factors affect transmission between and within plant communities of agricultural ecosystems, and both influence the processes of viral emergence and host range evolution.

7. BIBLIOGRAPHY

Adams, I., & Fox, A. (2016). Diagnosis of plant viruses using next-generation sequencing and metagenomic analysis. *In* A. Wang, & X. Zhou (Eds.), *Current research topics in plant virology* (pp. 323–335). Cham, Switzerland: Springer.

Agrawal, A., & Lively, C. M. (2002). Infection genetics: gene-for-gene versus matching-alleles models and all points in between. *Evolutionary Ecology Research*, 4(1), 91-107.

Agrawal, A. F., & Lively, C. M. (2003). Modelling infection as a two-step process combining gene-for-gene and matching-allele genetics. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 270(1512), 323-334.

Agrios, G. N. (2005). **Plant pathology.** (5th eds.) Ed. Elsevier. Amsterdam.

Agudelo-Romero, P., de la Iglesia, F., & Elena, S. F. (2008). The pleiotropic cost of host-specialization in Tobacco etch potyvirus. Infection, *Genetics and Evolution*, 8(6), 806-814.

Aguiar, E.R., Olmo, R.P., Paro, S., Ferreira, F.V., de Faria, I.J., Todjro, Y.M., Lobo, F.P., Kroon, E.G., Meignin, C., Gatherer, D., Imler, J.L., Marques, J.T. (2015). Sequence-independent characterization of viruses based on the pattern of viral small RNAs produced by the host. *Nucleic acids research*, 43(13), 6191-6206.

Ainsworth, G. C. (1935). Mosaic diseases of the cucumber. *Annals of Applied Biology*, 22(1), 55-67.

Alam, M. N. U., & Chowdhury, U. F. (2020). Short k-mer abundance profiles yield robust machine learning features and accurate classifiers for RNA viruses. *PloS ONE*, 15(9), e0239381.

Alcalá-Briseño, R. I., Lotrakul, P., & Valverde, R. A. (2019). Genome sequence and phylogenetic analysis of a novel comovirus from tabasco pepper (Capsicum frutescens). *Virus genes*, 55(6), 854-858.

Alexander H.M., Mauck K.E., Whitfield A.E., Garrett K.A., Malmstrom C.M. (2014). Plant-virus interactions and the agro-ecological interface. *European Journal of Plant Pathology*, 138(3), 529-547.

Alexander, A. (2017). genetic_diversity_diffs v1. 0.3. R package version, 1.0.3. https://github.com/laninsky/genetic_diversity_diffs. Accessed 09 July 2019.

Ali, A., Natsuaki, T., & Okuda, S. (2006). The complete nucleotide sequence of a Pakistani isolate of *Watermelon mosaic virus* provides further insights into the taxonomic status in the *Bean common mosaic virus* subgroup. *Virus Genes*, 32(3), 307-311.

Allan, B. F., Langerhans, R. B., Ryberg, W. A., Landesman, W. J., Griffin, N. W., Katz, R. S., Oberle, B. J., Schutzenhofer, M. R., Smyth, K. N., St. Maurice, A. d., Clark, L., Crooks, K. R., Hernandez, D. E., McLean, R. G., Ostfeld, R. S., & Chase, J. M. (2009). Ecological correlates of risk and incidence of West Nile virus in the United States. *Oecologia*, 158(4), 699-708.

Al-Naimi, F. A., Garrett, K. A., & Bockus, W. W. (2005). Competition, facilitation, and niche differentiation in two foliar pathogens. *Oecologia*, 143(3), 449-457.

Alonso-Prados, J. L., Luis-Arteaga, M., Alvarez, J. M., Moriones, E., Batlle, A., Laviña, A., García-Arenal, F., & Fraile, A. (2003). Epidemics of aphid-transmitted viruses in melon crops in Spain. *European Journal of Plant Pathology*, 109(2), 129-138.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403-410.

Anderson, P. K., Cunningham, A. A., Patel, N. G., Morales, F. J., Epstein, P. R., & Daszak, P. (2004). Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. *Trends in ecology & evolution*, 19(10), 535-544.

Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P., & Tyson, G. W. (2012). Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic acids research*, 40(12), e94-e94.

Aranda, M. A., & Freitas-Astúa, J. (2017). Ecology and diversity of plant viruses, and epidemiology of plant virus-induced diseases. *Annals of Applied Biology, 171 (1), pp. 1-4.*

Ashford, R.W. (2003) When is a reservoir not a reservoir? *Emerging Infectious Diseases* 9, 1495–1496

Atreya, C. D., & Pirone, T. P. (1993). Mutational analysis of the helper component-proteinase gene of a potyvirus: effects of amino acid substitutions, deletions, and gene replacement on virulence and aphid transmissibility. *Proceedings of the National Academy of Sciences*, 90(24), 11919-11923.

Ayme, V., Souche, S., Caranta, C., Jacquemond, M., Chadœuf, J., Palloix, A., & Moury, B. (2006). Different mutations in the genome-linked protein VPg of Potato virus Y confer virulence on the pvr23 resistance in pepper. *Molecular plant-microbe interactions*, 19(5), 557-563.

Baessler, C., & Klotz, S. (2006). Effects of changes in agricultural land-use on landscape structure and arable weed vegetation over the last 50 years. *Agriculture, ecosystems & environment*, 115(1-4), 43-50.

Barba, M., Czosnek, H., & Hadidi, A. (2014). Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses*, 6(1), 106-136.

Bates, D., Sarkar, D., Bates, M. D., & Matrix, L. (2007). The lme4 package. R package version, 2(1), 74. http://CRAN.R-project.org/package=lme4. Accessed 08 May 2018.

Bedhomme, S., Lafforgue, G., & Elena, S. F. (2012). Multihost experimental evolution of a plant RNA virus reveals local adaptation and host-specific mutations. *Molecular Biology and Evolution,* 29(5), 1481-1492.

Bedhomme, S., Hillung, J., & Elena, S. F. (2015). Emerging viruses: why they are not jacks of all trades?. *Current opinion in virology,* 10, 1-6.

Bejerman, N., Debat, H., & Dietzgen, R. G. (2020). The Plant Negative-Sense RNA Virosphere: Virus Discovery Through New Eyes. *Frontiers in Microbiology,* 11.

Belgorodski, N., Greiner, M., Tolksdorf, K., Schueller, K., Flor, M., Göhring, L., & Greiner, M. M. (2017). rriskDistributions: Fitting distributions to given data or known quantiles. R package version, 2.1.2. https://CRAN.R-project.org/package=rriskDistributions. Accessed 01 Jun 2018.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289-300.

Bera, S., Fraile, A., & García-Arenal, F. (2018). Analysis of fitness trade-offs in the host range expansion of an RNA virus, tobacco mild green mosaic virus. *Journal of virology,* 92(24).

Bernardo, P., Charles-Dominique, T., Barakat, M., Ortet, P., Fernandez, E., Filloux, D., Hartnady, P., Rebelo, T. A., Cousins, S. R., Mesleard, F., Cohez, D., Yavercovski, N., Varsani, A., Harkins, G. W., Peterschmitt, M., Malmstrom, C. M., Martin, D. P. and Roumagnac, P. (2018). Geometagenomics illuminates the impact of agriculture on the distribution and prevalence of plant viruses at the ecosystem scale. *The ISME journal,* 12(1), 173-184.

Bertin, S., Manglli, A., McLeish, M., & Tomassoli, L. (2020). Genetic variability of watermelon mosaic virus isolates infecting cucurbit crops in Italy. *Archives of virology,* 165(4), 937-946.

Bester, R., Cook, G., Breytenbach, J. H., Steyn, C., De Bruyn, R., & Maree, H. J. (2021). Towards the validation of high-throughput sequencing (HTS) for routine plant virus diagnostics: measurement of variation linked to HTS detection of citrus viruses and viroids. *Virology Journal,* 18(1), 1-19.

Blanco, E., Casado, M.A., Costa, M., Escribano, R., García, M., Génova, M., Gómez, A., Gómez, F., Moreno, J.C., Morla, C. (2005). Los Bosques Ibéricos. *Una Interpretación Geobotánica*. Madrid: Editorial Planeta.

Blüthgen, N., Fründ, J., Vázquez, D. P., & Menzel, F. (2008). What do interaction network metrics tell us about specialization and biological traits. *Ecology,* 89(12), 3387-3399.

Blüthgen, N. (2010). Why network analysis is often disconnected from community ecology: a critique and an ecologist's guide. *Basic and Applied Ecology,* 11(3), 185-195.

Bohlen, P. J., Groffman, P. M., Driscoll, C. T., Fahey, T. J., & Siccama, T. G. (2001). Plant–soil–microbial interactions in a northern hardwood forest. *Ecology,* 82(4), 965-978.

Boonham, N., Kreuze, J., Winter, S., van der Vlugt, R., Bergervoet, J., Tomlinson, J., & Mumford, R. (2014). Methods in virus diagnostics: from ELISA to next generation sequencing. *Virus research,* 186, 20-31.

Borer, E. T., Seabloom, E. W., Mitchell, C. E., & Power, A. G. (2010). Local context drives infection of grasses by vector-borne generalist viruses. *Ecology Letters,* 13(7), 810-818.

Brooks, D. R., & Boeger, W. A. (2019). Climate change and emerging infectious diseases: Evolutionary complexity in action. *Current Opinion in Systems Biology,* 13, 75-81.

Brunt, A., Crabtree, K., Dallwitz, M. J., Gibbs, A. J., & Watson, L. (1996). *Viruses of plants.* CAB International, Wallingford, UK.

Büchi, L., Christin, P. A., & Hirzel, A. H. (2009). The influence of environmental spatial structure on the life-history traits and diversity of species in a metacommunity. *Ecological Modelling*, 220(21), 2857-2864.

Burks, D. J., & Azad, R. K. (2020). Higher-order Markov models for metagenomic sequence classification. *Bioinformatics*, 36(14), 4130-4136.

Cabanillas, L., Arribas, M., & Lázaro, E. (2013). Evolution at increased error rate leads to the coexistence of multiple adaptive pathways in an RNA virus. *BMC evolutionary biology*, 13(1), 1-16.

Campbell, C. L., & Neher, D. A. (1994). Estimating disease severity and incidence. *Epidemiology and management of root diseases* (pp. 117-147). Springer, Berlin, Heidelberg.

Candresse, T., Filloux, D., Muhire, B., Julian, C., Galzi, S., Fort, G., Bernardo, P., Daugrois, J. H., Fernandez, E.., Martin, D. P., Varsani, A., & Roumagnac, P. (2014). Appearances can be deceptive: revealing a hidden viral infection with deep sequencing in a plant quarantine context. *PLoS ONE* 9:e102945.

Carbone, I., Liu, Y. C., Hillman, B. I., & Milgroom, M. G. (2004). Recombination and migration of Cryphonectria hypovirus 1 as inferred from gene genealogies and the coalescent. *Genetics*, 166(4), 1611-1629.

Cardoso, P., Silva, I., de Oliveira, N. G., & Serrano, A. R. (2004). Higher taxa surrogates of spider (Araneae) diversity and their efficiency in conservation. *Biological Conservation*, 117(4), 453-459.

Castroviejo, S. (1986-2012). Flora iberica 1-8, 10-15, 17-18, 21. Real Jardín Botánico, CSIC, Madrid.

Catalá, R., Carrasco-López, C., Perea-Resa, C., Hernández-Verdeja, T., & Salinas, J. (2019). Emerging roles of LSM complexes in posttranscriptional regulation of plant response to abiotic Stress. *Frontiers in plant science*, 10, 167.

Chang, S., Puryear, J., & Cairney, J. (1993). A simple and efficient method for isolating RNA from pine trees. *Plant molecular biology reporter,* 11(2), 113-116.

Chao, A., Colwell, R. K., Lin, C. W., & Gotelli, N. J. (2009). Sufficient sampling for asymptotic minimum species richness estimators. *Ecology,* 90(4), 1125-1133.

Chao, A., & Jost, L. (2012). Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology,* 93(12), 2533-2547.

Chao, A., Gotelli, N. J., Hsieh, T. C., Sander, E. L., Ma, K. H., Colwell, R. K., & Ellison, A. M. (2014). Rarefaction and extrapolation with Hill numbers:

a framework for sampling and estimation in species diversity studies. *Ecological monographs,* 84(1), 45-67.

Choset, H. (2000). Coverage of known spaces: The boustrophedon cellular decomposition. *Autonomous Robots,* 9(3), 247-253.

Claflin, S. B., Jones, L. E., Thaler, J. S., & Power, A. G. (2017). Crop-dominated landscapes have higher vector-borne plant virus prevalence. *Journal of Applied Ecology*, 54(4), 1190-1198.

Clark, M. F., & Adams, A. N. (1977). Characteristics of the microplate method of enzyme-linked immunosorbent assay for the detection of plant viruses. *Journal of general virology,* 34(3), 475-483.

Claverie, S., Bernardo, P., Kraberger, S., Hartnady, P., Lefeuvre, P., Lett, J. M., Galzi, S., Filloux, D., Harkins, G. W., Varsani, A., Martin, D. P., & Roumagnac P. (2018). From spatial metagenomics to molecular characterization of plant viruses: A geminivirus case study. *Advances in virus research*, *101*, 55-83.

Cooper, I., & Jones, R. A. (2006). Wild plants and viruses: under-investigated ecosystems. *Advances in virus research,* 67, 1-47.

Cowling, R. M., Rundel, P. W., Lamont, B. B., Arroyo, M. K., & Arianoutsou, M. (1996). Plant diversity in Mediterranean-climate regions. *Trends in Ecology & Evolution,* 11(9), 362-366.

Crescenzi, A., Fanigliulo, A., Comes, S., Masenga, V., Pacella, R., & Piazzolla, P. (2001). Necrosis of watermelon caused by *Watermelon mosaic virus*. *Journal of Plant Pathology,* 83(3), 227.

Crespo, O., Janssen, D., García, C., & Ruiz, L. (2017). Biological and molecular diversity of *Cucumber green mottle mosaic virus* in Spain. *Plant disease,* 101(6), 977-984.

Cronin, J. P., Welsh, M. E., Dekkers, M. G., Abercrombie, S. T., & Mitchell, C. E. (2010). Host physiological phenotype explains pathogen reservoir potential. *Ecology letters,* 13(10), 1221-1232.

Crouch, J. A., Tredway, L. P., Clarke, B. B., & Hillman, B. I. (2009). Phylogenetic and population genetic divergence correspond with habitat for the pathogen Colletotrichum cereale and allied taxa across diverse grass communities. *Molecular ecology,* 18(1), 123-135.

Da Silva, J., Coetzer, M., Nedellec, R., Pastore, C., & Mosier, D. E. (2010). Fitness epistasis and constraints on adaptation in a human immunodeficiency virus type 1 protein region. *Genetics,* 185(1), 293-303.

de Brevern, A. G., Meyniel, J. P., Fairhead, C., Neuvéglise, C., & Malpertuy, A. (2015). Trends in IT innovation to build a next generation bioinformatics solution to manage and analyse biological big data produced by NGS technologies. *BioMed research international*, 2015.

de Souza, R. S. C., Armanhi, J. S. L., & Arruda, P. (2020). From microbiome to traits: designing synthetic microbial communities for improved crop resiliency. *Frontiers in Plant Science*, 11, 1179.

Dean, J. D., Goodwin, P. H., & Hsiang, T. (2002). Comparison of relative RT-PCR and northern blot analyses to measure expression of β-1, 3-glucanase inNicotiana benthamiana infected withColltotrichum destructivum. *Plant Molecular Biology Reporter,* 20(4), 347-356.

Deng, P., Wu, Z., & Wang, A. (2015). The multifunctional protein CI of potyviruses plays interlinked and distinct roles in viral genome replication and intercellular movement. *Virology journal,* 12(1), 1-11.

Desbiez, C., & Lecoq, H. (2004). The nucleotide sequence of *Watermelon mosaic virus* (WMV, Potyvirus) reveals interspecific recombination between two related potyviruses in the 5′ part of the genome*. Archives of Virology,* 149(8), 1619-1632.

Desbiez, C., Costa, C., Wipf-Scheibel, C., Girard, M., & Lecoq, H. (2007). Serological and molecular variability of watermelon mosaic virus (genus Potyvirus). *Archives of virology,* 152(4), 775-781.

Desbiez, C., & Lecoq, H. (2008). Evidence for multiple intraspecific recombinants in natural populations of *Watermelon mosaic virus* (WMV, Potyvirus). *Archives of virology,* 153(9), 1749-1754.

Desbiez, C., Joannon, B., Wipf-Scheibel, C., Chandeysson, C., & Lecoq, H. (2009). Emergence of new strains of *Watermelon mosaic virus* in South-eastern France: Evidence for limited spread but rapid local population shift. *Virus Research,* 141(2), 201-208.

Desbiez, C., Joannon, B., Wipf-Scheibel, C., Chandeysson, C., & Lecoq, H. (2011). Recombination in natural populations of watermelon mosaic virus: new agronomic threat or damp squib?. *Journal of general virology,* 92(8), 1939-1948.

Desbiez, C., Verdin, E., Moury, B., Lecoq, H., Millot, P., Wipf-Scheibel, C., Mirzayeva, S., Sultanova, N., Balakishiyeva, G., Mammadov, A., Kheyr-Pour, A., & Huseynova, I., (2019). Prevalence and molecular diversity of the main viruses infecting cucurbit and solanaceous crops in Azerbaijan. *European Journal of Plant Pathology,* 153(2), 359-369.

Ding, S. W., & Voinnet, O. (2007). Antiviral immunity directed by small RNAs. *Cell,* 130(3), 413-426.

Dizney, L. J., & Ruedas, L. A. (2009). Increased host species diversity and decreased prevalence of Sin Nombre virus. *Emerging infectious diseases*, *15*(7), 1012.

Dolédec, S., Chessel, D., Ter Braak, C. J., & Champely, S. (1996). Matching species traits to environmental variables: a new three-table ordination method. *Environmental and Ecological Statistics,* 3(2), 143-166.

Dombrovsky, A., Huet, H., Chejanovsky, N., & Raccah, B. (2005). Aphid transmission of a potyvirus depends on suitability of the helper component and the N terminus of the coat protein. *Archives of virology,* 150(2), 287-298.

Donaire, L., Wang, Y., Gonzalez-Ibeas, D., Mayer, K. F., Aranda, M. A., & Llave, C. (2009). Deep-sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes. *Virology,* 392(2), 203-214.

Dray, S., & Dufour, A. B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of statistical software,* 22(4), 1-20. R package version 1.7-18. http://CRAN.R-project.org/package=ade4.

Dray, S., & Legendre, P. (2008). Testing the species traits–environment relationships: the fourth-corner problem revisited. *Ecology,* 89(12), 3400-3412.

Dubay, B. (2017). Developing Methodology For Elucidating Host Range Evolution In Plant RNA Viruses. Trabajo de fin de máster. Universidad Politécnica de Madrid.

Dybdahl, M. F., Jenkins, C. E., & Nuismer, S. L. (2014). Identifying the molecular basis of host-parasite coevolution: merging models and mechanisms. *The American Naturalist,* 184(1), 1-13.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput*. Nucleic acids research,* 32(5), 1792-1797.

Elena, S. F., Agudelo-Romero, P., & Lalić, J. (2009). The evolution of viruses in multi-host fitness landscapes. *The open virology journal,* 3, 1.

Elena, S. F., Fraile, A., & García-Arenal, F. (2014). Evolution and emergence of plant viruses. *Advances in Virus Research,* 88, 161-191.

Elena, S. F. (2017). Local adaptation of plant viruses: lessons from experimental evolution*. Molecular ecology,* 26(7), 1711-1719.

Ellouze, W., Mishra, V., Howard, R. J., Ling, K. S., & Zhang, W. (2020). Control of Cucumber green mottle mosaic virus in commercial greenhouse production with agricultural disinfectants and resistant cucumber varieties. *bioRxiv*.

Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research*, 8(3), 186-194.

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131(2), 479-491.

Excoffier, L., Laval, G., & Schneider, S. (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary bioinformatics,* 1, 117693430500100003.

Expósito-Rodríguez, M., Borges, A. A., Borges-Pérez, A., & Pérez, J. A. (2008). Selection of internal control genes for quantitative real-time RT-PCR studies during tomato development process. *BMC plant biology*, 8(1), 131.

Fan, H., Ives, A. R., Surget-Groba, Y., & Cannon, C. H. (2015). An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC genomics*, 16(1), 522.

FAO (2009). Global agriculture towards 2050. Food and Agriculture Organization of the United Nations: Rome-Italy.

Fitch, W. M. (2000). Homology: a personal view on some of the problems. *Trends in genetics*, 16(5), 227-231.

Fleming-Davies, A. E., Dukic, V., Andreasen, V., & Dwyer, G. (2015). Effects of host heterogeneity on pathogen diversity and evolution. *Ecology letters*, 18(11), 1252-1261.

Foster, S. D., Hosack, G. R., Monk, J., Lawrence, E., Barrett, N. S., Williams, A., & Przeslawski, R. (2020). Spatially balanced designs for transect-based surveys. *Methods in Ecology and Evolution*, 11(1), 95-105.

Fraile, A., & García-Arenal, F. (2010). The coevolution of plants and viruses: resistance and pathogenicity. *Advances in virus research*, 76, 1-32.

Fraile, A., Pagán, I., Anastasio, G., Sáez, E., & García-Arenal, F. (2011). Rapid genetic diversification and high fitness penalties associated with pathogenicity evolution in a plant virus. *Molecular Biology and Evolution*, 28(4), 1425-1437.

Fraile, A., & García-Arenal, F. (2016). Environment and evolution modulate plant virus pathogenesis. *Current opinion in virology*, 17, 50-56.

Frank, S. A. (1996). Models of parasite virulence. *The Quarterly review of biology*, 71(1), 37-78.

Fraser, C., Alm, E. J., Polz, M. F., Spratt, B. G., & Hanage, W. P. (2009). The bacterial species challenge: making sense of genetic and ecological diversity. *Science*, 323(5915), 741-746.

Fried, G., Petit, S., Dessaint, F., & Reboud, X. (2009). Arable weed decline in Northern France: crop edges as refugia for weed conservation?. *Biological Conservation*, 142(1), 238-243.

Futuyma, D. J., & Moreno, G. (1988). The evolution of ecological specialization. *Annual review of Ecology and Systematics*, 19(1), 207-233.

Gadhave, K. R., Gautam, S., Rasmussen, D. A., & Srinivasan, R. (2020). Aphid Transmission of Potyvirus: The Largest Plant-Infecting RNA Virus Genus. *Viruses*, 12(7), 773.

Gal-On, A. (2007) *Zucchini yellow mosaic virus*: insect transmission and pathogenicity—the tails of two proteins. *Molecular plant pathology*, 8, 139–150.

Gara, I. W., Kondo, H., Maeda, T., Inouye, N., & Tamada, T. (1997). Stunt disease of Habenaria radiata caused by a strain of *Watermelon mosaic virus* 2. *Japanese Journal of Phytopathology*, 63(2), 113-117.

García-Arenal, F., & McDonald, B. A. (2003). An analysis of the durability of resistance to plant viruses. *Phytopathology*, 93(8), 941-952.

García-Arenal, F., & Fraile, A. (2013). Trade-offs in host range evolution of plant viruses. *Plant Pathology*, 62, 2-9.

García-Arenal, F., & Zerbini, F. M. (2019). Life on the edge: geminiviruses at the interface between crops and wild plant hosts. *Annual review of virology*, 6, 411-433.

Gavilán, R. G., Vilches, B., Gutiérrez-Girón, A., Blanquer, J. M., & Escudero, A. (2018). Sclerophyllous versus deciduous forests in the Iberian Peninsula: A standard case of Mediterranean climatic vegetation distribution. *Geographical Changes in Vegetation and Plant Functional Types* (pp. 101-116). Springer, Cham.

Ge, H., Walhout, A. J., & Vidal, M. (2003). Integrating 'omic'information: a bridge between genomics and systems biology. *TRENDS in Genetics*, 19(10), 551-560.

Gibbs, M. J., Armstrong, J. S., & Gibbs, A. J. (2000). Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics*, 16(7), 573-582.

Godefroid, S., & Koedam, N. (2007). Urban plant species patterns are highly driven by density and function of built-up areas. *Landscape ecology*, 22(8), 1227-1239.

González, R., Butković, A., & Elena, S. F. (2020). From foes to friends: Viral infections expand the limits of host phenotypic plasticity. *Advances in virus research* (Vol. 106, pp. 85-121). Academic Press.

Goodall-Copestake, W. P., Tarling, G. A., & Murphy, E. J. (2012). On the comparison of population-level estimates of haplotype and nucleotide diversity: a case study using the gene cox1 in animals. *Heredity*, 109(1), 50-56.

Granda, E., Escudero, A., & Valladares, F. (2014). More than just drought: complexity of recruitment patterns in Mediterranean forests. *Oecologia*, 176(4), 997-1007.

Han, B. A., Schmidt, J. P., Bowden, S. E., & Drake, J. M. (2015). Rodent reservoirs of future zoonotic diseases. *Proceedings of the National Academy of Sciences*, 112(22), 7039-7044.

Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C., & Martiny, J. B. (2012). Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature Reviews Microbiology*, 10(7), 497-506.

Haydon, D. T., Cleaveland, S., Taylor, L. H., & Laurenson, M. K. (2002). Identifying reservoirs of infection: a conceptual and practical challenge. *Emerging infectious diseases*, 8(12), 1468-1473.

Hervé, M. (2020). RVAideMemoire: Testing and Plotting Procedures for Biostatistics. R package version 0.9-75. https://CRAN.R project.org/package=RVAideMemoire. Accessed 28 Jun 2019.

Holmes, E. C. (2009). The evolution and emergence of RNA viruses. Oxford University Press , Oxford, UK.

Hosseini P. R., Mills J. N., Prieur-Richard A. H., Ezenwa V. O., Bailly X., Rizzoli A., Suzán G., Vittecoq M., García-Peña G. E., Daszak P., Guégan J. F., & Roche B. (2017). Does the impact of biodiversity differ between emerging and endemic pathogens? The need to separate the concepts of hazard and risk. Philosophical Transactions of the Royal Society B: *Biological Sciences*, 372(1722), 20160129.

Hsieh, T. C., Ma, K. H., & Chao, A. (2016). iNEXT: an R package for rarefaction and extrapolation of species diversity (H ill numbers). *Methods in Ecology and Evolution*, 7(12), 1451-1456.

Hsu, H. T. (2002). Biological control of plant pathogens (viruses). *Encyclopedia of pest management* (ed. by David, P.), pp. 68. Marcel Dekker, New York.

Hudson, P. J., Perkins, S. E., & Cattadori, I. M. (2008). The emergence of wildlife disease and the application of ecology. *Infectious disease ecology: Effects of Ecosystems on Disease and of Disease on Ecosystems* (eds Ostfeld, R.S., Keesing, F. & Eviner, V.T.). Princeton University Press, Princeton,

Huelsenbeck, J. P., & Ronquist, F. (2001). mrbayes, version 3.0 b4: Bayesian influence of phylogeny. *Bioinformatics*, 17, 754-755.

Huggett, J., Dheda, K., Bustin, S., & Zumla, A. (2005). Real-time RT-PCR normalisation; strategies and considerations. *Genes & Immunity*, 6(4), 279-284.

Hurlbert, S. H. (1971). The nonconcept of species diversity: a critique and alternative parameters. *Ecology*, 52(4), 577-586.

Hutchinson, G.E., (1957). Concluding remarks*. Cold Spring Harb.* Symp. Quant. Biol. 22, 415–427.

Hyvönen, T. (2007). Can conversion to organic farming restore the species composition of arable weed communities?. *Biological Conservation*, 137(3), 382-390.

Ishibashi, K., Mawatari, N., Miyashita, S., Kishino, H., Meshi, T., & Ishikawa, M. (2012). Coevolution and hierarchical interactions of *Tomato mosaic virus* and the resistance gene Tm-1. *PLoS Pathog*, 8(10), e1002975.

Islam, W., Noman, A., Naveed, H., Alamri, S. A., Hashem, M., Huang, Z., & Chen, H. Y. (2020). Plant-insect vector-virus interactions under environmental change. *Science of The Total Environment*, 701, 135044.

Janzac, B., Montarry, J., Palloix, A., Navaud, O., & Moury, B. (2010). A point mutation in the polymerase of Potato virus Y confers virulence toward the Pvr4 resistance of pepper and a high competitiveness cost in susceptible cultivar. *Molecular plant-microbe interactions*, 23(6), 823-830.

Jenner, C. E., Wang, X., Ponz, F., & Walsh, J. A. (2002). A fitness cost for *Turnip mosaic virus* to overcome host resistance. *Virus research*, 86(1-2), 1-6.

Joannon, B., Lavigne, C., Lecoq, H., & Desbiez, C. (2010). Barriers to gene flow between emerging populations of *Watermelon mosaic virus* in southeastern France. *Phytopathology*, 100(12), 1373-1379.

Johnson, P. T., Preston, D. L., Hoverman, J. T., & LaFonte, B. E. (2013). Host and parasite diversity jointly control disease risk in complex communities. *Proceedings of the National Academy of Sciences*, 110(42), 16916-16921.

Johnson, P. T., Ostfeld, R. S., & Keesing, F. (2015). Frontiers in research on biodiversity and disease. *Ecology letters*, 18(10), 1119-1133.

Jones R. A. C. (2009). Plant virus emergence and evolution: origins, new encounter scenarios, factors driving emergence, effects of changing world conditions, and prospects for control. *Virus Res*. 141:113–30

Jones R. A. C. (2013). Virus diseases of pasture grasses in Australia: incidences, losses, epidemiology, and management. *Crop and Pasture Science*, 64(3), 216-233.

Jones R. A. C. (2014). Trends in plant virus epidemiology: opportunities from new or improved technologies. *Virus research*, 186, 3-19.

Jones R. A. C. & Naidu, R. A. (2019). Global dimensions of plant virus diseases: Current status and future perspectives. *Annual review of virology*, 6, 387-409.

Jones R. A. C. (2020). Disease Pandemics and Major Epidemics Arising from New Encounters between Indigenous Viruses and Introduced Crops. *Viruses*, 12(12), 1388.

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology*, 88(10), 2427-2439.

Juárez, M., Legua, P., Mengual, C. M., Kassem, M. A., Sempere, R. N., Gómez, P., Truniger, V., & Aranda, M. A. (2013). Relative incidence, spatial distribution and genetic diversity of cucurbit viruses in eastern Spain. *Annals of Applied Biology*, 162(3), 362-370.

Kassem, M. A., Sempere, R. N., Juárez, M., Aranda, M. A., & Truniger, V. (2007). Cucurbit aphid-borne yellows virus is prevalent in field-grown cucurbit crops of southeastern Spain. *Plant Disease*, 91(3), 232-238.

Keesing, F., Holt, R. D., & Ostfeld, R. S. (2006). Effects of species diversity on disease risk. *Ecology letters*, 9(4), 485-498.

Keesing F., Belden LK., Daszak P., Dobson A., Harvell CD., Holt RD., Hudson P., Jolles A., Jones KE., Mitchell CE., Myers SS., Bogich T., Ostfeld R.S. (2010). Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature*, 468(7324), 647-652.

King, A. M., Lefkowitz, E., Adams, M. J., & Carstens, E. B. (Eds.). (2011). *Virus taxonomy: ninth report of the International Committee on Taxonomy of Viruses*. Elsevier.

Kneitel, J. M., & Chase, J. M. (2004). Trade-offs in community ecology: linking spatial scales and species coexistence. *Ecology letters*, 7(1), 69-80.

Koonin, E. V., Dolja, V. V., Krupovic, M., Varsani, A., Wolf, Y. I., Yutin, N., Zerbini, F. M., & Kuhn, J. H. (2020). Global organization and proposed megataxonomy of the virus world. *Microbiology and molecular biology reviews: MMBR*, 84(2).

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution*, 35(6), 1547-1549.

Lacroix, C., Jolles, A., Seabloom, E. W., Power, A. G., Mitchell, C. E., & Borer, E. T. (2014). Non-random biodiversity loss underlies predictable increases in viral disease prevalence. *Journal of the Royal Society Interface*, 11(92), 20130947.

Lalić, J., & Elena, S. F. (2012). Magnitude and sign epistasis among deleterious mutations in a positive-sense plant RNA virus. *Heredity*, 109(2), 71-77.

Lam, H. M., Ratmann, O., & Boni, M. F. (2018). Improved algorithmic complexity for the 3SEQ recombination detection algorithm. *Molecular biology and evolution*, 35(1), 247-251.

Lambert, C., Braxton, C., Charlebois, R. L., Deyati, A., Duncan, P., La Neve, F., Malicki, H. D., Ribrioux, S., Rozelle, D. K., Michaels, B., Sun, W., Yang, Z., & Khan, A. S. (2018). Considerations for optimization of high-throughput sequencing bioinformatics pipelines for virus detection. *Viruses*, 10(10), 528.

Lana, J. M., & Iriarte-Goñi, I. (2015). Commons and the legacy of the past. Regulation and uses of common lands in twentieth century Spain. *International Journal of the Commons*, 9(2).

Lavorel, S., & Garnier, E. (2002). Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail. *Functional ecology*, 16(5), 545-556.

Lecoq, H. (1992). Les virus des cultures de melon et de courgette de plein champ. II. PHM, *Revue horticole*, (324), 15-25.

Lecoq, H., & Desbiez, C. (2008). Watermelon mosaic virus and Zucchini yellow mosaic virus. ‘‘*Encyclopedia of Virology, Third Edition”* (B. W. J. Mahy and M. H. V. Van Regenmortel, eds.), pp. 433–440. Elsevier, Oxford, UK, 5 Vols.

Lecoq, H., Fabre, F., Joannon, B., Wipf-Scheibel, C., Chandeysson, C., Schoeny, A., & Desbiez, C. (2011). Search for factors involved in the rapid

shift in Watermelon mosaic virus (WMV) populations in South-eastern France. *Virus Research*, 159(2), 115-123.

Lecoq, H., & Desbiez, C. (2012). Viruses of cucurbit crops in the Mediterranean region: an ever-changing picture. *Advances in virus research*, 84, 67-126.

Lefebvre, M., Theil, S., Ma, Y., & Candresse, T. (2019). The VirAnnot pipeline: A resource for automated viral diversity estimation and operational taxonomy units assignation for virome sequencing data. *Phytobiomes Journal*, 3(4), 256-259.

Legendre, P., Galzin, R., & Harmelin-Vivien, M. L. (1997). Relating behavior to habitat: solutions to thefourth-corner problem. *Ecology*, 78(2), 547-562.

Lenth, R. V. (2016) Least-squares means: The R package lsmeans *J. Stat. Softw.* 69 1 33. http://CRAN.R-project.org/package=lsmeans.

Levi, T., Keesing, F., Holt, R. D., Barfield, M., & Ostfeld, R. S. (2016). Quantifying dilution and amplification in a community of hosts for tick-borne pathogens. *Ecological Applications*, 26(2), 484-498.

Liang, X. Z., Lee, B. T., & Wong, S. M. (2002). Covariation in the capsid protein of Hibiscus chlorotic ringspot virus induced by serial passaging in a host that restricts movement leads to avirulence in its systemic host. *Journal of Virology*, 76(23), 12320-12324.

Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., & Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066), 355-359.

LoGiudice, K., Duerr, S. T., Newhouse, M. J., Schmidt, K. A., Killilea, M. E., & Ostfeld, R. S. (2008). Impact of host community composition on Lyme disease risk. *Ecology*, 89(10), 2841-2849.

Longdon, B., Brockhurst, M. A., Russell, C. A., Welch, J. J., & Jiggins, F. M. (2014). The evolution and genetics of virus host shifts. *PLoS Pathog*, 10(11), e1004395.

Løvdal, T., & Lillo, C. (2009). Reference gene selection for quantitative real-time PCR normalization in tomato subjected to nitrogen, cold, and light stress. *Analytical biochemistry*, 387(2), 238-242.

Luis Arteaga, M.P., Quiot, J.B., Leroux, J. (1976). Mise en évidence d'une souche de watermelon mosaic virus (WMV2) dans le Sud-Est de la France. *Annales de Phytopathologie* 8, 347–353.

Luis, A. D., Kuenzi, A. J., & Mills, J. N. (2018). Species diversity concurrently dilutes and amplifies transmission in a zoonotic host–pathogen system through competing mechanisms. *Proceedings of the National Academy of Sciences*, 115(31), 7979-7984.

Lynch, M., & Walsh, B. (1998). Genetics and analysis of quantitative traits Leal SM. Genetics and Analysis of Quantitative Traits. *Am J Hum Genet*. 2001;68(2):548-549.

Ma, Y., Marais, A., Lefebvre, M., Theil, S., Svanella-Dumas, L., Faure, C., & Candresse, T. (2019). Phytovirome analysis of wild plant populations: Comparison of double-stranded RNA and virion-associated nucleic acid metagenomic approaches. *Journal of virology*, 94(1), e01462-19.

Ma, Y., Marais, A., Lefebvre, M., Faure, C., & Candresse, T. (2020). Metagenomic analysis of virome cross-talk between cultivated Solanum lycopersicum and wild Solanum nigrum. *Virology*, 540, 38-44.

MacArthur, R. H., & MacArthur, J. W. (1961). On bird species diversity. *Ecology*, 42(3), 594-598.

Maclot, F., Candresse, T., Filloux, D., Malmstrom, C. M., Roumagnac, P., van der Vlugt, R., & Massart, S. (2020). Illuminating an ecological blackbox: Using High Throughput Sequencing to characterize the plant virome across scales. *Frontiers in Microbiology*, 11, 2575.

Hilario, E., & Mackay, J. (2007). *Protocols for nucleic acid analysis by nonradioactive probes*. 2nd ed. New York: Springer; 2007.

Malmstrom, C. M., McCullough, A. J., Johnson, H. A., Newton, L. A., & Borer, E. T. (2005). Invasive annual grasses indirectly increase virus incidence in California native perennial bunchgrasses. *Oecologia*, 145(1), 153-164.

Malmstrom, C. M., Melcher, U., & Bosque-Pérez, N. A. (2011). The expanding field of plant virus ecology: historical foundations, knowledge gaps, and research directions. *Virus research*, 159(2), 84-94.

Malmstrom, C. M., & Alexander, H. M. (2016). Effects of crop viruses on wild plants. *Current opinion in virology*, 19, 30-36.

Malpica, J. M., Sacristán, S., Fraile, A., & García-Arenal, F. (2006). Association and host selectivity in multi-host pathogens. *PloS ONE*, 1(1), e41.

Maree, H. J., Fox, A., Al Rwahnih, M., Boonham, N., & Candresse, T. (2018). Application of HTS for routine plant virus diagnostics: state of the art and challenges. *Frontiers in plant science*, 9, 1082.

Márquez, L. M., Redman, R. S., Rodriguez, R. J., & Roossinck, M. J. (2007). A virus in a fungus in a plant: three-way symbiosis required for thermal tolerance. *Science*, 315(5811), 513-515.

Martin, D., & Rybicki, E. (2000). RDP: detection of recombination amongst aligned sequences. *Bioinformatics*, 16(6), 562-563.

Martin, D. P., Posada, D., Crandall, K. A., & Williamson, C. (2005). A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Research & Human *Retroviruses*, 21(1), 98-102.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1), 10-12.

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

Martínez-Turiño, S., & García, J. A. (2020). Potyviral coat protein and genomic RNA: A striking partnership leading virion assembly and more. *Virus Assembly and Exit Pathways*, 108, 165.

Massart, S., Olmos, A., Jijakli, H., & Candresse, T. (2014). Current impact and future directions of high throughput sequencing in plant virus diagnostics. *Virus research*, 188, 90-96.

Massart S., Candresse T., Gil J., Lacomme C., Predajna L., Ravnikar M., & Wetzel T. (2017). A framework for the evaluation of biosecurity, commercial, regulatory, and scientific impacts of plant viruses and viroids identified by NGS technologies. *Frontiers in microbiology*, 8, 45.

Massart, S., Chiumenti, M., De Jonghe, K., Glover, R., Haegeman, A., Koloniuk, I., Komínek, P., Kreuze, J., Kutnjak, D., Lotos, L., Maclot, F., Maliogka, V., Maree, H.J., Olivier, T., Olmos, A., Pooggin, M.M., Reynard, J.-S., Ruiz-García, A.B., Safarova, D., Schneeberger, P.H.H., Sela, N., Turco, S., Vainio, E.J., Varallyay, E., Verdin, E., Westenberg, M., Brostaux, Y., & Candresse, T. (2019). Virus detection by high-throughput sequencing of small RNAs: large-scale performance testing of sequence analysis strategies. *Phytopathology*, 109(3), 488-497.

May, R. M., & Nowak, M. A. (1994). Superinfection, metapopulation dynamics, and the evolution of diversity. *Journal of Theoretical Biology*, 170(1), 95-114.

Maynard Smith J. (1992) 'Analyzing the Mosaic Structure of Genes' , *Journal of Molecular Evolution* , 34 : 126 –9 .

McArthur, D. B. (2019). Emerging infectious diseases. *Nursing Clinics*, 54(2), 297-311.

McLeish, M., Sacristán, S., Fraile, A., & Garcia-Arenal, F. (2017). Scale dependencies and generalism in host use shape virus prevalence. *Proceedings of the Royal Society B: Biological Sciences*, 284(1869), 20172066.

McLeish, M. J., Fraile, A., & García-Arenal, F. (2018). Ecological complexity in plant virus host range evolution. *Advances in virus research*, 101, 293-339.

McLeish, M. J., Fraile, A., & García-Arenal, F. (2019). Evolution of plant–virus interactions: host range and virus emergence. *Current opinion in virology*, 34, 50-55.

McLeish, M. J., Fraile, A., & García-Arenal, F. (2020a). Trends and gaps in forecasting plant virus disease risk. *Annals of Applied Biology*, 176(2), 102-108.

McLeish, M. J., Fraile, A., & García-Arenal, F. (2020b). Population genomics of plant viruses: the ecology and evolution of virus emergence. *Phytopathology®*, 111(1), 32-39.

McLeish, M., Peláez, A., Pagán, I., Gavilán, R., Fraile, A., & García-Arenal, F. (2021). Structuring of plant communities across agricultural landscape mosaics: the importance of connectivity and the scale of effect. *BMC Ecology and Evolution*, *21*(1), 1-12.

Medail, F., & Diadema, K. (2009). Glacial refugia influence plant diversity patterns in the Mediterranean Basin. *J Biogeogr*, 36(7), 1333.

Melcher, U., Muthukumar, V., Wiley, G. B., Min, B. E., Palmer, M. W., Verchot-Lubicz, J., Ali, A., Nelson, R. S., Roe, B. A., Thapa, V., & Pierce M. L., (2008). Evidence for novel viruses by analysis of nucleic acids in virus-like particle fractions from Ambrosia psilostachya. *Journal of Virological Methods*, 152(1-2), 49-55.

Melcher, U., & Grover, V. (2011). Genomic approaches to discovery of viral species diversity of non-cultivated plants. *Recent advances in plant virology*, 321-342.

Mendes, R. S., Evangelista, L. R., Thomaz, S. M., Agostinho, A. A., & Gomes, L. C. (2008). A unified index to measure ecological diversity and species rarity. *Ecography*, 31(4), 450-456.

Mihaljevic, J. R., Joseph, M. B., Orlofske, S. A., & Paull, S. H. (2014). The scaling of host density with richness affects the direction, shape, and detectability of diversity-disease relationships. *PloS ONE*, 9(5), e97812.

Miyashita, Y., Atsumi, G., & Nakahara, K. S. (2016). Trade-offs for viruses in overcoming innate immunities in plants. *Molecular Plant-Microbe Interactions*, 29(8), 595-598.

Mlotshwa, S., Pruss, G. J., & Vance, V. (2008). Small RNAs in viral infection and host defense. *Trends in plant science*, 13(7), 375-382.

Montarry, J., Doumayrou, J., Simon, V., & Moury, B. (2011). Genetic background matters: a plant–virus gene-for-gene interaction is strongly influenced by genetic contexts. *Molecular plant pathology*, 12(9), 911-920.

Montarry, J., Cartier, E., Jacquemond, M., Palloix, A., & Moury, B. (2012). Virus adaptation to quantitative plant resistance: erosion or breakdown?. *Journal of evolutionary biology*, 25(11), 2242-2252.

Moreno, I. M., Malpica, J. M., Dıaz-Pendón, J. A., Moriones, E., Fraile, A., & Garcıa-Arenal, F. (2004). Variability and genetic structure of the population of watermelon mosaic virus infecting melon in Spain. *Virology*, 318(1), 451-460.

Moreno-Pérez, M. G., García-Luque, I., Fraile, A., & García-Arenal, F. (2016). Mutations that determine resistance breaking in a plant RNA virus have pleiotropic effects on its fitness that depend on the host environment and on the type, single or mixed, of infection. *Journal of virology*, 90(20), 9128-9137.

Moury, B., & Simon, V. (2011). dN/dS-based methods detect positive selection linked to trade-offs between different fitness traits in the coat protein of potato virus Y. *Molecular biology and evolution*, 28(9), 2707-2717.

Moury, B., Fabre, F., Hébrard, E., & Froissart, R. (2017). Determinants of host species range in plant viruses. *Journal of General Virology*, 98(4), 862-873.

Moya-Ruiz, D., Rabadán, P., Juárez, M., & Gómez, P. (2021). Assessment of the current status of potyviruses in watermelon and pumpkin crops in Spain: Epidemiological impact of cultivated plants and mixed infections. *Plants*, *10*(1), 138.

Muratet, A., Machon, N., Jiguet, F., Moret, J., & Porcher, E. (2007). The role of urban structures in the distribution of wasteland flora in the greater Paris area, France. *Ecosystems*, 10(4), 661-671.

Muthukumar, V., Melcher, U., Pierce, M., Wiley, G. B., Roe, B. A., Palmer, Thapa, V., Ali, A., & Ding, T. (2009). Non-cultivated plants of the

Tallgrass Prairie Preserve of northeastern Oklahoma frequently contain virus-like sequences in particulate fractions. *Virus Research*, 141(2), 169-173.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in ecology and evolution*, 4(2), 133-142.

Nash, K. L., Allen, C. R., Angeler, D. G., Barichievy, C., Eason, T., Garmestani, A. S., Graham, N. A. J., Granholm, D., Knutson, M., Nelson, R. J., Nyström, M., Stow, C. A., Sundstrom S. M. (2014). Discontinuities, cross-scale patterns, and the organization of ecosystems. *Ecology*, 95(3), 654-667.

Nei, M., & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular biology and evolution*, 3(5), 418-426.

Nerva, L., Forgia, M., Ciuffo, M., Chitarra, W., Chiapello, M., Vallino, M., Varese, G. C., & Turina, M. (2019). The mycovirome of a fungal collection from the sea cucumber Holothuria polii. *Virus research*, 273, 197737.

Ng, S.H., Braxton, C., Eloit, M., Feng, S.F., Fragnoud, R., Mallet, L., Mee, E.T., Sathiamoorthy, S., Vandeputte, O., Khan, A.S. (2018). Current perspectives on high-throughput sequencing (HTS) for adventitious virus detection: upstream sample processing and library preparation. *Viruses*, 10(10), 566.

Nicaise, V. (2014). Crop immunity against viruses: outcomes and future challenges. *Frontiers in plant science*, 5, 660.

Nielsen, A., & Bascompte, J. (2007). Ecological networks, nestedness and sampling effort. *Journal of Ecology*, 1134-1141.

Novotny, V., Basset, Y., Miller, S. E., Weiblen, G. D., Bremer, B., Cizek, L., & Drozd, P. (2002). Low host specificity of herbivorous insects in a tropical forest. *Nature*, 416(6883), 841-844.

O'Dwyer, J. P., & Cornell, S. J. (2018). Cross-scale neutral ecology and the maintenance of biodiversity. *Scientific reports*, 8(1), 1-8.

Öckinger, E., Dannestam, Å., & Smith, H. G. (2009). The importance of fragmentation and habitat quality of urban grasslands for butterfly diversity. *Landscape and Urban Planning*, 93(1), 31-37

Oerke, E. C. (2006). Crop losses to pests. *The Journal of Agricultural Science*, 144(1), 31-43.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., & O'Hara, R. B. (2015). vegan: community ecology package. R package version, 2.2.1.2015. http://CRAN.R-project.org/package= vegan. Accessed 12 Mar 2020.

Ostfeld, R. S., & Keesing, F. (2000). Biodiversity series: the function of biodiversity in the ecology of vector-borne zoonotic diseases. *Canadian Journal of Zoology*, 78(12), 2061-2078.

Ostfeld, R. S., & LoGiudice, K. (2003). Community disassembly, biodiversity loss, and the erosion of an ecosystem service. *Ecology*, 84(6), 1421-1427.

Ostfeld, R. S., & Keesing, F. (2012). Effects of host diversity on infectious disease. *Annual review of ecology, evolution, and systematics*, 43.

Padidam, M., Sawyer, S., & Fauquet, C. M. (1999). Possible emergence of new geminiviruses by frequent recombination. *Virology*, 265(2), 218-225.

Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N.N., Kyrpides, N.C. (2016). *Uncovering Earth's virome*. *Nature*, *536*(7617), 425-430.

Pagán, I., González-Jara, P., Moreno-Letelier, A., Rodelo-Urrego, M., Fraile, A., Piñero, D., & García-Arenal, F. (2012). Effect of biodiversity changes in disease risk: exploring disease emergence in a plant-virus system. *PLoS Pathog*, 8(7), e1002796.

Pandey, V. C., Patel, D., Maiti, D., & Singh, D. P. (2020). Case studies of perennial grasses: Phytoremediation (holistic approach). *Phytoremediation potential of perennial Grasses* (pp. 337-349). Elsevier.

Pandit, S. N., Kolasa, J., & Cottenie, K. (2009). Contrasts between habitat generalists and specialists: an empirical extension to the basic metacommunity framework. *Ecology*, 90(8), 2253-2262.

Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, 26(3), 419-420. http://CRAN.R-project.org/package= pegas.

Parker, J., Rambaut, A., & Pybus, O. G. (2008). Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. Infection, *Genetics and Evolution*, 8(3), 239-246.

Parry, J. N., & Persley, D. M. (2005). Carrot as a natural host of *Watermelon mosaic virus*. *Australasian Plant Pathology*, 34(2), 283-284.

Peakall, R. O. D., & Smouse, P. E. (2006). GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular ecology notes*, 6(1), 288-295.

Pecman, A., Kutnjak, D., Gutiérrez-Aguirre, I., Adams, I., Fox, A., Boonham, N., & Ravnikar, M. (2017). Next generation sequencing for detection and discovery of plant viruses and viroids: comparison of two approaches. *Frontiers in microbiology*, 8, 1998.

Peng, Y., Kadoury, D., Gal-On, A., Huet, H., Wang, Y., & Raccah, B. (1998). Mutations in the HC-Pro gene of zucchini yellow mosaic potyvirus: effects on aphid transmission and binding to purified virions. *Journal of General Virology*, 79(4), 897-904.

Pérez-Padilla, V., Fortes, I. M., Romero-Rodríguez, B., Arroyo-Mateos, M., Castillo, A. G., Moyano, C., De León, L., & Moriones, E. (2020). Revisiting Seed Transmission of the Type Strain of Tomato yellow leaf curl virus in Tomato Plants. *Phytopathology*, 110(1), 121-129.

Peterson, G., Allen, C. R., & Holling, C. S. (1998). Ecological resilience, biodiversity, and scale. *Ecosystems*, 1(1), 6-18.

Posada, D., & Crandall, K. A. (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the National Academy of Sciences*, 98(24), 13757-13762.

Posada, D. (2002). Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Molecular biology and evolution*, 19(5), 708-717.

Posada, D. (2008). jModelTest: phylogenetic model averaging. *Molecular biology and evolution*, 25(7), 1253-1256.

Poulicard, N., Pinel-Galzi, A., Hebrard, E., & Fargette, D. (2010). Why Rice yellow mottle virus, a rapidly evolving RNA plant virus, is not efficient at breaking rymv1–2 resistance. *Molecular plant pathology*, 11(1), 145-154.

Poulicard, N., Pinel-Galzi, A., Traore, O., Vignols, F., Ghesquiere, A., Konate, G., Hebrard, E., & Fargette, D. (2012). Historical contingencies modulate the adaptability of Rice yellow mottle virus. *PLoS Pathog*, 8(1), e1002482.

Poulin, R., & Mouillot, D. (2003). Parasite specialization from a phylogenetic perspective: a new index of host specificity. *Parasitology*, 126(5), 473.

Power, A. G., & Mitchell, C. E. (2004). Pathogen spillover in disease epidemics. *The american naturalist*, 164(S5), S79-S89.

Power, A. G., Borer, E. T., Hosseini, P., Mitchell, C. E., & Seabloom, E. W. (2011). The community ecology of barley/cereal yellow dwarf viruses in Western US grasslands. *Virus research*, 159(2), 95-100.

Purcifull, D. E., & Hiebert, E. (1979). Serological distinction of watermelon mosaic virus isolates. *Phytopathology*, 69(2), 112-116.

R Core Team (2020). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. http://www.R-project.org/.

Ramachandran, P., & Perkins, T. J. (2013). Adaptive bandwidth kernel density estimation for next-generation sequencing data. *BMC proceedings* (Vol. 7, No. 7, pp. 1-10). BioMed Central.

Rambaut, A., & Drummond, A. (2012). FigTree: Tree figure drawing tool, v1. 4.2. Institute of Evolutionary Biology, University of Edinburgh.

Rambaut, A., Suchard, M. A., Xie, D., & Drummond, A. J. (2014). Tracer v1. 6. 2014.Retrieved from http://beast.bio.ed.ac.uk/Tracer.

Randolph, S. E., & Dobson, A. D. (2012). Pangloss revisited: a critique of the dilution effect and the biodiversity-buffers-disease paradigm. *Parasitology*, 139(7), 847-863.

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, 5(1), 1-20.

Revers, F., Le Gall, O., Candresse, T., & Maule, A. J. (1999). New advances in understanding the molecular biology of plant/potyvirus interactions. *Molecular Plant-Microbe Interactions*, 12(5), 367-376.

Revers, F., & García, J. A. (2015). Molecular biology of potyviruses. *Advances in virus research*, 92, 101-199.

Richardson, R. T., Sponsler, D. B., McMinn-Sauder, H., & Johnson, R. M. (2020). MetaCurator: A hidden Markov model-based toolkit for extracting and curating sequences from taxonomically-informative genetic markers. Methods in *Ecology and Evolution*, 11(1), 181-186.

Rico, P., Ivars, P., Elena, S. F., & Hernández, C. (2006). Insights into the selective pressures restricting Pelargonium flower break virus genome variability: evidence for host adaptation. *Journal of Virology*, 80(16), 8124-8132.

Roche, B., Rohani, P., Dobson, A. P., & Guégan, J. F. (2013). The impact of community organization on vector-borne pathogens. *The American Naturalist*, 181(1), 1-11.

Rodelo-Urrego, M., Pagán, I., González-Jara, P., Betancourt, M., Moreno-Letelier, A., Ayllón, M.A., Fraile, A., Piñero, D., & García-Arenal F.

(2013). Landscape heterogeneity shapes host-parasite interactions and results in apparent plant–virus codivergence. *Molecular Ecology*, 22(8), 2325-2340.

Rodelo-Urrego, M., García-Arenal, F., & Pagán, I. (2015). The effect of ecosystem biodiversity on virus genetic diversity depends on virus species: A study of chiltepin-infecting begomoviruses in Mexico. *Virus evolution*, 1(1).

Roenhorst, J. W., De Krom, C., Fox, A., Mehle, N., Ravnikar, M., & Werkman, A. W. (2018). Ensuring validation in diagnostic testing is fit for purpose: a view from the plant virology laboratory. *EPPO Bulletin*, 48(1), 105-115.

Rohwer, R. R., Hamilton, J. J., Newton, R. J., & McMahon, K. D. (2018). TaxAss: leveraging a custom freshwater database achieves fine-scale taxonomic resolution. *MSphere*, 3(5), e00327-18.

Romero-Calcerrada, R., & Perry, G. L. (2004). The role of land abandonment in landscape dynamics in the SPA 'Encinares del río Alberche y Cofio, Central Spain, 1984–1999. *Landscape and Urban Planning*, 66(4), 217-232.

Roossinck, M. J., Saha, P., Wiley, G. B., Quan, J., White, J. D., Lai, H., Chavarria, F., Shen, G. A., & Roe, B. A. (2010). Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. *Molecular Ecology*, 19, 81-88.

Roossinck, M. J. (2011). The good viruses: viral mutualistic symbioses. *Nature Reviews Microbiology*, 9(2), 99-108.

Roossinck, M. J. (2012). Plant virus metagenomics: biodiversity and ecology. *Annual review of genetics*, 46, 359-369.

Roossinck, M. J. (2015). Plants, viruses and the environment: ecology and mutualism. *Virology*, 479, 271-277.

Roossinck, M. J., & García-Arenal, F. (2015). Ecosystem simplification, biodiversity loss and plant virus emergence. *Current opinion in virology*, 10, 56-62.

Rosenzweig M. L. (1995). *Species Diversity in Space and Time*. Cambridge, MA: Cambridge Univ. Press

Sacristán, S., Fraile, A., & García-Arenal, F. (2004). Population dynamics of *Cucumber mosaic virus* in melon crops and in weeds in central Spain. *Phytopathology*, 94(9), 992-998.

Sambrook, J., Fritsch, E.F. & Maniatis, T. (1989) Molecular cloning: A laboratory manual. *Current protocols in molecular biology*, 2, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA.

Sandhya, S., Kishore, S., Sowdhamini, R., & Srinivasan, N. (2003). Effective detection of remote homologues by searching in sequence dataset of a protein domain fold. *FEBS letters*, 552(2-3), 225-230.

Sanjuán, R., Moya, A., & Elena, S. F. (2004). The contribution of epistasis to the architecture of fitness in an RNA virus. Proceedings of the National. *Academy of Sciences*, 101(43), 15376-15379.

Savary, S., Teng, P. S., Willocquet, L., & Nutter Jr, F. W. (2006). Quantification and modeling of crop losses: a review of purposes. *Annu. Rev. Phytopathol.,* 44, 89-112.

Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4), 592-593. R package version 2.7.1. http://CRAN.R-project.org/package= phangorn.

Schmidt, K. A., & Ostfeld, R. S. (2001). Biodiversity and the dilution effect in disease ecology. *Ecology*, 82(3), 609-619.

Schoelz, J. E., & Stewart, L. R. (2018). The role of viruses in the phytobiome. *Annual review of virology*, 5, 93-111.

Seabloom, E. W., Borer, E. T., Gross, K., Kendig, A. E., Lacroix, C., Mitchell, C. E., Mordecai, E. A., & Power, A. G. (2015). The community ecology of pathogens: coinfection, coexistence and community composition. *Ecology Letters*, 18(4), 401-415.

Sexton, J. P., Montiel, J., Shay, J. E., Stephens, M. R., & Slatyer, R. A. (2017). Evolution of ecological niche breadth. *Annual Review of Ecology, Evolution, and Systematics*, 48, 183-206.

Shah, N., Altschul, S. F., & Pop, M. (2018). Outlier detection in BLAST hits. *Algorithms for Molecular Biology*, 13(1), 1-9.

Sharp, P. M., & Hahn, B. H. (2010). The evolution of HIV-1 and the origin of AIDS. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1552), 2487-2494.

Shates, T. M., Sun, P., Malmstrom, C. M., Dominguez, C., & Mauck, K. E. (2019). Addressing research needs in the field of plant virus ecology by defining knowledge gaps and developing wild dicot study systems. *Frontiers in microbiology*, 9, 3305.

Shen, W., Shi, Y., Dai, Z., & Wang, A. (2020). The RNA-dependent RNA polymerase NIb of potyviruses plays multifunctional, contrasting roles during viral infection. *Viruses*, 12(1), 77.

Shipley, L. A., Forbey, J. S., & Moore, B. D. (2009). Revisiting the dietary niche: when is a mammalian herbivore a specialist?. *Integrative and Comparative Biology*, 49(3), 274-290.

Singhal, S., Gomez, S. M., & Burch, C. L. (2019). Recombination drives the evolution of mutational robustness. *Current opinion in systems biology*, 13, 142-149.

Soberón, J., & Peterson, A. T. (2005). Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodivers. Inform.* 2, 1–10

Sokal R., & Rohlf F. (1995) *Biometry. 3rd edn*. W.H. Freeman and Company, USA.

Sorel, M., Garcia, J. A., & German-Retana, S. (2014). The Potyviridae cylindrical inclusion helicase: a key multipartner and multifunctional protein. *Molecular plant-microbe interactions*, 27(3), 215-226.

Strauss, A. T., Civitello, D. J., Cáceres, C. E., & Hall, S. R. (2015). Success, failure and ambiguity of the dilution effect among competitors. *Ecology letters*, 18(9), 916-926.

Streicker, D. G., Turmelle, A. S., Vonhof, M. J., Kuzmin, I. V., McCracken, G. F., & Rupprecht, C. E. (2010). Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. *Science*, 329(5992), 676-679.

Streicker, D. G., Fenton, A., & Pedersen, A. B. (2013). Differential sources of host species heterogeneity influence the transmission and control of multihost parasites. *Ecology letters*, 16(8), 975-984.

Stukenbrock, E. H., & McDonald, B. A. (2008). The origins of plant pathogens in agro-ecosystems. *Annual Reviews of Phytopathology.*, 46, 75-100.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2), 437-460.

Tamburini, G., De Simone, S., Sigura, M., Boscutti, F., & Marini, L. (2016). Conservation tillage mitigates the negative effect of landscape simplification on biological control. *Journal of Applied Ecology*, 53(1), 233-241.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular biology and evolution*, 30(12), 2725-2729.

Tavert-Roudet, G., Anne, A., Barra, A., Chovin, A., Demaille, C., & Michon, T. (2017). The potyvirus particle recruits the plant translation initiation factor eIF4E by means of the VPg covalently linked to the viral RNA. *Molecular Plant-Microbe Interactions*, 30(9), 754-762.

Templeton, A. R., Crandall, K. A., & Sing, C. F. (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics*, 132(2), 619-633.

Tews, J., Brose, U., Grimm, V., Tielbörger, K., Wichmann, M. C., Schwager, M., & Jeltsch, F. (2004). Animal species diversity driven by habitat

heterogeneity/diversity: the importance of keystone structures. *Journal of biogeography*, 31(1), 79-92.

Thomas, R. (2019). Marine Biology: An Ecological Approach. *Ed-Tech Press*, Waltham Abbey

Thompson, P. L., Rayfield, B., & Gonzalez, A. (2017). Loss of habitat and connectivity erodes species diversity, ecosystem functioning, and stability in metacommunity networks. *Ecography*, 40(1), 98-108.

Tilman, D., Knops, J., Wedin, D., Reich, P., Ritchie, M., & Siemann, E. (1997). The influence of functional diversity and composition on ecosystem processes. *Science*, 277(5330), 1300-1302.

Tilman, D., Cassman, K. G., Matson, P. A., Naylor, R., & Polasky, S. (2002). Agricultural sustainability and intensive production practices. *Nature*, 418(6898), 671-677.

Tischendorf, L., & Fahrig, L. (2000). On the usage and measurement of landscape connectivity. *Oikos*, 90(1), 7-19.

Torre, C., Agüero, J., Gómez-Aix, C., & Aranda, M. A. (2020). Comparison of DAS-ELISA and qRT-PCR for the detection of cucurbit viruses in seeds. *Annals of Applied Biology*, 176(2), 158-169.

Travlos, I. S., Cheimona, N., Roussis, I., & Bilalis, D. J. (2018). Weed-species abundance and diversity indices in relation to tillage systems and fertilization. *Frontiers in Environmental Science*, 6, 11.

Tutin, T.; Heywood, V.; Burges, N.; Moore, D.; Valentine, D.; Walters, S.; Webb, D. (2010). Flora Europaea (2nd ed.). *Cambridge University Press*: Cambridge, UK, 2010.

Twerd, L., & Banaszak-Cibicka, W. (2019). Wastelands: their attractiveness and importance for preserving the diversity of wild bees in urban areas. *Journal of insect conservation*, 23(3), 573-588.

Van Buskirk, J., & Ostfeld, R. S. (1995). Controlling Lyme disease by modifying the density and species composition of tick hosts. *Ecological Applications*, 5(4), 1133-1140.

Van der Heijden, M. G., Boller, T., Wiemken, A., & Sanders, I. R. (1998). Different arbuscular mycorrhizal fungal species are potential determinants of plant community structure. *Ecology*, 79(6), 2082-2091.

Vázquez, D. P., Poulin, R., Krasnov, B. R., & Shenbrot, G. I. (2005). Species abundance and the distribution of specialization in host–parasite interaction networks. *Journal of Animal Ecology*, 74(5), 946-955.

Velez-Olmedo, J. B., Quiñonez, L. C., Vélez-Zambrano, S. M., Monteros-Altamirano, A., De Oliveira, A. S., & Resende, R. O. (2021). Low virus diversity and spread in wild Capsicum spp. accessions from Ecuador under natural inoculum pressure. *Archives of Virology*, 166(5), 1447-1453.

Verma, R. K., Mishra, M., Marwal, A., & Gaur, R. K. (2020). Identification, genetic diversity and recombination analysis of *Watermelon Mosaic Virus* isolates. *3 Biotech*, 10, 1-8.

Via, S., & Hawthorne, D. J. (2002). The genetic architecture of ecological specialization: correlated gene effects on host use and habitat choice in pea aphids. *The american naturalist*, 159(S3), S76-S88.

Viana, M., Mancy, R., Biek, R., Cleaveland, S., Cross, P. C., Lloyd-Smith, J. O., & Haydon, D. T. (2014). Assembling evidence for identifying reservoirs of infection. *Trends in ecology & evolution*, 29(5), 270-279.

Viana, D. S., & Chase, J. M. (2019). Spatial scale modulates the inference of metacommunity assembly processes. *Ecology*, 100(2), e02576.

Vurro, M., Bonciani, B., & Vannacci, G. (2010). Emerging infectious diseases of crop plants in developing countries: impact on agriculture and socio-economic consequences. *Food Security*, 2(2), 113-132.

Waggott, W. (1998). Long range PCR. *Clinical Applications of PCR* (pp. 81-91). Humana Press.

Wall, D. P., Fraser, H. B., & Hirsh, A. E. (2003). Detecting putative orthologs. *Bioinformatics*, 19(13), 1710-1711.

Wallis, C. M., Stone, A. L., Sherman, D. J., Damsteegt, V. D., Gildow, F. E., & Schneider, W. L. (2007). Adaptation of plum pox virus to a herbaceous

host (Pisum sativum) following serial passages. *Journal of General Virology*, 88(10), 2839-2845.

Wang, Y., Gaba, V., Yang, J., Palukaitis, P., & Gal-On, A. (2002). Characterization of synergy between *Cucumber mosaic virus* and potyviruses in cucurbit hosts. *Phytopathology*, 92(1), 51-58.

Wang, Y., & Hajimorad, M. R. (2016). Gain of virulence by *Soybean mosaic virus* on Rsv4-genotype soybeans is associated with a relative fitness loss in a susceptible host. *Molecular plant pathology*, 17(7), 1154-1159.

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. (2015). So many variables: joint modeling in community ecology. *Trends in Ecology & Evolution*, 30(12), 766-779.

Waterworth, H. E., & Hadidi, A. (1998). Economic losses due to plant viruses. *Plant virus disease control.* (ed. by A. Hadidi, R. K. Khetarpal and H. Koganezawa), pp. 1–13. The American Phytopathological Society, St. Paul, Minnesota.

Weiss N. (2015). wPerm: Permutation Tests. R package version, 1.0.1. http://CRAN.R-project.org/package= wPerm. Accessed 16 Sep 2019.

Weitz, J. S., Poisot, T., Meyer, J. R., Flores, C. O., Valverde, S., Sullivan, M. B., & Hochberg, M. E. (2013). Phage–bacteria infection networks. *Trends in microbiology*, 21(2), 82-91.

Wells, K., & Clark, N. J. (2019). Host specificity in variable environments. *Trends in parasitology*, 35(6), 452-465.

Welsh, M. E., Cronin, J. P., & Mitchell, C. E. (2020). Trait-based variation in host contribution to pathogen transmission across species and resource supplies. *Ecology*, 101(11), e03164.

Whitlock, M. C. (1996). The red queen beats the jack-of-all-trades: the limitations on the evolution of phenotypic plasticity and niche breadth. *The American Naturalist*, 148, S65-S77.

Whittaker, R. J., Willis, K. J., & Field, R. (2001). Scale and species richness: towards a general, hierarchical theory of species diversity. *Journal of biogeography*, 28(4), 453-470.

Wickham, H., Chang, W., & Wickham, M. H. (2016). Package 'ggplot2'. Create Elegant Data Visualisations Using the Grammar of Graphics. Version, 1.0.0. R package version, 2(1), 74. https://github.com/tidyverse/ggplot2. Accessed 24 Dec 2016.

Wilson, P. J., & Aebischer, N. J. (1995). The distribution of dicotyledonous arable weeds in relation to distance from the field edge. *Journal of applied ecology*, 295-310.

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome biology*, 20(1), 1-13.

Woolhouse, M. E. (2002). Population biology of emerging and re-emerging pathogens. *Trends in microbiology*, 10(10), s3-s7.

Woolhouse, M. E., Haydon, D. T., & Antia, R. (2005). Emerging pathogens: the epidemiology and evolution of species jumps. *Trends in ecology & evolution*, 20(5), 238-244.

Woolhouse, M. E., & Gowtage-Sequeria, S. (2005). Host range and emerging and reemerging pathogens. *Emerging infectious diseases*, 11(12), 1842.

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(3), e80.

Wu, Q., Luo, Y., Lu, R., Lau, N., Lai, E. C., Li, W. X., & Ding, S. W. (2010). Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. *Proceedings of the National Academy of Sciences*, 107(4), 1606-1611.

Wu, Q., Ding, S. W., Zhang, Y., & Zhu, S. (2015). Identification of viruses and viroids by next-generation sequencing and homology-dependent and

homology-independent algorithms. *Annual review of phytopathology*, 53, 425-444.

Wylie, S. J., Adams, M., Chalam, C., Kreuze, J., López-Moya, J. J., Ohshima, K., Praveen, S., Rabenstein, F., Stenger, D., Wang, A., Zerbini, FM., & Consortium, I. R. (2017). ICTV virus taxonomy profile: Potyviridae. *The Journal of general virology*, 98(3), 352.

Xu, C., Sun, X., Taylor, A., Jiao, C., Xu, Y., Cai, X., Wang, X., Ge, C., Pan, G., Wang, Q., Fei, Z.., Wang, Q. (2017). Diversity, distribution, and evolution of tomato viruses in China uncovered by small RNA sequencing. *Journal of virology*, 91(11), e00173-17.

Yockteng, R., Almeida, A. M., Yee, S., Andre, T., Hill, C., & Specht, C. D. (2013). A method for extracting high-quality RNA from diverse plants for next-generation sequencing and gene expression analyses. *Applications in plant sciences*, 1(12), 1300070.

Young, S. G., Tullis, J. A., & Cothren, J. (2013). A remote sensing and GIS-assisted landscape epidemiology approach to West Nile virus. *Applied Geography*, 45, 241-249.

Young H. S., Wood C. L., Kilpatrick A. M., Lafferty K. D., Nunn C. L., Vincent J. R. (2017) Conservation, biodiversity and infectiousdisease: scientific evidence and policy implications.*Philosophical Transactions of the Royal Society B: Biological Sciences* 372:20160124.

Yu, Q., Liu, J., Zheng, H., Jia, Y., Tian, H., & Ding, Z. (2016). Topoisomerase II-associated protein PAT1H1 is involved in the root stem cell niche maintenance in Arabidopsis thaliana. *Plant cell reports*, 35(6), 1297-1307.

Yuen, J., & Mila, A. (2015). Landscape-scale disease risk quantification and prediction. *Annual Review of Phytopathology*, 53, 471-484.

Zamora, J., Verdú, J. R., & Galante, E. (2007). Species richness in Mediterranean agroecosystems: spatial and temporal analysis for biodiversity conservation. *Biological conservation*, 134(1), 113-121.

8. ANNEXES

Table A1.1. RNA concentration (ng/µl) of each individual plant extract, purified by different tissue homogenisation methods and RNA extraction procedures as implemented in different kits.

	ID[1]	Homogenisation[2] I			Homogenisation[2] II			Homogenisation[2] III		
		Kit[3] A	Kit[3] B	Kit[3] C	Kit[3] A	Kit[3] B	Kit[3] C	Kit[3] A	Kit[3] B	Kit[3] C
Bromus rubeus L.	E1P3	70.4	40.5	159.1	72.9	16.8	17.9	67.3	4.3	59.7
	E1P29	57.2	86.6	246.0	38.1	4.8	14.9	11.3	4.1	81.4
	Q1P32	22.2	53.9	81.5	8.3	20.4	5.3	4.2	11.4	32.2
	Q1P38	60.6	50.3	107.9	31.0	31.8	17.1	9.7	12.5	24.9
	Q1P71	50.2	44.1	125.0	36.3	5.0	17.1	20.0	11.0	50.9
Carduus bourgeanus L.	L1F27	185.0	53.6	132.8	15.6	20.0	37.0	33.4	20.8	35.7
	L1F30	146.2	424.8	124.5	107.6	86.2	94.8	87.0	226.2	191.9
	L3P23	239.0	30.3	484.4	172.2	36.8	220.2	98.9	227.8	203.2
	L3P24	141.9	291.0	251.3	210.2	434.4	17.7	176.6	40.8	143.3
	L2P20	303.7	471.6	598.6	699.7	209.5	145.3	180.7	83.4	70.7
Cucumis melo L.	M1V3	190.7	265.2	431.6	498.6	98.1	247.3	944.9	48.1	335.8
	M1V4	456.5	392.2	669.9	1137.2	237.9	155.3	978.4	427.7	395.5
	M1V8	520.9	241.3	342.5	764.5	144.0	179.3	785.9	75.0	181.1
	M1V9	535.5	156.9	572.3	746.4	240.3	225.7	510.7	153.0	468.4
	M1V13	894.3	277.2	435.2	737.2	253.4	274.4	164.5	293.8	348.0
Papaver rhoeas L.	L3P20	335.8	573.2	585.0	374.8	182.9	158.9	212.1	73.0	382.8
	L2P8	188.1	115.3	377.5	85.8	25.3	44.5	101.7	33.1	7.0
	E3P10	219.0	56.7	47.1	53.1	4.4	12.6	63.4	7.8	13.1
	E3P53	104.1	316.5	73.1	34.3	117.8	22.5	67.9	10.9	24.7
	E3P127	273.7	85.2	91.9	27.0	6.6	91.0	40.4	11.6	42.9
Quercus ilex L.	Q3F15	340	11.4	606	101.5	20.6	239.5	40.7	4.3	216.8
	Q3F19	48.6	27	265.8	7.2	13.7	152.5	82	38.5	220.7
	Q3F5	53.7	33.7	73.5	49.7	48.5	2065.4	20.6	22.9	51.8
	Q1P1	33.3	14.2	189	61.9	6.8	112.7	21.4	10.3	161
	Q1P30	86	16.2	223.1	6.9	5.7	104.5	29.6	4.9	144.4
Stipa parviflora L.	E1P23	79.9	93.1	53.9	137.8	8	35	155.7	5.6	42.9
	E1P75	295.3	385.9	156.9	104	23.9	61	48.5	7.4	48.3
	E1P76	136.4	318.6	344.7	151.7	14.8	136.9	77.7	66.4	62.6
	E1P89	177.3	20.3	15.5	15.4	8.6	3.1	45.6	6.8	6.9
	Q1P69	116.4	131.5	33.7	31.5	18.2	8.7	142	10.3	3.9

[1]ID: individual plant extract code.

[2]Tissue homogenisation method: I; mortar with glass powder, II; Eppendorf tube without glass powder, III: Eppendorf tube with glass powder.

[3]Kit: A; Direct-zol™ RNA MiniPrep Plus, B; Rneasy Mini Kit. Kit, C; Agilent Plant RNA Isolation Mini Kit.

Table A1.2. Absorbance ratio at 260/230 nm of each individual plant extract, purified by different tissue homogenisation methods and RNA extraction procedures as implemented in different kits.

	ID[1]	Homogenisation[2] I			Homogenisation[2] II			Homogenisation[2] III		
		Kit[3] A	Kit[3] B	Kit[3] C	Kit[3] A	Kit[3] B	Kit[3] C	Kit[3] A	Kit[3] B	Kit[3] C
B. rubeus	E1P3	1.5	2.1	2.1	1.6	1.9	2.0	1.6	1.4	1.9
	E1P29	1.7	2.1	2.1	1.6	2.2	2.6	1.4	1.6	2.3
	Q1P32	1.6	1.8	2.0	1.3	1.4	1.6	1.0	1.5	1.9
	Q1P38	1.5	1.9	2.1	1.6	1.9	2.2	1.1	1.4	1.8
	Q1P71	1.6	2.0	2.0	1.3	1.6	1.9	1.4	1.5	2.1
C. bourgeanus	L1F27	1.6	1.9	1.9	1.4	1.7	2.1	1.5	1.8	2.1
	L1F30	1.7	2.1	2.1	1.7	2.0	2.1	1.7	2.1	2.1
	L3P23	1.8	1.9	2.1	1.8	2.1	2.2	1.8	2.1	2.2
	L3P24	2.0	2.1	2.2	1.9	2.1	2.3	2.0	2.1	2.2
	L2P20	1.9	2.1	2.1	2.0	2.2	2.1	1.9	2.1	2.1
C. melo	M1V3	2.1	2.1	2.1	2.0	2.0	2.2	2.1	2.0	2.1
	M1V4	2.0	2.1	2.2	2.1	2.1	2.1	2.1	2.1	2.1
	M1V8	2.0	2.1	2.1	2.1	2.0	2.2	2.0	2.0	2.1
	M1V9	2.0	2.1	2.1	2.1	2.1	2.1	2.0	2.1	2.1
	M1V13	2.0	2.1	2.1	2.1	2.1	2.1	2.0	2.1	2.1
P. rhoeas	L3P20	2.0	2.2	2.2	2.0	2.1	2.2	1.7	2.1	2.2
	L2P8	1.9	2.1	2.1	1.9	1.8	2.0	1.9	2.1	2.4
	E3P10	2.0	2.0	2.1	1.8	2.4	1.7	1.4	2.0	2.0
	E3P53	1.9	2.1	2.0	1.8	2.2	2.3	1.8	1.9	2.2
	E3P127	2.0	2.1	2.0	1.9	1.9	2.0	1.6	2.0	2.0
Q. ilex	Q3F15	1.4	1.4	1.6	1.5	1.4	1.7	1.3	1.1	1.7
	Q3F19	1.5	1.6	1.7	1.2	1.7	1.7	1.4	1.3	1.7
	Q3F5	1.4	1.7	1.5	1.4	1.8	1.5	1.3	1.7	1.3
	Q1P1	1.3	1.8	1.6	1.5	1.6	1.6	1.4	1.7	1.6
	Q1P30	1.3	1.9	1.6	1.1	1.3	1.6	1.4	1.3	1.6
S. parviflora	E1P23	2.1	2.1	2.2	1.8	2.1	2.2	2.0	1.4	2.1
	E1P75	2.1	2.1	2.2	2.1	2.0	2.2	1.7	2.2	2.3
	E1P76	2.1	2.1	2.1	2.1	2.2	2.1	2.0	1.9	2.3
	E1P89	2.0	2.1	2.6	2.1	1.8	8.5	1.6	2.3	2.2
	Q1P69	1.7	2.1	2.1	1.9	2.2	1.6	1.6	1.9	1.8

[1]ID: individual plant extract code.
[2]Tissue homogenisation method: I; mortar with glass powder, II; Eppendorf tube without glass powder, III: Eppendorf tube with glass powder.
[3]Kit: A; Direct-zol™ RNA MiniPrep Plus, B; Rneasy Mini Kit. Kit, C; Agilent Plant RNA Isolation Mini Kit.

Table A1.3. Absorbance ratio at 260/230 nm of each individual plant extract, purified by different tissue homogenisation methods and RNA extraction procedures as implemented in different kits.

	ID[1]	Homogenisation[2] I			Homogenisation[2] II			Homogenisation[2] III		
		Kit[3] A	Kit[3] B	Kit[3] C	Kit[3] A	Kit[3] B	Kit[3] C	Kit[3] A	Kit[3] B	Kit[3] C
B. rubeus	**E1P3**	0.9	1.7	2.3	1.1	1.8	2.2	1.0	1.3	2.4
	E1P29	1.4	2.1	2.4	1.1	1.0	2.5	0.7	2.0	2.4
	Q1P32	1.1	1.3	1.3	1.8	1.0	0.5	1.0	0.3	1.6
	Q1P38	0.9	0.2	1.9	0.8	0.4	1.3	0.8	0.2	1.2
	Q1P71	1.0	1.5	1.7	1.2	1.0	1.0	0.8	0.9	1.2
C. bourgeanus	**L1F27**	1.0	0.6	1.4	0.9	0.2	1.6	0.8	0.3	1.2
	L1F30	1.2	1.5	1.7	1.1	0.9	1.8	1.2	1.1	1.8
	L3P23	1.4	0.4	2.1	1.6	0.3	2.1	1.8	1.7	2.2
	L3P24	2.0	1.5	2.2	1.9	2.3	0.8	2.2	0.4	2.0
	L2P20	1.8	2.1	2.2	2.2	1.5	1.9	1.9	0.9	1.7
C. melo	**M1V3**	2.0	1.7	2.1	2.1	1.4	1.7	2.3	0.3	1.9
	M1V4	2.2	2.3	2.3	2.3	2.2	1.8	2.3	2.2	2.1
	M1V8	2.2	1.4	2.1	2.3	1.2	1.6	2.1	0.9	1.9
	M1V9	2.1	1.7	2.2	2.2	2.1	2.2	2.1	1.3	2.2
	M1V13	2.3	1.7	2.1	2.2	1.8	2.1	2.2	2.4	2.1
P. rhoeas	**L3P20**	1.9	1.8	2.3	1.7	1.3	2.2	1.2	0.9	2.3
	L2P8	1.5	2.1	2.0	1.9	0.4	1.0	1.8	0.6	0.4
	E3P10	2.0	0.4	1.6	1.8	0.1	0.5	0.9	0.3	0.7
	E3P53	1.3	1.6	1.4	1.1	2.1	1.0	1.3	1.4	1.1
	E3P127	2.0	1.9	1.2	1.1	0.8	1.3	0.9	0.2	0.9
Q. ilex	**Q3F15**	0.9	0.5	0.7	0.6	0.7	0.8	0.5	0.5	0.8
	Q3F19	0.6	1.0	1.0	0.5	1.0	0.8	0.9	1.3	0.9
	Q3F5	0.6	0.9	2.6	0.7	1.0	1.0	1.0	1.1	1.3
	Q1P1	0.5	1.3	0.7	0.6	0.7	0.6	0.6	1.0	0.7
	Q1P30	0.5	0.8	0.7	0.6	0.7	0.7	0.6	0.6	0.6
S. parviflora	**E1P23**	1.8	2.2	2.7	1.2	1.5	2.7	2.2	1.3	2.5
	E1P75	2.1	2.3	1.8	2.0	0.9	1.7	0.9	0.1	1.2
	E1P76	2.1	1.8	2.0	2.0	0.4	2.3	1.5	1.4	1.4
	E1P89	2.1	1.4	0.6	1.6	0.4	0.2	1.2	0.7	0.3
	Q1P69	1.1	1.4	1.3	2.1	0.3	1.1	1.1	2.2	0.5

[1]ID: individual plant extract code.
[2]Tissue homogenisation method: I; mortar with glass powder, II; Eppendorf tube without glass powder, III: Eppendorf tube with glass powder.
[3]Kit: A; Direct-zol™ RNA MiniPrep Plus, B; Rneasy Mini Kit. Kit, C; Agilent Plant RNA Isolation Mini Kit.

Table A2. Least Significant Difference (LSD) results among each class of each fixed factor in RNA extract quality.

Response variable	Variable	Estimate	S.E.	*P*-value
Concentration	Homogenisation I[1] vs Homogenisation II[1]	0.641	0.137	< 0.001
	Homogenisation I vs Homogenisation III[1]	0.921	0.135	< 0.001
	Homogenisation II vs Homogenisation III	0.280	0.134	0.0930
	Kit A[2] vs Kit B[2]	0.664	0.134	< 0.001
	Kit A vs Kit C[2]	0.172	0.142	0.447
	Kit B vs Kit C	0.836	0.144	<0.001
Absorbance 260/230	Homogenisation I vs Homogenisation II	0.031	0.063	0.875
	Homogenisation I vs Hmogenisation III	0.104	0.063	0.226
	Homogenisation II vs Homogenisation III	0.136	0.063	0.083
	Kit A vs Kit B	0.183	0.063	0.011
	Kit A vs Kit C	0.357	0.063	<0.001
	Kit B vs Kit C	0.173	0.063	0.018
Absorbance 260/280	Homogenisation I vs Homogenisation II	0.239	0.079	0.008
	Homogenisation I vs Hmogenisation III	0.301	0.079	<0.001
	Homogenisation II vs Homogenisation III	0.062	0.079	0.710
	Kit A vs Kit B	0.270	0.079	0.002
	Kit A vs Kit C	0.117	0.079	0.302
	Kit B vs Kit C	0.387	0.079	<0.001

[1]Tissue homogenisation methods: I; mortar with glass powder, II; Eppendorf without glass powder, III: Eppendorf with glass powder.
[2]RNA extraction procedures: A; Direct-zol™ RNA MiniPrep Plus, B; Rneasy Mini Kit. Kit, C; Agilent Plant RNA Isolation Mini Kit.

Table A3. Information on HTS libraries comprising HTS pool code, plant taxonomy, habitat and site of collection and number of RNA extracts from individual plant samples pooled at each library.

ID[1]	Species	Family	Habitat	Site of collection	RNA extracts pooled [2]	RIN
PV001	*Amaranthus sp* L.	*Amaranthaceae*	Cr	M1V	8	NA
PV002	*Convolvulus arvensis* L.	*Convolvulaceae*	Cr	M1V	11	7.5
PV003	*C. melo*	*Cucurbitaceae*	Cr	M1V	13	7.1
PV003bgi	*C. melo*	*Cucurbitaceae*	Cr	M1V	13	7.1
PV004	*Cyperus longus* L.	*Cyperaceae*	Cr	M1V	7	7.3
PV005	*Artemisia herba alta* Asso	*Asteraceae*	Oa	Q1F	12	4.3
PV006	*A. herba alta*	*Asteraceae*	Oa	Q1F	19	6.5
PV007	*Teucrium pseudochamaepitys* L.	*Lamiaceae*	Oa	Q1F	17	3.4
PV008	*Jasminum fruticans* L.	*Oleaceae*	Oa	Q1F	11	6.2
PV009	*Quercus coccifera* L.	*Fagaceae*	Oa	Q1F	13	7.6
PV009bgi	*Q. coccifera*	*Fagaceae*	Oa	Q1F	13	7.6
PV010	*Q. ilex*	*Fagaceae*	Oa	Q1F	10	7.6
PV011	*Portulaca oleraceae* L.	*Portulacaceae*	Cr	M1V	7	NA
PV012	*Chenopodium album* L.	*Amaranthaceae*	Cr	M2V	7	4.5
PV013	*Solanum nigrum* L.	*Solanaceae*	Cr	M2V	9	4.3
PV014	*Datura stramonium* L.	*Solanaceae*	Cr	M2V	8	5.4
PV015	*C. melo*	*Cucurbitaceae*	Cr	M2V	11	7.1
PV016	*C. melo*	*Cucurbitaceae*	Cr	M3V	16	7.0
PV017	*Lithospermum arvense* L.	*Boraginaceae*	Cr	M3V	12	7.2
PV018	*C. melo*	*Cucurbitaceae*	Cr	M4V	16	7.5
PV020	*L. arvense*	*Boraginaceae*	Cr	M4V	10	7.6
PV021	*C. melo*	*Cucurbitaceae*	Cr	M1V	13	7.0
PV022	*Taraxacum officinale* L..	*Asteraceae*	Wl	E4F	23	3.1

[1]HTS pool code.
[2] number of RNA extracts from individual plant samples pooled at each library.

Table A3. Continued.

ID[1]	Species	Family	Habitat	Site of collection	RNA extracts pooled [2]	RIN
PV023	*Leontodon sp* L.	*Asteraceae*	Wl	E4F	24	2.8
PV024	*Verbascum sinuatum* L.	*Scrophulariaceae*	Wl	E4F	15	7.0
PV025	*Geranium sp* L.	*Geraniaceae*	Wl	E4F	8	4.4
PV026	*Hieracium pilosella* L.	*Asteraceae*	Wl	E4F	7	6.3
PV027	*Festuca sp* L.	*Poaceae*	Wl	E4F	14	4.3
PV028	*Rubia peregrina* L.	*Rubiaceae*	Ed	L1V	6	4.6
PV029	*Picris echioides* L.	*Asteraceae*	Ed	L1V	8	3.8
PV030	*Milium vernale* Bieb.	*Poaceae*	Ed	L1V	7	3.0
PV031	*C. arvensis*	*Convolvulaceae*	Ed	L2F	9	3.1
PV032	*Brachypodium retusum* PERS.	*Poaceae*	Ed	L2F	10	2.7
PV033	*Amaranthus sp*	*Amaranthaceae*	Ed	L2F	10	2.5
PV034	*B. retusum*	*Poaceae*	Ed	L4F	8	2.0
PV035	*Q. coccifera*	*Fagaceae*	Oa	Q2F	17	3.4
PV036	*Q. ilex*	*Fagaceae*	Oa	Q2F	31	6.3
PV037	*Asparagus acutifolius* L.	*Asparagaceae*	Oa	Q2F	27	6.6
PV041	*Avenula bromoides* ((Gouan) H.Scholz)	*Poaceae*	Oa	Q3F	20	5.2
PV042	*Q. coccifera*	*Fagaceae*	Oa	Q3F	14	2.6
PV046	*C. bourgeanus*	*Asteraceae*	Cr	Z1V	7	2.5
PV047	*Zea mays* L.	*Poaceae*	Cr	Z2V	13	3.3
PV048	Desconocida 4	Desconocida	Cr	Z2V	9	6.0
PV049	*D. stramonium*	*Solanaceae*	Cr	Z1V	11	3.8
PV050	*Xanthium strumarium* L.	*Asteraceae*	Cr	Z1V	18	3.4
PV051	*Lotus corniculatus* L.	*Fabaceae*	Oa	Q4P	13	3.3
PV052	*Astragalus incanus* L.	*Fabaceae*	Oa	Q4P	10	2.5

[1]HTS pool code.
[2] number of RNA extracts from individual plant samples pooled at each library.

Table A3. Continued.

ID[1]	Species	Family	Habitat	Site of collection	RNA extracts pooled [2]	RIN
PV053	*A. incanus*	*Fabaceae*	Oa	Q4P	7	3.1
PV054	*Thymus vulgaris* L.	*Lamiaceae*	Oa	Q4P	8	7.5
PV055	*T. vulgaris*	*Lamiaceae*	Oa	Q4P	8	7.6
PV056	*Melilotus sp* L.	*Fabaceae*	Oa	Q4P	5	7.6
PV057	*Medicago orbicularis* (L.) Bartal.	*Fabaceae*	Wl	E4P	16	8.4
PV058	*Vicia sp* L.	*Fabaceae*	Wl	E4P	11	8.2
PV059	*Tragopogon sp* L.	*Asteraceae*	Wl	E4P	10	7.0
PV060	*Reseda lutea* L.	*Resedaceae*	Wl	E4P	5	2.6
PV061	*C. melo*	*Cucurbitaceae*	Cr	M5V	26	7.6
PV062	*C. melo*	*Cucurbitaceae*	Cr	M7V	46	8.4
PV063	*C. melo*	*Cucurbitaceae*	Cr	M8V	27	NA
PV064	*A. herba alta*	*Asteraceae*	Wl	E1F	21	NA
PV065	*Diplotaxis erucoides* DC.	*Brassicaceae*	Wl	E1F	19	NA
PV066	*C. bourgeanus*	*Asteraceae*	Wl	E1F	8	NA
PV067	*Geranium sp*	*Geraniaceae*	Wl	E1F	9	NA
PV068	*Leontodon sp*	*Asteraceae*	Wl	E1F	8	NA
PV069	*Plantago sp* L.	*Plantaginaceae*	Wl	E1F	11	NA
PV070	*Descurainia sophia* (L.) Webb ex Prantl	*Brassicaceae*	Wl	E1F	6	NA
PV071	*Salvia verbenaca* L.	*Lamiaceae*	Wl	E1F	10	NA
PV072	*T. officinale*	*Asteraceae*	Wl	E1F	8	NA
PV074	*Leontodon sp*	*Asteraceae*	Wl	E2F	11	NA
PV075	*T. officinale*	*Asteraceae*	Wl	E2F	6	NA
PV076	*V. sinuatum*	*Scrophulariaceae*	Wl	E2F	11	NA

[1]HTS pool code.
[2] number of RNA extracts from individual plant samples pooled at each library.

Table A3. Continued.

ID[1]	Species	Family	Habitat	Site of collection	RNA extracts pooled [2]	RIN
PV078	*Bromus sp* L.	*Poaceae*	Wl	E4F	8	2.3
PV079	*C. arvensis*	*Convolvulaceae*	Wl	E4F	8	5.4
PV080	*Odontites luteus* (L.) Clairv.	*Orobanchaceae*	Oa	Q4F	6	5.3
PV081	*L. arvense*	*Boraginaceae*	Ed	L3V	1	8.5
PV082	*Marrubium vulgare* L.	*Lamiaceae*	Oa	Q1F	14	5.8
PV083	*T. pseudochamaepitys*	*Lamiaceae*	Oa	Q1P	11	7.2
PV084	*Teucrium capitatum* L.	*Lamiaceae*	Oa	Q1P	10	7.7
PV085	*Avena sterilis* L.	*Poaceae*	Wl	E4P	27	2.7
PV086	*V. sinuatum*	*Scrophulariaceae*	Wl	E4P	19	8.1
PV087	*Echium vulgare* L.	*Boraginaceae*	Wl	E4P	7	5.5
PV088	*Andryala arenaria* (DC.) Boiss. & Reuter	*Asteraceae*	Oa	Q1P	9	8.2
PV089	*M. vulgare*	*Lamiaceae*	Oa	Q1P	9	7.0
PV090	*Dactylis glomerata* L.	*Poaceae*	Oa	Q4P	13	7.0
PV091	*Galium verum* L.	*Rubiaceae*	Oa	Q4P	8	6.9
PV092	Anchusa undulata L.	*Boraginaceae*	Wl	E1F	8	7.6
PV093	*J. fruticans*	*Oleaceae*	Oa	Q1P	8	7.6
PV094	*Hypericum pubescens* Boiss.	*Hypericaceae*	Oa	Q1P	6	4.9
PV095	*Diplotaxis virgata* (Cav.) DC.	*Brassicaceae*	Oa	Q1P	5	2.6
PV096	*M. vernale*	*Poaceae*	Ed	L2V	7	7.2
PV097	*C. arvensis*	*Convolvulaceae*	Ed	L2V	5	6.7
PV098	*P. echioides*	*Asteraceae*	Wl	E2F	11	4.2
PV099	*A. undulata*	*Boraginaceae*	Wl	E2F	13	5.5

[1]HTS pool code.
[2] number of RNA extracts from individual plant samples pooled at each library.

Table A3. Continued.

ID[1]	Species	Family	Habitat	Site of collection	RNA extracts pooled [2]	RIN
PV100	*Aphyllanthes monspeliensis* L.	*Asparagaceae*	Oa	Q4P	10	6.8
PV101	*A. acutifolius*	*Asparagaceae*	Oa	Q2P	8	3.9
PV102	*Asphodelus aestivus* BROT	*Liliaceae*	Oa	Q2P	9	4.3
PV103	*Centaurea melitensis* L.	*Asteraceae*	Oa	Q2P	7	4.3
PV104	*D. glomerata*	*Poaceae*	Oa	Q2P	8	6.9
PV105	*G. verum*	*Rubiaceae*	Oa	Q2P	5	6.0
PV106	*Helichrysum stoechas* (L.) Moench	*Asteraceae*	Oa	Q2P	5	6.9
PV107	*Q. coccifera*	*Fagaceae*	Oa	Q2P	4	7.1
PV108	*Q. ilex*	*Fagaceae*	Oa	Q2P	14	6.8
PV109	*R. lutea*	*Resedaceae*	Oa	Q2P	1	2.6
PV110	*Teucrium botrys* L.	*Lamiaceae*	Oa	Q2P	7	6.1
PV111	*T. pseudochamaepitys*	*Lamiaceae*	Oa	Q2P	1	6.5
PV112	*Centranthus calcitrapae* (L.) Dufr.	*Valerianaceae*	Oa	Q2P	8	7.1
PV113	*C. bourgeanus*	*Asteraceae*	Oa	Q2P	4	7.0
PV114	*Staehelina dubia* L.	*Asteraceae*	Oa	Q2P	7	4.1
PV115	*Bromus sp*	*Poaceae*	Ed	L1P	3	7.1
PV116	*Daucus sp* L.	*Apiaceae*	Ed	L1P	3	6.5
PV117	*D. erucoides*	*Brassicaceae*	Ed	L1P	4	7.2
PV118	Grass 1	*Poaceae*	Ed	L1P	5	6.8
PV119	*Lactuca serriola* L.	*Asteraceae*	Ed	L1P	2	7.6
PV120	*Leontodon sp*	*Asteraceae*	Ed	L1P	1	7.4
PV121	*Lepidium draba* (L.) Desv.	*Brassicaceae*	Ed	L1P	3	7.6
PV122	*P. echioides*	*Asteraceae*	Ed	L1P	2	7.1
PV123	*Rumex pulcher* L.	*Polygonaceae*	Ed	L1P	4	7.0

[1]HTS pool code.
[2] number of RNA extracts from individual plant samples pooled at each library.

Table A3. Continued.

ID[1]	Species	Family	Habitat	Site of collection	RNA extracts pooled [2]	RIN
PV124	*Malva sylvestris* L.	*Malvaceae*	Ed	L1P	3	6.2
PV125	*Senecio jacobaea* L.	*Asteraceae*	Ed	L1P	5	7.1
PV126	*Silybum marianum* (L.) Gaertner	*Asteraceae*	Ed	L1P	2	5.7
PV127	*Torilis nodosa* (L.) Gaertn.	*Apiaceae*	Ed	L1P	6	6.7
PV128	*Vicia sp*	*Fabaceae*	Ed	L1P	2	8.3
PV129	*Bromus sp*	*Poaceae*	Ed	L2P	3	7.1
PV130	*C. bourgeanus*	*Asteraceae*	Ed	L2P	3	6.7
PV131	*C. album*	*Amaranthaceae*	Ed	L2P	4	4.4
PV132	*C. arvensis*	*Convolvulaceae*	Ed	L2P	2	5.3
PV133	*Diplotaxis sp* DC.	*Brassicaceae*	Ed	L2P	6	5.4
PV134	*D. sophia*	*Brassicaceae*	Ed	L2P	2	7.3
PV135	*Fumaria parviflora* Lam.	*Papavaraceae*	Ed	L2P	4	6.6
PV136	*L. serriola*	*Asteraceae*	Ed	L2P	1	8.3
PV137	*Lolium perenne* L.	*Poaceae*	Ed	L2P	3	6.7
PV138	*Potentilla sp* L.	*Rosaceae*	Ed	L2P	2	3.6
PV139	*R.peregrina*	*Rubiaceae*	Ed	L2P	2	7.0
PV140	*S. marianum*	*Asteraceae*	Ed	L2P	3	4.4
PV141	*Bromus sp*	*Poaceae*	Ed	L3P	2	3.2
PV142	*C. arvensis*	*Convolvulaceae*	Ed	L3P	6	6.9
PV143	*D. erucoides*	*Brassicaceae*	Ed	L3P	13	3.4
PV144	*F. parviflora*	*Papavaraceae*	Ed	L3P	9	7.2
PV145	*L. draba*	*Brassicaceae*	Ed	L3P	1	4.1
PV146	*M. sylvestris*	*Malvaceae*	Ed	L3P	1	8.2

[1]HTS pool code.
[2] number of RNA extracts from individual plant samples pooled at each library.

Table A3. Continued.

ID[1]	Species	Family	Habitat	Site of collection	RNA extracts pooled [2]	RIN
PV147	*P. rhoeas*	*Papaveraceae*	Ed	L3P	1	4.2
PV148	*Scandix pecten-veneris* L.	*Apiaceae*	Ed	L3P	1	7.1
PV149	*S. marianum*	*Asteraceae*	Ed	L3P	2	6.8
PV150	*Anthriscus caucalis* Bieb	*Apiaceae*	Ed	L4P	7	4.6
PV151	*Aristolochia pistolochia* L.	*Aristolochiaceae*	Ed	L4P	3	6.9
PV152	*Brachypodium phoenicoides* (L.) Roemer & Schultes	*Poaceae*	Ed	L4P	10	7.1
PV153	*C. bourgeanus*	*Asteraceae*	Ed	L4P	6	7.1
PV154	*C. arvensis*	*Convolvulaceae*	Ed	L4P	5	7.2
PV155	*F. parviflora*	*Papavaraceae*	Ed	L4P	1	6.9
PV156	*Hirschfeldia incana* L.	*Brassicaceae*	Ed	L4P	1	7.7
PV157	*L. draba*	*Brassicaceae*	Ed	L4P	9	6.9
PV158	*R.peregrina*	*Rubiaceae*	Ed	L4P	1	7.7
PV159	*C. bourgeanus*	*Asteraceae*	Ed	L1V	2	5.5
PV160	*C. album*	*Amaranthaceae*	Ed	L1V	2	8.1
PV161	*C. arvensis*	*Convolvulaceae*	Ed	L1V	4	7.0
PV162	*Conyza canadensis* (L.) Cronquist	*Asteraceae*	Ed	L1V	2	6.8
PV163	*L. perenne*	*Poaceae*	Ed	L1V	3	2.3
PV164	*Sonchus oleraceus* L.	*Asteraceae*	Ed	L1V	2	7.4
PV165	*A. undulata*	*Boraginaceae*	Ed	L2V	5	8.1
PV166	*C. bourgeanus*	*Asteraceae*	Ed	L2V	1	2.5
PV167	*C. album*	*Amaranthaceae*	Ed	L2V	2	7.7
PV168	*C. canadensis*	*Asteraceae*	Ed	L2V	2	7.0
PV169	*G. verum*	*Rubiaceae*	Ed	L2V	1	7.3

[1]HTS pool code.
[2] number of RNA extracts from individual plant samples pooled at each library.

Table A3. Continued.

ID[1]	Species	Family	Habitat	Site of collection	RNA extracts pooled [2]	RIN
PV170	*L. serriola*	*Asteraceae*	Ed	L2V	2	7.1
PV171	*L. perenne*	*Poaceae*	Ed	L2V	2	4.1
PV172	*P. echioides*	*Asteraceae*	Ed	L2V	2	6.4
PV173	*R.peregrina*	*Rubiaceae*	Ed	L2V	3	7.7
PV174	*S. marianum*	*Asteraceae*	Ed	L2V	1	4.7
PV175	*Amaranthus sp*	*Amaranthaceae*	Ed	L3V	2	8.1
PV176	*Anacyclus clavatus* Desf. Pers.	*Asteraceae*	Ed	L3V	2	5.8
PV177	*Bassia scoparia* (L.) Voss	*Amaranthaceae*	Ed	L3V	2	6.5
PV178	*C. album*	*Amaranthaceae*	Ed	L3V	6	8.1
PV179	*C. canadensis*	*Asteraceae*	Ed	L3V	1	6.5
PV180	*Cynodon dactylon* (L.) Pers.	*Poaceae*	Ed	L3V	2	5.2
PV181	*D. erucoides*	*Brassicaceae*	Ed	L3V	2	7.0
PV182	*L. serriola*	*Asteraceae*	Ed	L3V	10	6.6
PV183	*P. echioides*	*Asteraceae*	Ed	L3V	1	6.9
PV184	*S. marianum*	*Asteraceae*	Ed	L3V	2	6.7
PV185	*S. nigrum*	*Solanaceae*	Ed	L3V	2	4.3
PV186	*S. oleraceae*	*Asteraceae*	Ed	L3V	1	7.8
PV187	*B. phoenicoides*	*Poaceae*	Ed	L2V	2	6.4
PV188	*C. arvensis*	*Convolvulaceae*	Ed	L3V	6	5.9
PV189	*Geranium sp*	*Geraniaceae*	Ed	L1P	1	4.3
PV190	*R.peregrina*	*Rubiaceae*	Ed	L1P	1	7.4
PV191	*Allium sativum* L.	*Amaryllidaceae*	Ed	L2P	1	4.3
PV192	*A. undulata*	*Boraginaceae*	Ed	L2P	1	7.9

[1]HTS pool code.
[2] number of RNA extracts from individual plant samples pooled at each library.

Table A3. Continued.

ID[1]	Species	Family	Habitat	Site of collection	RNA extracts pooled [2]	RIN
PV193	*G. verum*	*Rubiaceae*	Ed	L2P	1	6.9
PV194	*P. rhoeas*	*Papaveraceae*	Ed	L2P	1	4.7
PV195	*P. echioides*	*Asteraceae*	Ed	L2P	1	7.1
PV196	*Amaranthus sp*	*Amaranthaceae*	Ed	L1F	2	3.4
PV197	*A. caucalis*	*Apiaceae*	Ed	L1F	2	8.7
PV198	*B. retusum*	*Poaceae*	Ed	L1F	2	2.7
PV199	*C. bourgeanus*	*Asteraceae*	Ed	L1F	2	7.6
PV200	*C. album*	*Amaranthaceae*	Ed	L1F	2	4.9
PV201	*C. arvensis*	*Convolvulaceae*	Ed	L1F	8	6.8
PV202	*Conyza bonariensis* (L.) Cronquist	*Asteraceae*	Ed	L1F	6	4.9
PV203	*C. dactylon*	*Poaceae*	Ed	L1F	4	3.3
PV204	*D. erucoides*	*Brassicaceae*	Ed	L1F	3	2.5
PV205	*Geranium sp*	*Geraniaceae*	Ed	L1F	1	7.3
PV206	*L. serriola*	*Asteraceae*	Ed	L1F	2	4.3
PV207	*P. echioides*	*Asteraceae*	Ed	L1F	8	8.1
PV208	*R.peregrina*	*Rubiaceae*	Ed	L1F	2	3.3
PV209	*R. pulcher*	*Polygonaceae*	Ed	L1F	4	5.2
PV210	*C. dactylon*	*Poaceae*	Ed	L2F	2	2.3
PV211	*S. nigrum*	*Solanaceae*	Ed	L2F	1	1.9
PV212	*Amaranthus sp*	*Amaranthaceae*	Ed	L3F	5	2.5
PV213	*C. bourgeanus*	*Asteraceae*	Ed	L3F	2	4.8
PV214	*C. arvensis*	*Convolvulaceae*	Ed	L3F	10	3.2
PV215	*C. bonariensis*	*Asteraceae*	Ed	L3F	3	2.3
PV216	*C. dactylon*	*Poaceae*	Ed	L3F	7	3.4

[1]HTS pool code.
[2] number of RNA extracts from individual plant samples pooled at each library.

Table A3. Continued.

ID[1]	Species	Family	Habitat	Site of collection	RNA extracts pooled [2]	RIN
PV217	*D. stramonium*	*Solanaceae*	Ed	L3F	1	2.4
PV218	*Daucus sp*	*Apiaceae*	Ed	L3F	3	2.2
PV219	*D. erucoides*	*Brassicaceae*	Ed	L3F	8	2.4
PV220	*Eruca vesicaria* (L.) Cav.	*Brassicaceae*	Ed	L3F	3	2.4
PV221	*L. serriola*	*Asteraceae*	Ed	L3F	2	2.4
PV222	*M. sylvestris*	*Malvaceae*	Ed	L1F	2	2.5
PV223	*S. marianum*	*Asteraceae*	Ed	L3F	4	6.7
PV224	*Vicia sp*	*Fabaceae*	Ed	L3F	1	2.3
PV225	*Cirsium arvense* L.	*Asteraceae*	Ed	L4F	4	6.8
PV226	*C. arvensis*	*Convolvulaceae*	Ed	L4F	1	7.3
PV227	*Daucus sp*	*Apiaceae*	Ed	L4F	4	5.1
PV228	Desconocida 4	Desconocida	Ed	L4F	1	7.5
PV229	*D. erucoides*	*Brassicaceae*	Ed	L4F	2	5.7
PV230	*E. vulgare*	*Boraginaceae*	Ed	L4F	1	7.1
PV231	*L. draba*	*Brassicaceae*	Ed	L4F	6	7.1
PV232	*L. perenne*	*Poaceae*	Ed	L4F	1	2.5
PV233	*Phalaris minor* Retz.	*Poaceae*	Ed	L4F	2	4.1
PV234	*R. pulcher*	*Polygonaceae*	Ed	L4F	2	7.5
PV235	*S. marianum*	*Asteraceae*	Ed	L4F	3	6.9
PV236	*Trifolium campestre* Schreb.	*Fabaceae*	Ed	L4F	1	5.1
PV237	*Erodium cicutarium* (L.) L'Hérit. ex Alton	*Geraniaceae*	Ed	L4F	2	8.8
PV252	*R. peregrina*	*Rubiaceae*	Wl	E2F	1	8.0
PV257	Desconocida 5	Desconocida	Wl	E3F	9	8.3

[1]HTS pool code.
[2] number of RNA extracts from individual plant samples pooled at each library.

Table A3. Continued.

ID[1]	Species	Family	Habitat	Site of collection	RNA extracts pooled [2]	RIN
PV262	*G. verum*	*Rubiaceae*	Wl	E3F	9	3.6
PV268	*Sisymbrium runcinatum* Lag.	*Brassicaceae*	Wl	E3F	1	7.7
PV269	*Teucrium chamaedrys* L.	*Lamiaceae*	Wl	E3F	3	8.4
PV270	*A. acutifolius*	*Asparagaceae*	Wl	E4F	1	4.4
PV277	A. clavatus	*Asteraceae*	Wl	E1P	5	8.3
PV296	*S. parviflora*	*Poaceae*	Wl	E1P	6	8.1
PV298	*T. capitatum*	*Lamiaceae*	Wl	E1P	15	7.8
PV306	Asteriscus aquaticus (L.) LESS	*Asteraceae*	Wl	E2P	3	5.7
PV312	*C. melitensis*	*Asteraceae*	Wl	E2P	7	8.3
PV313	*C. calcitrapae*	*Valerianaceae*	Wl	E2P	1	8.6
PV330	Phlomis lychnitis L.	*Lamiaceae*	Wl	E2P	1	7.7
PV350	*E. cicutarium*	*Geraniaceae*	Wl	E3P	3	9.2
PV352	Eryngium campestre L.	*Apiaceae*	Wl	E3P	7	8.6
PV356	*Bromus madritensis* L.	*Poaceae*	Wl	E3P	13	7.3
PV359	Medicago sp L.	*Fabaceae*	Wl	E3P	6	7.8
PV369	*T. pseudochamaepitys*	*Lamiaceae*	Wl	E3P	3	8.7
PV380	*D. glomerata*	*Poaceae*	Wl	E4P	6	7.6
PV386	*P. rhoeas*	*Papaveraceae*	Wl	E4P	7	8.7
PV390	Thapsia villosa L.	*Apiaceae*	Wl	E4P	4	6.3
PV418	*A. pistolochia*	*Aristolochiaceae*	Oa	Q1P	1	4.1
PV421	*A. aquaticus*	*Asteraceae*	Oa	Q1P	4	8.3
PV423	*Bromus sp*	*Poaceae*	Oa	Q1P	6	6.9
PV426	*E. cicutarium*	*Geraniaceae*	Oa	Q1P	3	8.2
PV428	*B. madritensis*	*Poaceae*	Oa	Q1P	2	8.1

[1]HTS pool code.
[2] number of RNA extracts from individual plant samples pooled at each library.

Table A3. Continued.

ID[1]	Species	Family	Habitat	Site of collection	RNA extracts pooled [2]	RIN
PV430	*P. rhoeas*	*Papaveraceae*	Oa	Q1P	1	8.1
PV437	*Geranium sp*	*Geraniaceae*	Oa	Q2P	11	8.8
PV440	*T. villosa*	*Apiaceae*	Oa	Q2P	3	5.7
PV441	*Leontodon sp*	*Asteraceae*	Oa	Q2P	5	8.0
PV454	*Klasea pinnatifida* (Cav.) Cass.	*Asteraceae*	Oa	Q3P	11	8.2
PV457	*Medicago sp*	*Fabaceae*	Oa	Q3P	1	8.1
PV464	*R. peregrina*	*Rubiaceae*	Oa	Q3P	5	7.9
PV473	*E. campestre*	*Apiaceae*	Oa	Q4P	3	8.4
PV495	A. clavatus	*Asteraceae*	Cr	H2P	4	6.7
PV498	*R. peregrina*	*Rubiaceae*	Cr	H3P	9	8.0
PV500	Desconocida 5	Desconocida	Cr	H3P	4	3.6
PV501	*Bromus sp*	*Poaceae*	Cr	H3P	1	7.2
PV514	*D. erucoides*	*Brassicaceae*	Cr	M3V	3	2.9
PV516	*P. echioides*	*Asteraceae*	Cr	M3V	1	8.0
PV519	Desconocida 4	Desconocida	Cr	M4V	4	8.0
PV544	*S. runcinatum*	*Brassicaceae*	Cr	C1F	4	4.6
PV546	*K. pinnatifida*	*Asteraceae*	Cr	C2F	1	7.8
PV557	*E. cicutarium*	*Geraniaceae*	Cr	C2F	1	7.8
PV558	*P. lychnitis*	*Lamiaceae*	Oa	Q2P	1	8.1
PV559	*S. parviflora*	*Poaceae*	Oa	Q1P	1	6.5
PV560	*T. chamaedrys*	*Lamiaceae*	Oa	Q2P	1	8.6
PV561	*Dittrichia viscosa* (L.) Greuter	*Asteraceae*	Wl	E2F/E2P	8	8.8
PV562	*Foeniculum vulgare* Mill.	*Apiaceae*	Wl	E3F/E3P	8	8.6
PV563	*F. vulgare*	*Apiaceae*	Wl	E2P	6	4.4

[1]HTS pool code.
[2] number of RNA extracts from individual plant samples pooled at each library.

Table A3. Continued.

ID[1]	Species	Family	Habitat	Site of collection	RNA extracts pooled [2]	RIN
PV564	*Phlomis purpurea* L.	*Lamiaceae*	Wl	E2P	8	7.3
PV565	*P. purpurea*	*Lamiaceae*	Wl	E3P	4	6.8
PV566	*Retama sphaerocarpa* L.	*Fabaceae*	Wl	E2F/E2P	7	8.0
PV567	*R. pulcher*	*Polygonaceae*	Wl	E2P	4	8.5
PV568	*R. pulcher*	*Polygonaceae*	Wl	E3F/E3P	5	8.4
PV569	*S. marianum*	*Asteraceae*	Wl	E2F/E2P	8	8.2
PV570	*Crataegus monogyna* Jacq.	*Rosaceae*	Oa	Q3F/Q3P/Q4P	3	3.2
PV571	*Bupleurum rigidum* L.	*Apiaceae*	Oa	Q3F/Q3P	8	8.3
PV572	*Trifolium tomentosum* L.	*Fabaceae*	Wl	E1P	3	4.6
PV573	T. tomentosum L.	*Fabaceae*	Wl	E2P	4	7.6
PV574	*Helianthemum cinereum* (Cav.) Pers.	*Cistaceae*	Wl	Q3P	8	7.9
PV575	*Lavandula latifolia* Medik.	*Lamiaceae*	Wl	Q3F	3	4.8
PV576	*L. latifolia*	*Lamiaceae*	Wl	Q4F	3	8.3
PV577	*Lithodora fruticosa* (L.) Griseb.	*Boraginaceae*	Oa	Q3F/Q3P	3	6.1
PV578	*Origanum vulgare* L.	*Labiatae*	Oa	Q1P/Q2P/Q3P	4	8.3
PV579	*Rhamnus lycioides* L.	*Rhamnaceae*	Oa	Q1F	8	8.1
PV580	*R. lycioides*	*Rhamnaceae*	Oa	Q2F	5	7.9
PV581	*R. lycioides*	*Rhamnaceae*	Oa	Q2P	5	8.6
PV582	*R. lycioides*	*Rhamnaceae*	Oa	Q3F/Q3P	6	8.1
PV583	*Astragalus sesameus* L.	*Fabaceae*	Wl	E1P	6	3.0
PV584	*A. sesameus*	*Fabaceae*	Wl	E3P	5	8.6
PV585	*Borago officinalis* L.	*Boraginaceae*	Wl	E2P	6	6.7
PV586	*F. parviflora*	*Papavaraceae*	Cr	H1P	8	7.7
PV587	*H. incana*	*Brassicaceae*	Cr	H1P	8	6.7

[1]HTS pool code.
[2] number of RNA extracts from individual plant samples pooled at each library.

Table A3. Continued.

ID[1]	Species	Family	Habitat	Site of collection	RNA extracts pooled [2]	RIN
PV588	*Hordeum vulgare* L.	*Poaceae*	Cr	H1P	8	6.0
PV589	*H. vulgare*	*Poaceae*	Cr	H1P	8	5.6

[1]HTS pool code.
[2] number of RNA extracts from individual plant samples pooled at each library.

Table A4.1. List of virus OTUs detected by HTS presented according to their number of detections in the ecosystem.

Habitat	Virus	Detections[1]
Cr/Ed/Oa/Wl	*Cucumber mosaic virus*	188/306 (61%)
Cr/Ed/Oa/Wl	*Pelargonium zonate spot virus*	143/306 (47%)
Cr/Ed/Oa/Wl	*Tobacco mild green mosaic virus*	110/306 (36%)
Cr/Ed/Oa/Wl	*Rubus chlorotic mottle virus*	103/306 (34%)
Cr/Ed/Oa/Wl	*Tobacco mosaic virus*	81/306 (26%)
Cr/Ed/Oa/Wl	*Watermelon mosaic virus*	66/306 (22%)
Cr/Ed/Oa/Wl	*Tomato aspermy virus*	59/306 (19%)
Cr/Ed/Oa/Wl	*Pepper mild mottle virus*	54/306 (18%)
Cr/Ed/Oa/Wl	*Turnip yellows virus*	51/306 (17%)
Cr/Ed/Oa/Wl	*Brassica yellows virus*	37/306 (12%)
Cr/Ed/Oa/Wl	*Youcai mosaic virus*	28/306 (9%)
Cr/Ed/Oa/Wl	*Tomato mosaic virus*	22/306 (7%)
Cr/Ed/Oa/Wl	*Rehmannia mosaic virus*	20/306 (7%)
Cr/Ed/Oa/Wl	*Turnip mosaic virus*	18/306 (6%)
Cr/Ed/Oa/Wl	*Tomato brown rugose fruit virus*	14/306 (5%)
Cr/Ed/Oa/Wl	*Plum pox virus*	10/306 (3%)
Cr/Ed/Oa/Wl	*Tomato mottle mosaic virus*	9/306 (3%)
Cr/Ed/Oa	*Beet western yellows virus*	45/306 (15%)
Cr/Ed/Oa	*Beet mild yellowing virus*	38/306 (12%)
Cr/Ed/Oa	*Beet chlorosis virus*	35/306 (11%)
Cr/Ed/Oa	*Gayfeather mild mottle virus*	33/306 (11%)
Cr/Ed/Wl	*Barley yellow dwarf virus*	10/306 (3%)
Ed/Oa/Wl	*Turnip crinkle virus*	30/306 (10%)
Ed/Oa/Wl	*Wasabi mottle virus*	9/306 (3%)
Cr/Ed	*Soybean mosaic virus*	9/306 (3%)
Cr/Ed	*Cereal yellow dwarf virus*	3/306 (1%)
Cr/Ed	*Cucumber leaf spot virus*	3/306 (1%)
Cr/Ed	*Pothos latent virus*	3/306 (1%)
Cr/Ed	*Wheat yellow dwarf virus*	3/306 (1%)
Cr/Ed	*Zucchini yellow mosaic virus*	2/306 (1%)
Ed/Oa	*Maize dwarf mosaic virus*	31/306 (10%)
Ed/Oa	*Cucumber green mottle mosaic virus*	2/306 (1%)
Ed/Wl	*Cucurbit aphid-borne yellows virus*	8/306 (3%)
Ed/Wl	*Brome mosaic virus*	3/306 (1%)
Ed/Wl	*Cardamine chlorotic fleck virus*	3/306 (1%)
Ed/Wl	*Ixeridium yellow mottle virus 1*	3/306 (1%)
Ed/Wl	*Pelargonium leaf curl virus*	3/306 (1%)
Ed/Wl	*Pelargonium necrotic spot virus*	3/306 (1%)
Ed/Wl	*Tomato bushy stunt virus*	3/306 (1%)
Ed/Wl	*Artichoke mottled crinkle virus*	2/306 (1%)
Ed/Wl	*Carnation Italian ringspot virus*	2/306 (1%)
Ed/Wl	*Cucumber necrosis virus*	2/306 (1%)

[1]Detections are the proportion of libraries where virus OTUs were detected.

Table A4.1. Continued.

Habitat	Virus	Detections[1]
Ed/Wl	*Eggplant mottled crinkle virus*	2/306 (1%)
Ed/Wl	*Grapevine Algerian latent virus*	2/306 (1%)
Ed/Wl	*Ixeridium yellow mottle virus 2*	2/306 (1%)
Ed/Wl	*Moroccan pepper virus*	2/306 (1%)
Ed/Wl	*Turnip yellow mosaic virus*	2/306 (1%)
Oa/Wl	*Turnip vein-clearing virus*	6/306 (2%)
Oa/Wl	*Ribgrass mosaic virus*	5/306 (2%)
Oa/Wl	*Alfalfa mosaic virus*	3/306 (1%)
Oa/Wl	*Eggplant mottled dwarf virus*	3/306 (1%)
Cr	*Calla lily latent virus*	6/306 (2%)
Cr	*Barley virus G*	1/306 (0%)
Cr	*Wheat streak mosaic virus*	1/306 (0%)
Ed	*Potato leafroll virus*	21/306 (7%)
Ed	*Tobacco virus 2*	21/306 (7%)
Ed	*Parietaria mottle virus*	18/306 (6%)
Ed	*Melon aphid-borne yellows virus*	9/306 (3%)
Ed	*Lettuce yellow mottle virus*	6/306 (2%)
Ed	*Pepo aphid-borne yellows virus*	6/306 (2%)
Ed	*Suakwa aphid-borne yellows virus*	4/306 (1%)
Ed	*Chickpea chlorotic stunt virus*	3/306 (1%)
Ed	*Sorghum mosaic virus*	3/306 (1%)
Ed	*Bell pepper mottle virus*	2/306 (1%)
Ed	*Carrot red leaf virus*	2/306 (1%)
Ed	*Cotton leafroll dwarf virus*	2/306 (1%)
Ed	*Cowpea polerovirus 1*	2/306 (1%)
Ed	*Tobacco necrosis virus A*	2/306 (1%)
Ed	*Tropical soda apple mosaic virus*	2/306 (1%)
Ed	*Yam spherical virus*	2/306 (1%)
Ed	*Barley yellow striate mosaic virus*	1/306 (0%)
Ed	*Beet black scorch virus*	1/306 (0%)
Ed	*Iranian johnsongrass mosaic virus*	1/306 (0%)
Ed	*Lettuce big-vein associated virus*	1/306 (0%)
Ed	*Olive latent virus 1*	1/306 (0%)
Ed	*Olive mild mosaic virus*	1/306 (0%)
Ed	*Potato necrosis virus*	1/306 (0%)
Ed	*Potato virus Y*	1/306 (0%)
Ed	*Sugarcane yellow leaf virus*	1/306 (0%)
Ed	*Tobacco necrosis virus D*	1/306 (0%)
Ed	*Wheat leaf yellowing-associated virus*	1/306 (0%)
Oa	*Prunus necrotic ringspot virus*	3/306 (1%)
Oa	*Obuda pepper virus*	2/306 (1%)
Oa	*Paprika mild mottle virus*	2/306 (1%)
Oa	*Bermuda grass latent virus*	1/306 (0%)
Oa	*Dulcamara mottle virus*	1/306 (0%)

[1]Detections are the proportion of libraries where virus OTUs were detected.

Table A4.1. Continued.

Habitat	Virus	Detections[1]
Wl	*Cymbidium ringspot virus*	1/306 (0%)
Wl	*Havel river virus*	1/306 (0%)
Wl	*Petunia asteroid mosaic virus*	1/306 (0%)
Wl	*Turnip rosette virus*	1/306 (0%)

[1]Detections are the proportion of libraries where virus OTUs were detected.

Table A4.2. List of virus OTUs detected by HTS presented according to their number of detections in Crop habitat.

Habitat	Virus	Detections[1]
Cr	*Cucumber mosaic virus*	19/37 (51%)
Cr	*Watermelon mosaic virus*	14/37 (38%)
Cr	*Rubus chlorotic mottle virus*	9/37 (24%)
Cr	*Soybean mosaic virus*	8/37 (22%)
Cr	*Tobacco mild green mosaic virus*	8/37 (22%)
Cr	*Calla lily latent virus*	6/37 (16%)
Cr	*Pepper mild mottle virus*	6/37 (16%)
Cr	*Gayfeather mild mottle virus*	5/37 (14%)
Cr	*Pelargonium zonate spot virus*	5/37 (14%)
Cr	*Tomato aspermy virus*	5/37 (14%)
Cr	*Beet mild yellowing virus*	4/37 (11%)
Cr	*Brassica yellows virus*	4/37 (11%)
Cr	*Barley yellow dwarf virus*	4/37 (11%)
Cr	*Turnip yellows virus*	4/37 (11%)
Cr	*Beet chlorosis virus*	3/37 (8%)
Cr	*Beet western yellows virus*	3/37 (8%)
Cr	*Tobacco mosaic virus*	3/37 (8%)
Cr	*Turnip mosaic virus*	2/37 (5%)
Cr	*Barley virus G*	1/37 (3%)
Cr	*Cucumber leaf spot virus*	1/37 (3%)
Cr	*Cereal yellow dwarf virus*	1/37 (3%)
Cr	*Pothos latent virus*	1/37 (3%)
Cr	*Plum pox virus*	1/37 (3%)
Cr	*Rehmannia mosaic virus*	1/37 (3%)
Cr	*Tomato brown rugose fruit virus*	1/37 (3%)
Cr	*Tomato mottle mosaic virus*	1/37 (3%)
Cr	*Tomato mosaic virus*	1/37 (3%)
Cr	*Wheat streak mosaic virus*	1/37 (3%)
Cr	*Wheat yellow dwarf virus*	1/37 (3%)
Cr	*Youcai mosaic virus*	1/37 (3%)
Cr	*Zucchini yellow mosaic virus*	1/37 (3%)

[1] Detections are the proportion of libraries where virus OTUs were detected.

Table A4.3. List of virus OTUs detected by HTS presented according to their number of detections in Edge habitat.

Habitat	Virus	Detections[1]
Ed	*Pelargonium zonate spot virus*	124/133 (93%)
Ed	*Cucumber mosaic virus*	107/133 (80%)
Ed	*Rubus chlorotic mottle virus*	85/133 (64%)
Ed	*Tomato aspermy virus*	46/133 (35%)
Ed	*Turnip yellows virus*	44/133 (33%)
Ed	*Tobacco mild green mosaic virus*	43/133 (32%)
Ed	*Beet western yellows virus*	40/133 (30%)
Ed	*Tobacco mosaic virus*	35/133 (26%)
Ed	*Pepper mild mottle virus*	33/133 (25%)
Ed	*Beet mild yellowing virus*	32/133 (24%)
Ed	*Brassica yellows virus*	31/133 (23%)
Ed	*Beet chlorosis virus*	30/133 (23%)
Ed	*Maize dwarf mosaic virus*	30/133 (23%)
Ed	*Watermelon mosaic virus*	28/133 (21%)
Ed	*Gayfeather mild mottle virus*	25/133 (19%)
Ed	*Potato leafroll virus*	21/133 (16%)
Ed	*Tobacco virus 2*	21/133 (16%)
Ed	*Parietaria mottle virus*	18/133 (14%)
Ed	*Tomato mosaic virus*	16/133 (12%)
Ed	*Melon aphid-borne yellows virus*	9/133 (7%)
Ed	*Rehmannia mosaic virus*	9/133 (7%)
Ed	*Turnip crinkle virus*	8/133 (6%)
Ed	*Turnip mosaic virus*	8/133 (6%)
Ed	*Cucurbit aphid-borne yellows virus*	7/133 (5%)
Ed	*Lettuce yellow mottle virus*	6/133 (5%)
Ed	*Pepo aphid-borne yellows virus*	6/133 (5%)
Ed	*Plum pox virus*	6/133 (5%)
Ed	*Tomato brown rugose fruit virus*	6/133 (5%)
Ed	*Tomato mottle mosaic virus*	6/133 (5%)
Ed	*Barley yellow dwarf virus*	5/133 (4%)
Ed	*Youcai mosaic virus*	5/133 (4%)
Ed	*Suakwa aphid-borne yellows virus*	4/133 (3%)
Ed	*Chickpea chlorotic stunt virus*	3/133 (2%)
Ed	*Sorghum mosaic virus*	3/133 (2%)
Ed	*Bell pepper mottle virus*	2/133 (2%)
Ed	*Cardamine chlorotic fleck virus*	2/133 (2%)
Ed	*Cotton leafroll dwarf virus*	2/133 (2%)
Ed	*Cucumber leaf spot virus*	2/133 (2%)
Ed	*Cowpea polerovirus 1*	2/133 (2%)
Ed	*Carrot red leaf virus*	2/133 (2%)
Ed	*Cereal yellow dwarf virus*	2/133 (2%)
Ed	*Ixeridium yellow mottle virus 1*	2/133 (2%)
Ed	*Pothos latent virus*	2/133 (2%)

[1] Detections are the proportion of libraries where virus OTUs were detected.

Table A4.3. Continued.

Habitat	Virus	Detections[1]
Ed	*Tobacco necrosis virus A*	2/133 (2%)
Ed	*Tropical soda apple mosaic virus*	2/133 (2%)
Ed	*Wheat yellow dwarf virus*	2/133 (2%)
Ed	*Yam spherical virus*	2/133 (2%)
Ed	*Artichoke mottled crinkle virus*	1/133 (1%)
Ed	*Beet black scorch virus*	1/133 (1%)
Ed	*Brome mosaic virus*	1/133 (1%)
Ed	*Barley yellow striate mosaic virus*	1/133 (1%)
Ed	*Cucumber green mottle mosaic virus*	1/133 (1%)
Ed	*Carnation Italian ringspot virus*	1/133 (1%)
Ed	*Cucumber necrosis virus*	1/133 (1%)
Ed	*Eggplant mottled crinkle virus*	1/133 (1%)
Ed	*Grapevine Algerian latent virus*	1/133 (1%)
Ed	*Iranian johnsongrass mosaic virus*	1/133 (1%)
Ed	*Ixeridium yellow mottle virus 2*	1/133 (1%)
Ed	*Lettuce big-vein associated virus*	1/133 (1%)
Ed	*Moroccan pepper virus*	1/133 (1%)
Ed	*Olive latent virus 1*	1/133 (1%)
Ed	*Olive mild mosaic virus*	1/133 (1%)
Ed	*Pelargonium leaf curl virus*	1/133 (1%)
Ed	*Pelargonium necrotic spot virus*	1/133 (1%)
Ed	*Potato necrosis virus*	1/133 (1%)
Ed	*Potato virus Y*	1/133 (1%)
Ed	*Soybean mosaic virus*	1/133 (1%)
Ed	*Sugarcane yellow leaf virus*	1/133 (1%)
Ed	*Tomato bushy stunt virus*	1/133 (1%)
Ed	*Tobacco necrosis virus D*	1/133 (1%)
Ed	*Turnip yellow mosaic virus*	1/133 (1%)
Ed	*Wheat leaf yellowing-associated virus*	1/133 (1%)
Ed	*Wasabi mottle virus*	1/133 (1%)
Ed	*Zucchini yellow mosaic virus*	1/133 (1%)

[1] Detections are the proportion of libraries where virus OTUs were detected.

Table A4.4 List of virus OTUs detected by HTS presented according to their number of detections in Oakwood habitat.

Habitat	Virus	Detections[1]
Oa	*Cucumber mosaic virus*	40/71 (56%)
Oa	*Tobacco mild green mosaic virus*	34/71 (48%)
Oa	*Tobacco mosaic virus*	25/71 (35%)
Oa	*Watermelon mosaic virus*	20/71 (28%)
Oa	*Turnip crinkle virus*	14/71 (20%)
Oa	*Pepper mild mottle virus*	11/71 (15%)
Oa	*Youcai mosaic virus*	10/71 (14%)
Oa	*Pelargonium zonate spot virus*	7/71 (10%)
Oa	*Rehmannia mosaic virus*	7/71 (10%)
Oa	*Tomato aspermy virus*	7/71 (10%)
Oa	*Tomato brown rugose fruit virus*	6/71 (8%)
Oa	*Rubus chlorotic mottle virus*	4/71 (6%)
Oa	*Wasabi mottle virus*	4/71 (6%)
Oa	*Gayfeather mild mottle virus*	3/71 (4%)
Oa	*Prunus necrotic ringspot virus*	3/71 (4%)
Oa	*Tomato mosaic virus*	3/71 (4%)
Oa	*Turnip mosaic virus*	3/71 (4%)
Oa	*Turnip vein-clearing virus*	3/71 (4%)
Oa	*Beet chlorosis virus*	2/71 (3%)
Oa	*Beet mild yellowing virus*	2/71 (3%)
Oa	*Beet western yellows virus*	2/71 (3%)
Oa	*Obuda pepper virus*	2/71 (3%)
Oa	*Paprika mild mottle virus*	2/71 (3%)
Oa	*Plum pox virus*	2/71 (3%)
Oa	*Ribgrass mosaic virus*	2/71 (3%)
Oa	*Turnip yellows virus*	2/71 (3%)
Oa	*Alfalfa mosaic virus*	1/71 (1%)
Oa	*Bermuda grass latent virus*	1/71 (1%)
Oa	*Brassica yellows virus*	1/71 (1%)
Oa	*Cucumber green mottle mosaic virus*	1/71 (1%)
Oa	*Dulcamara mottle virus*	1/71 (1%)
Oa	*Eggplant mottled dwarf virus*	1/71 (1%)
Oa	*Maize dwarf mosaic virus*	1/71 (1%)
Oa	*Tomato mottle mosaic virus*	1/71 (1%)

[1] Detections are the proportion of libraries where virus OTUs were detected.

Table A4.5 List of virus OTUs detected by HTS presented according to their number of detections in Wasteland habitat.

Habitat	Virus taxon	Detections[1]
WI	*Tobacco mild green mosaic virus*	25/65 (38%)
WI	*Cucumber mosaic virus*	22/65 (34%)
WI	*Tobacco mosaic virus*	18/65 (28%)
WI	*Youcai mosaic virus*	12/65 (18%)
WI	*Turnip crinkle virus*	8/65 (12%)
WI	*Pelargonium zonate spot virus*	7/65 (11%)
WI	*Rubus chlorotic mottle virus*	5/65 (8%)
WI	*Turnip mosaic virus*	5/65 (8%)
WI	*Pepper mild mottle virus*	4/65 (6%)
WI	*Wasabi mottle virus*	4/65 (6%)
WI	*Watermelon mosaic virus*	4/65 (6%)
WI	*Rehmannia mosaic virus*	3/65 (5%)
WI	*Ribgrass mosaic virus*	3/65 (5%)
WI	*Turnip vein-clearing virus*	3/65 (5%)
WI	*Alfalfa mosaic virus*	2/65 (3%)
WI	*Brome mosaic virus*	2/65 (3%)
WI	*Eggplant mottled dwarf virus*	2/65 (3%)
WI	*Pelargonium leaf curl virus*	2/65 (3%)
WI	*Pelargonium necrotic spot virus*	2/65 (3%)
WI	*Tomato bushy stunt virus*	2/65 (3%)
WI	*Tomato mosaic virus*	2/65 (3%)
WI	*Artichoke mottled crinkle virus*	1/65 (2%)
WI	*Brassica yellows virus*	1/65 (2%)
WI	*Barley yellow dwarf virus-PAV*	1/65 (2%)
WI	*Cucurbit aphid-borne yellows virus*	1/65 (2%)
WI	*Cardamine chlorotic fleck virus*	1/65 (2%)
WI	*Carnation Italian ringspot virus*	1/65 (2%)
WI	*Cucumber necrosis virus*	1/65 (2%)
WI	*Cymbidium ringspot virus*	1/65 (2%)
WI	*Eggplant mottled crinkle virus*	1/65 (2%)
WI	*Grapevine Algerian latent virus*	1/65 (2%)
WI	*Havel river virus*	1/65 (2%)
WI	*Ixeridium yellow mottle virus 1*	1/65 (2%)
WI	*Ixeridium yellow mottle virus 2*	1/65 (2%)
WI	*Moroccan pepper virus*	1/65 (2%)
WI	*Petunia asteroid mosaic virus*	1/65 (2%)
WI	*Plum pox virus*	1/65 (2%)
WI	*Tomato aspermy virus*	1/65 (2%)
WI	*Tomato brown rugose fruit virus*	1/65 (2%)
WI	*Tomato mottle mosaic virus*	1/65 (2%)
WI	*Turnip rosette virus*	1/65 (2%)

[1] Detections are the proportion of libraries where virus OTUs were detected.

Table A4.6. List of virus genera detected by HTS presented according to their number of detections in the ecosystem.

Habitat	Virus	Detections[1]
Cr/Ed/Oa/Wl	*Cucumovirus*	192/306 (63%)
Cr/Ed/Oa/Wl	*Tobamovirus*	168/306 (55%)
Cr/Ed/Oa/Wl	*Anulavirus*	143/306 (47%)
Cr/Ed/Oa/Wl	*Potyvirus*	107/306 (35%)
Cr/Ed/Oa/Wl	*Sobemovirus*	103/306 (34%)
Cr/Ed/Oa/Wl	*Polerovirus*	91/306 (30%)
Cr/Ed/Wl	*Luteovirus*	10/306 (3%)
Ed/Oa/Wl	*Carmovirus*	30/306 (10%)
Ed/Oa/Wl	*Tymovirus*	3/306 (1%)
Cr/Ed	*Aureusvirus*	3/306 (1%)
Ed/Wl	*Tombusvirus*	4/306 (1%)
Ed/Wl	*Betacarmovirus*	3/306 (1%)
Ed/Wl	*Bromovirus*	3/306 (1%)
Oa/Ed	*Ilarvirus*	21/306 (7%)
Oa/Wl	*Alfamovirus*	3/306 (1%)
Oa/Wl	*Nucleorhabdovirus*	3/306 (1%)
Cr	*Tritimovirus*	1/306 (0%)
Ed	*Varicosavirus*	6/306 (2%)
Ed	*Necrovirus*	3/306 (1%)
Ed	*Alphanecrovirus*	1/306 (0%)
Ed	*Cytorhabdovirus*	1/306 (0%)
Oa	*Panicovirus*	1/306 (0%)

[1] Detections are the proportion of libraries where virus OTUs were detected.

Table A4.7. List of virus genera detected by HTS presented according to their number of detections in Crop habitat.

Habitat	Virus	Detections[1]
Cr	*Cucumovirus*	19/37 (51%)
Cr	*Potyvirus*	17/37 (46%)
Cr	*Tobamovirus*	12/37 (32%)
Cr	*Sobemovirus*	9/37 (24%)
Cr	*Polerovirus*	6/37 (16%)
Cr	*Anulavirus*	5/37 (14%)
Cr	*Luteovirus*	4/37 (11%)
Cr	*Aureusvirus*	1/37 (3%)
Cr	*Tritimovirus*	1/37 (3%)

[1] Detections are the proportion of libraries where virus OTUs were detected.

Table A4.8. List of virus genera detected by HTS presented according to their number of detections in Edge habitat.

Habitat	Virus	Detections[1]
Ed	*Anulavirus*	124/133 (93%)
Ed	*Cucumovirus*	111/133 (83%)
Ed	*Sobemovirus*	85/133 (64%)
Ed	*Polerovirus*	80/133 (60%)
Ed	*Tobamovirus*	73/133 (55%)
Ed	*Potyvirus*	58/133 (44%)
Ed	*Ilarvirus*	18/133 (14%)
Ed	*Carmovirus*	8/133 (6%)
Ed	*Varicosavirus*	6/133 (5%)
Ed	*Luteovirus*	5/133 (4%)
Ed	*Necrovirus*	3/133 (2%)
Ed	*Aureusvirus*	2/133 (2%)
Ed	*Betacarmovirus*	2/133 (2%)
Ed	*Alphanecrovirus*	1/133 (1%)
Ed	*Bromovirus*	1/133 (1%)
Ed	*Cytorhabdovirus*	1/133 (1%)
Ed	*Tombusvirus*	1/133 (1%)
Ed	*Tymovirus*	1/133 (1%)

[1] Detections are the proportion of libraries where virus OTUs were detected.

Table A4.9. List of virus genera detected by HTS presented according to their number of detections in Oakwood habitat.

Habitat	Virus	Detections[1]
Oa	*Tobamovirus*	46/71 (65%)
Oa	*Cucumovirus*	40/71 (56%)
Oa	*Potyvirus*	23/71 (32%)
Oa	*Carmovirus*	14/71 (20%)
Oa	*Anulavirus*	7/71 (10%)
Oa	*Sobemovirus*	4/71 (6%)
Oa	*Ilarvirus*	3/71 (4%)
Oa	*Polerovirus*	2/71 (3%)
Oa	*Alfamovirus*	1/71 (1%)
Oa	*Nucleorhabdovirus*	1/71 (1%)
Oa	*Panicovirus*	1/71 (1%)
Oa	*Tymovirus*	1/71 (1%)

[1] Detections are the proportion of libraries where virus OTUs were detected.

Table A4.10. List of virus genera detected by HTS presented according to their number of detections in Wasteland habitat.

Habitat	Virus	Detections[1]
Wl	*Tobamovirus*	37/65 (57%)
Wl	*Cucumovirus*	22/65 (34%)
Wl	*Potyvirus*	9/65 (14%)
Wl	*Carmovirus*	8/65 (12%)
Wl	*Anulavirus*	7/65 (11%)
Wl	*Sobemovirus*	5/65 (8%)
Wl	*Polerovirus*	3/65 (5%)
Wl	*Tombusvirus*	3/65 (5%)
Wl	*Alfamovirus*	2/65 (3%)
Wl	*Bromovirus*	2/65 (3%)
Wl	*Nucleorhabdovirus*	2/65 (3%)
Wl	*Betacarmovirus*	1/65 (2%)
Wl	*Luteovirus*	1/65 (2%)
Wl	*Tymovirus*	1/65 (2%)

[1] Detections are the proportion of libraries where virus OTUs were detected.

Table A4.11. List of virus families detected by HTS presented according to their number of detections in the ecosystem.

Virus	Habitat	Detections[1]
Cr/Ed/Oa/Wl	*Bromoviridae*	223/306 (73%)
Cr/Ed/Oa/Wl	*Virgaviridae*	168/306 (55%)
Cr/Ed/Oa/Wl	*Potyviridae*	107/306 (35%)
Cr/Ed/Oa/Wl	*Solemoviridae*	103/306 (34%)
Cr/Ed/Oa/Wl	*Luteoviridae*	95/306 (31%)
Cr/Ed/Oa/Wl	*Tombusviridae*	41/306 (13%)
Ed/Oa/Wl	*Rhabdoviridae*	10/306 (3%)
Ed/Oa/Wl	*Tymoviridae*	3/306 (1%)

[1] Detections are the proportion of libraries where virus OTUs were detected.

Table A4.12. List of virus families detected by HTS presented according to their number of detections in Crop habitat.

Habitat	Virus	Detections[1]
Cr	*Bromoviridae*	21/37 (57%)
Cr	*Potyviridae*	17/37 (46%)
Cr	*Virgaviridae*	12/37 (32%)
Cr	*Solemoviridae*	9/37 (24%)
Cr	*Luteoviridae*	8/37 (22%)
Cr	*Tombusviridae*	1/37 (3%)

[1] Detections are the proportion of libraries where virus OTUs were detected.

Table A4.13. List of virus families detected by HTS presented according to their number of detections in Edge habitat.

Habitat	Virus	Detections[1]
Ed	*Bromoviridae*	131/133 (99%)
Ed	*Solemoviridae*	85/133 (64%)
Ed	*Luteoviridae*	81/133 (61%)
Ed	*Virgaviridae*	73/133 (55%)
Ed	*Potyviridae*	58/133 (44%)
Ed	*Tombusviridae*	14/133 (11%)
Ed	*Rhabdoviridae*	7/133 (5%)
Ed	*Tymoviridae*	1/133 (1%)

[1] Detections are the proportion of libraries where virus OTUs were detected.

Table A4.14. List of virus families detected by HTS presented according to their number of detections in Oakwood habitat.

Habitat	Virus	Detections[1]
Oa	*Virgaviridae*	46/71 (65%)
Oa	*Bromoviridae*	44/71 (62%)
Oa	*Potyviridae*	23/71 (32%)
Oa	*Tombusviridae*	15/71 (21%)
Oa	*Solemoviridae*	4/71 (6%)
Oa	*Luteoviridae*	2/71 (3%)
Oa	*Rhabdoviridae*	1/71 (1%)
Oa	*Tymoviridae*	1/71 (1%)

[1] Detections are the proportion of libraries where virus OTUs were detected.

Table A4.15. List of virus families detected by HTS presented according to their number of detections in Wasteland habitat.

Habitat	Virus	Detections[1]
Wl	*Virgaviridae*	37/65 (57%)
Wl	*Bromoviridae*	27/65 (42%)
Wl	*Tombusviridae*	11/65 (17%)
Wl	*Potyviridae*	9/65 (14%)
Wl	*Solemoviridae*	5/65 (8%)
Wl	*Luteoviridae*	4/65 (6%)
Wl	*Rhabdoviridae*	2/65 (3%)
Wl	*Tymoviridae*	1/65 (2%)

[1] Detections are the proportion of libraries where virus OTUs were detected.

Table A5. Number of samples sent for sequencing for each host plant species.

Spp	Familia	Cr	Ed	Oa	Wl	Total
Amaranthus sp	*Amaranthaceae*	6 M1V/ 1 M3V	0	0	0	7
A. undulata	*Boraginaceae*	0	0	0	1 E1F	1
B. retusum	*Poaceae*	0	1 L4F	0	0	1
Bromus sp	*Poaceae*	0	0	0	1 E1P	1
C. bourgeanus	*Asteraceae*	0	1 L2P	0	0	1
C. album	*Amaranthaceae*	5 M2V/ 1 M3V	1 L2V	0	0	7
C. tinctoria	*Euphorbiaceae*	1 M3V	0	0	0	1
C. salviifolius	*Cistaceae*	0	0	2 Q2P	0	2
C. arvensis	*Convolvulaceae*	7 M1V/ 1 M3V	0	0	0	8
C. canadensis	*Asteraceae*	0	1 L2V	0	0	1
C. melo	*Cucurbitaceae*	8 M1V/ 10 M2V/ 7M3V/ 7 M4V	0	0	0	32
C. dactylon	*Poaceae*	0	1 L3F	0	0	1
D. erucoides	*Brassicaceae*	1 M3V	1 L3P /2 L3V	0	0	4
H. stoechas	*Asteraceae*	0	0	4 Q2P	0	4
B. madritensis	*Poaceae*	0	0	0	1 E4P	1
L. arvense	*Boraginaceae*	2 M3V/ 3M4V	0	0	0	5
P. minor	*Poaceae*	4 M2V	0	0	0	4
P. echioides	*Asteraceae*	1 M3V	4 L2V	0	0	5
Q. coccifera	*Fagaceae*	0	0	7 Q2F	0	7
R. lycioides	*Rhamnaceae*	0	0	1 Q3P	0	1
S. nigrum	*Solanaceae*	3 M2V	0	0	0	3
S. dubia	*Asteraceae*	0	0	5 Q2P	0	5
T. officinale	*Asteraceae*	0	0	1 Q2P	0	1
T. botrys	Lamiaceae	0	0	5 Q2P	0	5
Total		**68**	**12**	**25**	**3**	**108**

Table A6. Accession number, specimen code, CP gene sequence status, host species, sampling year and location of the nucleotide sequences determined for WMV CP gene.

Accession	Code	CP sequence	Host species	Sampling year	Location	Latitude	Longitude
MN814346	M1V_17	Complete	*Amaranthus sp.*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN814347	M1V_19	Complete	*Amaranthus sp.*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN814348	M1V_2	Complete	*Amaranthus sp.*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN814349	M1V_48	Complete	*Amaranthus sp.*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN814350	M1V_7	Complete	*Amaranthus sp.*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN839547	M1V_14	Complete	*Amaranthus sp.*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN814407	M3V_33	Complete	*Amaranthus sp.*	2015	Zarza de Tajo	40.045900	-3.125000
MN814390	E1F_106	Complete	*A. undulata*	2015	Colmenar de Oreja	40.059138	-3.500323
MN814408	E1P_87	Complete	*Bromus sp.*	2016	Colmenar de Oreja	40.059138	-3.500323
MN814415	E4P_31	Complete	*B. madritensis*	2016	Guadalajara	40.494167	-3.131139
MN814389	L4F_2	Complete	*B. retusum*	2015	Carabaña	40.245459	-3.252617
MN814398	L2V_34	Complete	*C. album*	2015	Colmenar de Oreja	40.048758	-3.476462
MN814361	M2V_14	Complete	*C. album*	2015	Colmenar de Oreja	40.050578	-3.503040
MN814362	M2V_24	Complete	*C. album*	2015	Colmenar de Oreja	40.050578	-3.503040
MN814363	M2V_28	Complete	*C. album*	2015	Colmenar de Oreja	40.050578	-3.503040
MN814365	M2V_30	Complete	*C. album*	2015	Colmenar de Oreja	40.050578	-3.503040
MN814364	M2V_34	Complete	*C. album*	2015	Colmenar de Oreja	40.050578	-3.503040
MN814409	M3V_42	Complete	*C. album*	2015	Zarza de Tajo	40.045900	-3.125000
MN814351	M1V_11	Complete	*C. arvensis*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN814352	M1V_12	Complete	*C. arvensis*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN814353	M1V_16	Complete	*C. arvensis*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN814354	M1V_21	Complete	*C. arvensis*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN814355	M1V_5	Complete	*C. arvensis*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN839549	M1V_49	Complete	*C. arvensis*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN839539	M1V_43	Complete	*C. arvensis*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN814413	M3V_13	Complete	*C. arvensis*	2015	Zarza de Tajo	40.045900	-3.125000

Table A6. Continued.

Accession	Code	CP sequence	Host species	Sampling year	Location	Latitude	Longitude
MN814396	L2P_24	Complete	*C. bourgeanus*	2016	Colmenar de Oreja	40.048758	-3.476462
MN814399	L2V_4	Complete	*C. canadensis*	2015	Colmenar de Oreja	40.048758	-3.476462
MN814406	L3F_33	Complete	*C. dactylon*	2015	Zarza de Tajo	40.050352	-3.125794
MN814369	M2V_1	Complete	*C. melo*	2015	Colmenar de Oreja	40.050578	-3.503040
MN814370	M2V_13	Complete	*C. melo*	2015	Colmenar de Oreja	40.050578	-3.503040
MN814371	M2V_22	Complete	*C. melo*	2015	Colmenar de Oreja	40.050578	-3.503040
MN814372	M2V_27	Complete	*C. melo*	2015	Colmenar de Oreja	40.050578	-3.503040
MN814373	M2V_7	Complete	*C. melo*	2015	Colmenar de Oreja	40.050578	-3.503040
MN839540	M2V_4	Complete	*C. melo*	2015	Colmenar de Oreja	40.050578	-3.503040
MN839550	M2V_38	Complete	*C. melo*	2015	Colmenar de Oreja	40.050578	-3.503040
MN839543	M2V_50	Complete	*C. melo*	2015	Colmenar de Oreja	40.050578	-3.503040
MN839541	M2V_43	Complete	*C. melo*	2015	Colmenar de Oreja	40.050578	-3.503040
MN839542	M2V_46	Complete	*C. melo*	2015	Colmenar de Oreja	40.050578	-3.503040
MN814381	M4V_17	Complete	*C. melo*	2015	Santa Cruz de la Zarza	40.032298	-3.182801
MN814382	M4V_30	Complete	*C. melo*	2015	Santa Cruz de la Zarza	40.032298	-3.182801
MN814383	M4V_40	Complete	*C. melo*	2015	Santa Cruz de la Zarza	40.032298	-3.182801
MN814384	M4V_45	Complete	*C. melo*	2015	Santa Cruz de la Zarza	40.032298	-3.182801
MN814385	M4V_49	Complete	*C. melo*	2015	Santa Cruz de la Zarza	40.032298	-3.182801
MN839545	M4V_51	Complete	*C. melo*	2015	Santa Cruz de la Zarza	40.032298	-3.182801
MN839544	M4V_27	Complete	*C. melo*	2015	Santa Cruz de la Zarza	40.032298	-3.182801
MN814356	M1V_13	Complete	*C. melo*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN814357	M1V_18	Complete	*C. melo*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN814358	M1V_27	Complete	*C. melo*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN814359	M1V_4	Complete	*C. melo*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN814360	M1V_50	Complete	*C. melo*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN839538	M1V_29	Complete	*C. melo*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN839548	M1V_40	Complete	*C. melo*	2015	Villarrubia de Santiago	40.031840	-3.345220
MN839546	M1V_3	Complete	*C. melo*	2015	Villarrubia de Santiago	40.031840	-3.345220

Table A6. Continued.

Accession	Code	CP sequence	Host species	Sampling year	Location	Latitude	Longitude
MN814374	M3V_11	Complete	*C. melo*	2015	Zarza de Tajo	40.045900	-3.125000
MN814375	M3V_17	Complete	*C. melo*	2015	Zarza de Tajo	40.045900	-3.125000
MN814376	M3V_21	Complete	*C. melo*	2015	Zarza de Tajo	40.045900	-3.125000
MN814377	M3V_32	Complete	*C. melo*	2015	Zarza de Tajo	40.045900	-3.125000
MN814378	M3V_6	Complete	*C. melo*	2015	Zarza de Tajo	40.045900	-3.125000
MN839552	M3V_34	Complete	*C. melo*	2015	Zarza de Tajo	40.045900	-3.125000
MN839551	M3V_19	Complete	*C. melo*	2015	Zarza de Tajo	40.045900	-3.125000
MN814412	Q2P_93	Complete	*C. salviifolius*	2016	San Martín de la Vega	40.225155	-3.530613
MN814411	Q2P_141	Complete	*C. salviifolius*	2016	San Martín de la Vega	40.225155	-3.530613
MN814410	M3V_26	Complete	*C. tinctoria*	2015	Zarza de Tajo	40.045900	-3.125000
MN814397	L3P_2	Complete	*D. erucoides*	2016	Zarza de Tajo	40.050352	-3.125794
MN814404	L3V_52	Complete	*D. erucoides*	2015	Zarza de Tajo	40.050352	-3.125794
MN814405	L3V_89	Complete	*D. erucoides*	2015	Zarza de Tajo	40.050352	-3.125794
MN814414	M3V_36	Complete	*D. erucoides*	2015	Zarza de Tajo	40.045900	-3.125000
MT254410	Q2P_55	Complete	*H. stoechas*	2016	San Martín de la Vega	40.225155	-3.530613
MT254411	Q2P_76	Complete	*H. stoechas*	2016	San Martín de la Vega	40.225155	-3.530613
MT254417	Q2P_88	Partial	*H. stoechas*	2016	San Martín de la Vega	40.225155	-3.530613
MT254418	Q2P_132	Partial	*H. stoechas*	2016	San Martín de la Vega	40.225155	-3.530613
MN814386	M4V_35	Complete	*L. arvense*	2015	Santa Cruz de la Zarza	40.032298	-3.182801
MN814387	M4V_42	Complete	*L. arvense*	2015	Santa Cruz de la Zarza	40.032298	-3.182801
MN814388	M4V_50	Complete	*L. arvense*	2015	Santa Cruz de la Zarza	40.032298	-3.182801
MN814379	M3V_41	Complete	*L. arvense*	2015	Zarza de Tajo	40.045900	-3.125000
MN814380	M3V_48	Complete	*L. arvense*	2015	Zarza de Tajo	40.045900	-3.125000
MN814400	L2V_14	Complete	*P. echioides*	2015	Colmenar de Oreja	40.048758	-3.476462
MN814401	L2V_25	Complete	*P. echioides*	2015	Colmenar de Oreja	40.048758	-3.476462
MN814402	L2V_27	Complete	*P. echioides*	2015	Colmenar de Oreja	40.048758	-3.476462

Table A6. Continued.

Accession	Code	CP sequence	Host species	Sampling year	Location	Latitude	Longitude
MN814403	L2V_31	Complete	*P. echioides*	2015	Colmenar de Oreja	40.048758	-3.476462
MN814420	M3V_46	Complete	*P. echioides*	2015	Zarza de Tajo	40.045900	-3.125000
MN814418	M2V_19	Complete	*P. minor*	2015	Colmenar de Oreja	40.050578	-3.503040
MN814416	M2V_10	Complete	*P. minor*	2015	Colmenar de Oreja	40.050578	-3.503040
MN814419	M2V_37	Complete	*P. minor*	2015	Colmenar de Oreja	40.050578	-3.503040
MN814417	M2V_18	Complete	*P. minor*	2015	Colmenar de Oreja	40.050578	-3.503040
MT254420	Q2F_29	Partial	*Q. coccifera*	2016	San Martín de la Vega	40.225155	-3.530613
MT254421	Q2F_34	Partial	*Q. coccifera*	2016	San Martín de la Vega	40.225155	-3.530613
MT254422	Q2F_103	Partial	*Q. coccifera*	2016	San Martín de la Vega	40.225155	-3.530613
MT254423	Q2F_109	Partial	*Q. coccifera*	2016	San Martín de la Vega	40.225155	-3.530613
MT254424	Q2F_112	Partial	*Q. coccifera*	2016	San Martín de la Vega	40.225155	-3.530613
MT254425	Q2F_137	Partial	*Q. coccifera*	2016	San Martín de la Vega	40.225155	-3.530613
MT254416	Q2F_144	Complete	*Q. coccifera*	2016	San Martín de la Vega	40.225155	-3.530613
MN814421	Q3P_2	Complete	*R. lycioides*	2016	Villar del Olmo	40.320294	-3.205903
MN814391	Q2P_136	Complete	*S. dubia*	2016	San Martín de la Vega	40.225155	-3.530613
MN814392	Q2P_14	Complete	*S. dubia*	2016	San Martín de la Vega	40.225155	-3.530613
MN814393	Q2P_26	Complete	*S. dubia*	2016	San Martín de la Vega	40.225155	-3.530613
MN814394	Q2P_29	Complete	*S. dubia*	2016	San Martín de la Vega	40.225155	-3.530613
MN814395	Q2P_63	Complete	*S. dubia*	2016	San Martín de la Vega	40.225155	-3.530613
MN814366	M2V_26	Complete	*S. nigrum*	2015	Colmenar de Oreja	40.050578	-3.503040
MN814367	M2V_36	Complete	*S. nigrum*	2015	Colmenar de Oreja	40.050578	-3.503040
MN814368	M2V_42	Complete	*S. nigrum*	2015	Colmenar de Oreja	40.050578	-3.503040
MT254412	Q2P_36	Complete	*T. botrys*	2016	San Martín de la Vega	40.225155	-3.530613
MT254413	Q2P_41	Complete	*T. botrys*	2016	San Martín de la Vega	40.225155	-3.530613
MT254414	Q2P_64	Complete	*T. botrys*	2016	San Martín de la Vega	40.225155	-3.530613
MT254415	Q2P_72	Complete	*T. botrys*	2016	San Martín de la Vega	40.225155	-3.530613
MT254419	Q2P_149	Complete	*T. botrys*	2016	San Martín de la Vega	40.225155	-3.530613
MN814422	Q2P_13	Complete	*T. officinale*	2016	San Martín de la Vega	40.225155	-3.530613

Table A7. Analysis of molecular variance (AMOVA) results of WMV population grouped by habitat and host groups.

Level	Source of variation	d.f.[1]	Sum of squares	Mean squares	Estimated variability	Percentage of variation	*P*-value[2]
Habitat	Between groups	2	411.91	205.95	6.48	22%	$P < 0.05$*
	Within populations	105	2,360.36	22.48	22.48	78%	$P < 0.05$*
	Total	107	2,772.27		28.96	100%	
Host	Between groups	1	200.04	200.04	3.90	14%	$P<0.05$*
	Within populations	106	2,572.23	24.27	24.27	86%	$P<0.05$*
	Total	107	2,772.27		28.17	100%	

[1]d.f.: degrees of freedom.
[2]*P*-value: significant values are marked with an * ($P < 0.05$).