Wild plant species associated viromes : towards improved characterization strategies and variability in various ecological environments

Yuxin Ma

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

Ecole doctorale Sciences de la Vie et de la Santé

Mention Biologie Végétale

Par

Yuxin MA

Wild plant species associated viromes: towards improved characterization strategies and variability in various ecological environments

Sous la direction de : **Thierry CANDRESSE**

Soutenue le 12 Septembre 2019

Membres du jury :

Mme OGLIASTRO Mylène	Directrice de Recherche INRA, Montpellier	Présidente
M. MASSART Sébastien	Professeur de l'Université de Liège (Belgique)	Rapporteur
M. ROUMAGNAC Philippe	Directeur de Recherche CIRAD, Montpellier	Rapporteur
M. CANDRESSE Thierry	Directeur de Recherche INRA, Villenave d'Ornon	Directeur de these
Mme VACHER Corinne	Directrice de Recherche INRA, Pessac	Invitée

RESUME SUBSTANTIEL

Les approches de métagénomique basées sur les techniques de séquençage haut débit (high throughput sequencing, HTS) ont ouvert une nouvelle ère pour la découverte non biaisée et la caractérisation des virus. Comme pour les autres virus, de tels efforts de métagénomique montrent que la diversité des virus de plante (phytovirus) était jusqu'à très récemment largement sous-estimée. Il est donc nécessaire d'explorer la diversité de virus associés aux populations végétales et de comprendre les forces évolutives structurant cette diversité dans le temps et dans l'espace. Dans le même temps, le développement de telles études est toujours confronté à des questions d'ordre méthodologique concernant, par exemple, le choix des populations d'acides nucléiques cibles, la reproductibilité des analyses ou l'implémentation d'une stratégie pour la comparaison fiable de la richesse virale dans différent environnements. Dans cette thèse le phytovirome associé à des populations végétales échantillonnées dans différents écosystèmes, avec un focus sur les espèces sauvages ou les adventices, a été caractérisé par des approches de métagénomique par HTS. Dans ces expériences, l'analyse bioinformatique de la complexité des viromes a été conduite par deux stratégies, l'une classique basée sur l'annotation Blast des contigs pour l'identification des familles virales présentes dans un échantillon et l'autre, nouvelle, implémentée dans le pipeline "VirAnnot" et qui permet de classifier les séquences virales identifiées en OTU (operational taxonomic units) représentant un proxy des espèces virales.

Le pipeline VirAnnot a été décrit et validé au cours de cette thèse. Il permet l'identification automatisée des OTU viraux et a été utilisé dans toutes les analyses rapportées dans ce mémoire. Dans son principe, une analyse RPS-Blast est utilisée pour détecter des motifs protéiques viraux conservés codés par les contigs. Les contigs ainsi identifiés sont alors alignés et une analyse de clustering permet de regrouper dans un OTU toutes les séquences proches partageant une identité protéique ou nucléotidique supérieure à une valeur seuil. Un seuil de 10% de divergence a ainsi été validé comme permettant d'identifier des OTU basés sur la séquence conservée de la RNA polymérase RNA dépendante virale (RdRp) représentant dans de nombreuses familles virales une approximation raisonnable des espèces virales telles que décrites dans la taxonomie établie par l'ICTV (International Committee for the Taxonomy of Viruses).

Grace à la stratégie implémentée dans VirAnnot, deux approches d'enrichissement en séquences virales, la purification d'ARN bicaténaires (double-stranded RNA, dsRNA) ou d'acides nucléiques associés aux virions (Virion-associated nucleic acids, VANA), ont été comparées pour la description du virome associé à des pools complexes représentatifs des espèces végétales les plus prévalentes dans divers écosystèmes cultivés ou sauvages. Une large diversité virale, dominée par

des virus dsRNA nouveaux a été détectée dans tous les sites d'étude, avec une très large variabilité inter-sites qui a limité la capacité à tirer des conclusions quant à l'impact des pratiques culturales sur la diversité virale. Une tendance à la présence d'une diversité plus grande des virus dsRNA dans les sites non cultivés (118 *vs* 77 OTUs unique) a cependant été observée. L'approche basée sur les dsRNA purifiés a constamment révélé une diversité virale plus large que celle basée sur les VANA, et ce quel que soit le critère d'évaluation utilisé. Par ailleurs, des analyses de dissimilarité ont montré que les deux approches sont largement reproductibles mais qu'elles ne donnent pas systématiquement des résultats totalement convergents. Ces résultats illustrent des propriétés des phytoviromes associées aux écosystèmes étudiés et montrent l'intérêt de l'approche par les OTU pour estimer et comparer précisément la richesse des populations virales, permettant de raisonner les choix méthodologiques pour l'étude des phytoviromes et, vraisemblablement, des autres viromes.

Par une approche de HTS de dsRNAs purifiés, nous avons analysé le virome associé à des espèces cultivées et aux adventices et espèces sauvages environnantes dans un contexte horticultural du sud-ouest de la France. Les variations temporelles du virome ont été analysées par une approche de ré-échantillonnage sur deux années successives des mêmes populations végétales. Au total, 126 échantillons composites espèce-spécifiques représentant un total de 48 espèces végétales ont été ainsi analysés. Les données HTS obtenues ont été annotées au niveau des familles virales par Blast ainsi que par une approche de clustering des OTUs. Une large diversité virale a été identifiée, avec un total de 231 OTUs représentant 18 familles virales. La majorité des virus ainsi identifiés correspond à des virus à génome dsRNA jusqu'à présent non caractérisés. Pour les virus ssRNA (single stranded RNA), la proportion de virus nouveaux n'a été que de 48.1%. Les infections virales se sont révélées fréquentes, avec 86.5% des échantillons composites présentant au moins un OTU. Le nombre d'OTU uniques augmente linéairement avec le nombre d'échantillons analysés pour une espèce donnée, suggérant que le virome de chaque espèce végétale est probablement beaucoup plus grand et que sa description pourrait nécessiter l'analyse de plusieurs centaines de plantes pour une espèce donnée. La structure globale du virome s'est révélée relativement stable au fil du temps, en particulier s'agissant du ratio des virus ssRNA *versus* dsRNA et du nombre de familles virales détectées. Cependant, la composition du virome en termes d'OTU s'est révélée remarquablement dynamique, 68.8% des OTU n'étant détectés que dans un seul échantillon et seulement 6 OTUs (2.6%) étant détectés de façon reproductible au long des deux ans de l'étude. La recherche des OTUs partagés entre espèces cultivées et sauvages a par ailleurs démontré une sur-représentation des virus ssRNA.

Bien que les virus dsRNA soient abondants et divers dans les phytoviromes, leur origine reste une question car ils pourraient être soit des virus infectant les plantes échantillonnées soit des mycovirus infectant des champignons associés à ces plantes. Afin de tenter d'éclairer cette question, j'ai analysé en parallèle le virome associé à des pools végétaux complexes et le virome (mycovirome) associé à des pools fongiques cultivés à partir des mêmes échantillons végétaux. La diversité fongique (mycobiome) associée à ces deux types d'échantillons a également été analyse par une approche de barcoding. L'objectif était de tenter de répondre à deux questions : (1) quelle est l'origine, fongique ou végétale, des virus dsRNA identifiés et (2) quel est l'effet des pratiques culturales sur les communautés virales et fongiques associées aux diverses populations végétales échantillonnées. Les communautés fongiques identifiées montrent une riche diversité et sont dominées par des *Dothideomycetes*, *Sordariomycetes* and *Leotiomycetes*. L'analyse des viromes par séquençage HTS de dsRNAs purifiés a révélé un total de 18 familles virales comprenant des virus ssRNA de polarité positive ou négative, des virus dsRNA et des virus dsDNA, pour un total de 249 OTUs RdRp. Les approches de culturomique fongique ont capturé de l'ordre de 10% de la diversité fongique présente dans les échantillons végétaux mais les analyses de virome n'ont révélé pratiquement aucune corrélation entre les phytoviromes directement obtenus à partir des échantillons végétaux et les mycoviromes obtenus à partir des cultures fongiques. Les compositions des mycobiomes et, encore plus, des viromes ont montré une grande spécificité de site. Les comparaisons entre sites ont montré une plus grande diversité des mycobiomes dans les sites non cultivés alors que les viromes ont montré une richesse en familles virales plus importante dans les sites cultivés, suggérant que mycobiomes and phytoviromes sont structurés par des forces évolutives différentes. Des analyses complémentaires seront nécessaires pour confirmer ces données dans d'autres environnements et pour commencer à identifier les forces contribuant à structure les populations virales et fongiques associées aux plantes.

Il a été suggéré que certaines familles virales pouvaient être préférentiellement associées aux environnements cultivés ou, au contraire, aux environnements non-cultivés, soulevant à nouveau la question des forces évolutives contribuant à la composition des communautés virales. Sous ce type de questionnement général existent en métagénomique et en écologie virale des nombreuses sous-questions, comme celles de la contribution de la diversité des populations d'hôtes à la diversité des viromes, de la contribution des changements dans les populations d'hôtes à l'évolution de la pathogénicité des virus ou à l'émergence de nouveaux virus, de la contribution du virome au fonctionnement des populations végétales ou au phénotype étendu des holobiontes les hébergeant.

De la même façon, alors que les plantes sauvages et les adventices poussant à proximité des cultures constituent un réservoir potentiel pour de futures épidémies ou pour l'émergence de virus nouveaux, la fréquence et la directionnalité des flux de virus entre ces deux compartiments restent très peu documentées. Nous avons étudié par des approches de métagénomique HTS la diversité et les échanges de populations virales entre deux espèces botaniquement proches, la tomate (*Solanum lycopersicum*) et la morelle noire (*Solanum nigrum*). Une large variabilité du virome a été observée mais sans pouvoir relier cette diversité à un hôte particulier ou au contexte d'échantillonnage. Seuls 17.9% des OTU ont été trouvés partagés entre la tomate et la morelle. L'assemblage de contigs très longs a permis une analyse détaillée des populations de plusieurs virus. Deux souches (NTN et C1) du virus Y de la pomme de terre (potato virus Y, PVY) ont été fréquemment détectées dans les populations de tomate et dans les populations de morelle échantillonnées dans les champs de tomate. Par contre, le PVY s'est révélé rare dans les populations de morelle poussant au sein d'autres cultures, suggérant que les infections du PVY sont une conséquence d'un phénomène de spill-over depuis les cultures de tomate. Des populations très diverses du virus 1 du flétrissement de la fève (broad wilt bean virus 1, BBWV1) présentant des pseudo-recombinaisons entre segments génomiques ont été détectées uniquement chez la morelle, suggérant l'existence de barrière limitant le transfert du BBWV1 vers la tomate. Un nouvel Ilarvirus infectant préférentiellement la morelle mais retrouvé chez la tomate a également été identifié et nommé Solanum nigrum ilarvirus 1. Les résultats obtenus enrichissent nos connaissances du virome de ces deux espèces et des flux viraux entre elles.

L'objectif central de cette thèse était d'explorer la diversité des phytovirus par des approches de métagénomique HTS, d'étudier la prévalence et la dynamique des populations virales dans le temps et dans l'espace et d'aborder la question des forces structurant les phytoviromes dans divers environnements. L'utilisation des approches de métabarcoding a aussi permis d'analyser en parallèle les mycobiomes. Les résultats acquis les plus significatifs concernent principalement deux aspects. L'un, méthodologique, permet aujourd'hui de mieux raisonner le choix des approches mises en œuvre et de mesurer précisément la diversité virale par l'identification d'OTU qui représentent un proxy acceptable des espèces virales. L'autre, porte sur la description fine et la comparaison de viromes (et dans un cas, de mycobiomes) entre différentes situations contrastées, apportant de nouvelles connaissances sur la diversité, la dynamique et les forces structurant ces communautés virales.

LES VIROMES ASSOCIES AUX PLANTES SAUVAGES : VERS DES STRATEGIES DE CARACTERISATION OPTIMISEES ET VARIABILITE DANS DIVERS ENVIRONNEMENTS

RESUME :

Les approches de métagénomique basées sur l'utilisation des techniques de séquençage haut débit ont ouvert une nouvelle ère pour la découverte non biaisée et la caractérisation génomique des virus. Comme pour les autres virus, de telles études montrent que la diversité des virus phytopathogènes a jusqu'à tout récemment été fortement sous-estimée. Ces virus constituant une composante potentiellement importante des écosystèmes naturels ou des agrosystèmes anthropisés, il est important d'explorer la diversité des virus associés aux populations végétales et de comprendre les forces structurant cette diversité dans le temps et dans l'espace. Dans le même temps, le développement de telles études reste confronté à des questions d'ordre méthodologique concernant, par exemple, le choix des populations d'acides nucléiques à séquencer, la reproductibilité des analyses ou la disponibilité d'une stratégie permettant de comparer de façon fiable la richesse virale dans différents environnements. Dans le présent travail, le virome associé à des populations végétales échantillonnées dans différents écosystèmes, avec un focus sur les adventices et les plantes sauvages, a été caractérisé par des approches de métagénomique par séquençage haut débit. Dans ces travaux, l'analyse bioinformatique de la richesse du virome a été conduite par deux approches, l'une classique basée sur l'annotation Blast pour l'identification des familles virales présentes dans un échantillon, et l'autre, décrite et validée ici, qui permet de classifier les séquences virales métagénomiques en unités taxonomiques opérationnelles (operational taxonomic units, OTUs) utilisées comme proxy des espèces virales. Toujours dans une perspective méthodologique, les résultats obtenus avec des pools complexes de plantes représentatifs de la diversité végétale au site d'échantillonnage (approche « tondeuse à gazon ») ont permis de comparer les performances des deux techniques actuellement utilisées pour enrichir les séquences virales, la purification d'ARN bicaténaires (double-stranded RNA, dsRNA) ou d'acides nucléiques associés aux virions (virion-associated nucleic acids, VANA). Les résultats obtenus par les deux approches ont mis en évidence des viromes riches mais montrent que l'approche dsRNA devrait être préférée pour l'analyse de tels pools complexes car elle permet de façon reproductible une description plus complète du phytovirome, à l'exception des virus ADN. Les viromes caractérisés montrent, pour les populations végétales de milieux cultivés ou non gérés tempérés échantillonnées, une forte incidence virale (jusqu'à 86.5% dans 126 pools monospécifiques collectés sur une période de deux ans) et confirment la prédominance des virus

dsRNA qui représentent plus de 70% des OTU identifiés. Alors qu'une proportion significative des virus ssRNA détectés sont déjà connus, plus de 90% des virus dsRNA détectés sont nouveaux et n'avaient pas été caractérisés auparavant. Un effort important en culturomique visant à comparer le phytovirome avec le mycovirome de cultures fongiques obtenues à partir des mêmes échantillons végétaux a révélé un nombre remarquablement faible d'OTUs partagés, renforçant le questionnement sur la nature, phytovirus ou mycovirus, des virus dsRNA identifiés dans les viromes des plantes. La composition en OTU des viromes analysés s'est révélée variable entre sites d'échantillonnage mais aussi très dynamique dans le temps, avec seulement une très faible fraction des OTUs ré-échantillonnés de façon stable dans la même population végétale sur une période de deux ans. Pris dans leur ensemble, ces travaux exploratoires permettent de mieux raisonner les choix méthodologiques pour l'étude des viromes associés aux plantes et étendent notre connaissance de la diversité des phytovirus, en particulier dans des espèces végétales sauvages largement négligées, apportant des points de référence importants pour de nouveaux travaux en écologie et en évolution virale.

MOTS CLES : Métagénomique, virus phytopathogène, diversité, virome

WILD PLANT SPECIES ASSOCIATED VIROMES: TOWARDS IMPROVED CHARACTERIZATION STRATEGIES AND VARIABILITY IN VARIOUS ECOLOGICAL ENVIRONMENTS

ABSTRACT:

Metagenomics based on high throughput sequencing (HTS) has opened a new era of unbiased discovery and genomic characterization of viruses. As for other viruses, such metagenomic studies indicate that the diversity of plant viruses was until recently far underestimated. As potentially important components of unmanaged and cultivated ecosystems, there is a need to explore the diversity of the viruses associated with plant populations and to understand the drivers shaping their diversity in space and time. At the same time, the development of such studies is still faced by methodological questions concerning, for example, the choice of target nucleic acids populations, the reproducibility of the analyses or the implementation of a strategy to accurately compare virus richness in different environments. In the present thesis the phytovirome associated with plant populations sampled in various ecosystems, with an emphasis on wild plant or weed species was characterized using HTS-based metagenomics. In these experiments, the bioinformatic analysis of the virome complexity was performed using two strategies, a classical one based on Blast-based contigs annotation for the identification of the viral families present in a sample and a novel one, described and validated here, and which allows to classify the metagenomic viral sequences into operational taxonomic units (OTUs) as a proxy to viral species. Also from the methodological perspective, the results obtained using complex plant pools such as those used in the "lawn-mower" sampling strategy allowed to compare the performance of the two currently used viral enrichment methods, double-stranded RNA (dsRNA) or Virion-associated nucleic acids (VANA) purification. The results indicate both of approaches uncovered rich viromes and suggest that the dsRNA approach should be preferred when analyzing complex plant pools since it consistently provided a more comprehensive description of the analysed phytoviromes, with the exception of the DNA viruses. The virome characterization results obtained showed, for the temperate plant populations from unmanaged and cultivated sampling sites, a high virus incidence (up to 86.5% in 126 single species pools collected over a two-year period) and confirmed the predominance of dsRNA viruses with greater than 70% of the phytovirome OTUs. While a significant proportion of detected single-stranded RNA (ssRNA) viruses are already known agents, more than 90% of the dsRNA viruses are novel and had not previously been characterized. A large scale culturomics effort to contrast the phytovirome with

the mycovirome of fungal cultures obtained from the same plant samples revealed an extremely low number of shared OTUs, further deepening the debate about the phytovirus or mycovirus nature of the dsRNA viruses identified in plant viromes. The OTU composition of the analyzed phytoviromes varied significantly between sampling sites but was also shown to be highly dynamic over time, with a very low proportion of OTUs consistently re-sampled in the same plant population over a 2 years period. Taken together, these exploratory studies allow a more reasoned choice of methodology for the study of plant-associated viromes and expand our knowledge of plant virus diversity especially in neglected wild plant populations, providing important references for the further viral ecology and evolution studies.

KEYWORDS: Metagenomics, plant virus, diversity, virome

Unité de recherche

UMR 1332 Biologie du Fruit et Pathologie, INRA & Université de Bordeaux, Campus INRA de la Grande Ferrade, 71 avenue Edouard Bourleaux, CS20032, 33882 VILLENAVE d'ORNON cedex

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my supervisor **Dr. Thierry Candresse**, whose immense knowledge, valuable suggestions, patient guidance and constant encouragement make me successfully complete this thesis. His conscientious academic spirit, great enthusiasm for scientific research, modest, open-minded personality inspire me both in academic study and daily life. He gave me enormous help and advice during the whole process of the thesis writing, which has made my accomplishment possible.

I would like to extend my warmly gratitude to dear **Dr. Armelle Marais**, who has given me huge help and valuable suggestions for all my experiments, and patient revisions for all my publications. Many times it was her who sent me messages to comfort my homesickness, and it was her who always gave me great encouragements when I was lost and frustrated, which made me feel so warm in this foreign country. She is wise, capable, straightforward and thoughtful and I am lucky to have her!

I would like to express my deepest gratitude to the committees of my mid-term examination and the final defense: **Dr. Phillipe Roumagnac** and **Dr. Corinne Vacher** for their valuable advices during the mid-thesis committee which ensure the following work be in a right direction; **Dr. Mylène Ogliastro** and **Prof. Sébastien Massart** great efforts on evaluation of the manuscript with their expertise and gave the opportunity to defend my PhD. Their insightful comments pushed me to widen my research from various perspectives. I am also grateful to **Dr. Derrick Robinson**, my tutor from the Univ. Bordeaux, for his warm and sincere suggestions on my PhD study and social relationship in the lab and university.

Next, I acknowledge the contribution of my dear colleagues Chantal Faure and Laurence Svanella-Dumas, for the multiple samplings with me during my PhD, Marie Lefebvre and Sébastien Theil for their great help on data processing. I also want to thank Tania Fort for her suggestions and patient instructions for the fungi metabarcoding experiments and subsequent data processing, analysis and the article revision. Here, I would also like to thank my colleague at the same time my best friend Shuo Liu who has provided many help for my experiments and constant encouragements and confidence for the completion of this thesis. Sincere thanks also to Charlie Pauvert for insightful suggestions and discussions.

My sincere thanks also go to all my dear colleagues in the virology team. They are kind, thoughtful and open-minded and provided me selfless help for my study and life in France and great

convenience during the thesis-writing period. Luc has given me a lot of concern and help (microwave, furniture) for my good life in Bordeaux. And every first day of the new year, Veronique and Stephane invite me for family diner to share this warm moment with me. Sylvie, Val, Timi, Olivier and other researches gave me a lot of useful advices on my PhD study. A lot of thanks to dear Amandine Bordat, Amandine Barra, David, Pat, Thierry, Coralie, Jocelyne, Pascal, Genevieve, Melo, Michel, Guillaume, Vincent and Quynh for their support. Many thanks to our secretaries: Claudine, Florence, Chantal, Dominique, Veronique, they helped me with so many administrative issues that I could focus on my study.

I thank to my dear fellow labmates Vincent, Bernadette, Cecile, Xavier, Justine, Yasmine, Ahmed, Zoé, Guillaume, Maria, Flora, Fanny, Roxane, Solenne for stimulating discussions, and for all the fun we have had in the last four years. Thanks to my fellow doctoral students: Xiaozhou, Weiya, Ye, Nan, Lei, Jianqiao, Xiaojun and Hang for their warm company and support. I'd like to say thanks to my dear Pierre for his love, company, constant encouragements and trust. He led me into the French groups and helped me integrate into the French life and discover the culture differences. I will always cherish these precious memories.

Last but not the least, I would like to express my deep love and gratitude to my parents for supporting me spiritually throughout writing this thesis and my life in general.

At last, I would like to express my sincere gratitude to the China Scholarship Council (CSC) which financed my study and life in France for a four-year period.

SCIENTIFIC PRODUCTION

• *List of Publications*

- **Ma, Y.**, Marais, A., Lefebvre, M., Theil, S., Svanella-Dumas, L., Faure, C. and Candresse, T. (2019). Phytovirome analysis of wild plant populations: comparison of double-stranded RNA (dsRNA) and Virion-associated nucleic acids (VANA) metagenomic approaches. ***Journal of Virology (In press).***
- **Ma, Y.**, Marais, A., Theil, S., Svanella-Dumas, L., Faure, C., Bergey, Bernard., Couture, C., Contreras, S., Laizet, Y., and Candresse, T. (2019). Crop and wild plants/weed species-associated viromes in a horticultural context: diversity, prevalence and stability over a two-year period. ***To be submitted***.
- **Ma, Y.**, Fort, T., Marais, A., Lefebvre, M., Theil, S., Svanella-Dumas, L., Faure, C., Vacher, C. and Candresse, T. (2019). Different patterns in leaf-associated viromes and mycobiomes of wild plant populations between cultivated and unmanaged environments. ***To be submitted***.
- **Ma, Y.**, Marais, A., Lefebvre, M., Theil, S., Faure, C., Svanella-Dumas, L., and Candresse, T. (2019). Virome cross-talk between cultivated *Solanum lycopersicum* and wild *Solanum nigrum*. ***Submitted***.
- Lefebvre, M., Theil, S., **Ma, Y**. and Candresse, T. (2019). The VirAnnot pipeline: a resource for automated viral diversity estimation and operational taxonomy units (OTU) assignation for virome sequencing data. ***Phytobiomes Journal***. doi:10.1094/PBIOMES-07-19-0037-A.

• *List of Oral Communications*

- JAN 2019, 17èmes Rencontres de Virologie Végétale — Aussois, France
- JUN 2018, 10es journées des doctorants du département (SPE) de l'INRA ***-First Prize*** — Nice, France
- APR 2018, Journée Scientifique de l'École Doctorale SVS — Bordeaux, France

• *List of Posters*

- APR 2017, Journée Scientifique de l'École Doctorale SVS — Arcachon, France
- JAN 2017, 16^{ièmes} Rencontres de Virologie Végétale — Aussois, France.
- DEC 2018, International Phytobiomes Conference — Montpellier, France

FORMATIONS SUIVIES

- ***Catégorie : 2.1 Connaitre son environnement scientifique et socio-économique***

  - Advice to Young PhD Students (16 novembre 2017)

    6 heures enregistrées par : Mathématiques et Informatique.

  - Scientific Python (beginner) (20 février 2018)

    12 heures enregistrées par : Mathématiques et Informatique.

- ***Catégorie : 2.2 Développer ses compétences en langues***

  - Français Langue Etrangère (Autumn 2015) Campus Talence - Bât A21 ***17.5 heures***
  - Français Langue Etrangère (Spring 2016) Campus Talence - Bât A21 ***15 heures***
  - Français Langue Etrangère (novembre 2017) Campus Talence - Bât A21 ***40 heures***
  - Anglais -Intermédiaire (2017) Université Bordeaux Montaigne ***8 heures***

- ***Total participation: 98.5 heures / 4 module***

CONTENTS

GENERAL INTRODUCTION

GENERAL INTRODUCTION

A Brief Overview of Viruses

Viruses are infectious biological entities that replicate only inside living cells of their host(s). Collectively, they can infect all life forms including animals, plants, bacteria, fungi and archaea (Koonin *et al.*, 2006). The vast majority of viruses are much smaller than bacteria with a diameter of between 20 and 300 nanometers. That is the reason why the Chamberland bacteria filters cannot intercept viruses and why the crushed leaf extracts from infected tobacco plants remain infectious after filtration, as was found by Dmitri Ivanovsky in 1892 and validated by Martinus Beijerinck in 1898. That was the first evidence of the existence of a virus which is now known as tobacco mosaic virus (TMV) (Creager *et al.*, 1999).

Outside the cell, viruses exist as free particles called virions, which are typically nucleic acids surrounded by a protective protein shell called a capsid (Breitbart and Rohwer, 2005, Gelderblom, 1996). Some viruses also have a lipid "envelope" derived from the host cell membrane. Virion shapes can be simple helical and icosahedral forms for some viruses and be more complex structures for others. The virus nucleic acids vary between different species, as they can be either DNA or RNA, single stranded (ss) or double stranded (ds), positive sense (+) or negative sense (-) and can be linear or circular (Abrescia *et al.*, 2012; Lodish *et al.*, 2000; Zhang *et al.*, 2019). The viral genome can be encoded on a single molecule or segmented. The diversity of the ways viruses store their genetic information is therefore extremely large and without equivalent in cellular organisms. Notably, viral genomes share no absolutely conserved genes such as 16S Ribosomal RNA gene for Bacteria (Olsen *et al.*, 1986) and ITS region for fungi (Schoch *et al.*, 2012) that could be used as a target for barcoding approaches.

Based on these various types of genomes and on the mechanism(s) of replication (Baltimore, 1974, Temin and Baltimore, 1972), the Baltimore classification system has been largely used for the classification of viruses. This classification places viruses into seven groups: double-stranded DNA (dsDNA) viruses, single-stranded DNA (ssDNA) viruses, double-stranded RNA (dsRNA) viruses, positive sense (+) ssRNA viruses, negative sense (-) ssRNA viruses, ssRNA-RT viruses (that use a virally encoded reverse transcriptase, an RNA-dependent DNA polymerase, to produce DNA from the initial virion RNA genome), dsDNA-RT viruses (viruses that transcribe their DNA genome into a pre-genomic RNA used as a template during genome replication by a virally encoded reverse transcriptase). Because their small genome size and high mutation rates make it difficult to determine viral ancestry beyond the order level, the Linnaean hierarchical system was

only partially accepted by the International Committee on Taxonomy of Viruses (ICTV) with the hierarchical ranks of order, family, genus and species. The Baltimore classification was therefore used as an independent, longtime supplement to the official virus taxonomy. However, the latest 2018b taxonomy release of ICTV expands the available taxonomic ranks to 15 including Class, Phylum, Kingdom and Realm and comprises 14 orders, 143 families, 64 subfamilies, 846 genera, and 4,958 species including subviral agents such as satellites and viroids (Siddell *et al.*, 2019), with a current push to further define higher order taxa.

Plant viruses: diversity was much underestimated

While as indicated above, the number of recognized viral species remains relatively small, large-scale environmental metagenomics studies, enabled by high-throughput sequencing technologies, have revealed that viruses are the most abundant biological entities on earth with an estimated number of $10^{30\text{-}31}$ species (Breitbart and Rohwer, 2005). As another example, it has been estimated that ca. 1.7 million undiscovered viral species belonging to key zoonotic families may exist in mammals and birds (Carroll *et al.*, 2018). Yet most of what we know about viral populations comes from the marine environment and human microbiome (Brum *et al.*, 2015; Reyes *et al.*, 2010). Plant viruses, or plant-associated viruses, have been targeted by more limited efforts than animal viruses and are consequently generally less well understood, while their diversity is similarly grossly underestimated. For example, the newly released 2018b ICTV taxonomy listed a total of 4,958 virus species, however only about 1,337 species (27.0%) are plant viruses (Siddell *et al.*, 2019). This percentage is to be compared with the ca. 5,500 known mammalian species contrasted with an estimated ca 400,000 described plant species and ca. 2,000 new plant species described each year (Bachman, 2016).

Two main reasons may explain this underestimation of plant virus diversity. The first one is that viruses have traditionally been thought of as pathogens, which has led to biased studies, largely focusing efforts on viruses causing visible symptoms in economically important crops (Wren *et al.*, 2006). Indeed, the Viral Identification Data Exchange (VIDE) Database shows that most known plant viruses were initially identified from cultivated crop species (http://bio-mirror.im.ac.cn/mirrors/pvo/vide/sppindex.htm). However, crop species only comprise a minute fraction of all plant species. In addition, while the efforts of pathologists and virologists have historically been focused towards symptomatic hosts, there is also evidence that only a small fraction of viruses causes obvious symptoms and which were known as acute or chronic plant viruses (Roossinck, 2005). These acute plant viruses are frequently transmitted horizontally and

their infection may cause the death of the host, or lead to recovery of the host or conversion to chronic infections (Roossinck, 2010). In contrast to acute viruses, persistent viruses, previously called cryptic viruses (Boccardo *et al.*, 1987), are generally asymptomatic, transmitted vertically through host cell division and sexual reproduction and generally exist with low titer in host plants. Most currently known persistent viruses have double-stranded RNA genomes encoding only an RNA dependent RNA polymerase (RdRp) and a coat protein such as viruses in family *Partitiviridae* (Roossinck, 2010). As an exception, persistent viruses in the *Endornaviridae* family, that were previously regarded as dsRNA viruses, have recently been reclassified as ssRNA viruses on the basis of phylogenetic affinities, the genomic dsRNA found in infected plants being now interpreted as replication intermediates (Siddell *et al.*, 2019). Persistent viruses have been revealed to be abundant in plants, however they were so far poorly studied (Roossinck, 2010; Roossinck, 2012; Roossinck, 2015). Taken together, these biases towards cultivated crops and acute/chronic viruses suggest that similar to other parts of virology, there exists a huge gap in our understanding of plant virus diversity, evolution and ecology (Wren *et al.*, 2006).

To bridge this gap, some recent studies have started to analyze virus populations with a focus on wild plant populations, which are likely to represent reservoirs for both novel epidemies of known viruses and for novel, emerging agents (Anderson *et al.*, 2004; Cooper and Jones, 2006; Elena *et al.*, 2014; McLeish *et al.*, 2019; Stobbe and Roossinck, 2014). The few metagenomic studies performed to date have shown that virus occurrence is quite common in wild plants, independently of the presence of symptoms, with as high as 70% prevalence and most them are novel to science (Roossinck, 2010). These viruses were usually detected from the *Partitiviridae*, *Endornaviridae*, *Totiviridae* and *Chrysoviridae* families, the latter two families being only considered as suspected/potential persistent plant viruses because all or the vast majority of the viruses they contain are known as fungal viruses. However, members of these families are frequently and abundantly detected from plants (Roossinck, 2015). Consequently, it remains to be determined whether all or part of the persistent viruses identified from plant viromes have a fungal origin and infect fungi associated with the analyzed plant samples or whether they are plant viruses replicating in the sampled plants. It has been argued that the latter hypothesis seems more plausible given the very small amount of fungal tissue that is found in plants harboring endophytes (Roossinck, 2010).

The second reason for our limited knowledge of plant virus diversity lies with the limitation of traditional virus detection technologies, which are based either on virus biological properties related to the interaction(s) with host(s) and/or vector(s) or on intrinsic properties of the virus itself

such as its coat protein and nucleic acid(s) (Naidu and Hughes, 2003). Precipitation/agglutination tests, enzyme-linked immunosorbent assays (ELISA) and immunoblotting are all generally based on the detection of the viral coat protein and molecular hybridization assays and polymerase chain reaction (and its many variants) are based on the detection of the virus nucleic acid(s) (Boonham *et al.*, 2014; Naidu and Hughes, 2003; Yadav and Khurana, 2016). These methods have for a long time helped virologists detect and identify viruses but have serious limitations, such as requiring *a priori* knowledge of the virus, being often slow and/or labor-intensive, often lacking sensitivity or throughput. These limitations have over the years imposed strong restrictions to the virus discovery process, highlighted for example by the range of diseases for which the causal virus was only identified recently or even still remains to be identified. This situation has dramatically changed with the development of next-generation sequencing (NGS) technologies (Delwart, 2007; Mokili *et al.*, 2012; Roossinck *et al.*, 2015; Rosario and Breitbart, 2011; Wu *et al.*, 2015).

The next generation sequencing combines three major improvements: (i) instead of cloning of DNA fragments, it relies on the preparation of libraries in a cell free system; (ii) thousands-to-millions of sequencing reactions are conducted in parallel and (iii) the sequencing output is directly detected without the need for electrophoresis, base interrogation being performed cyclically and in parallel (van Dijk *et al.*, 2014). Thanks to the very high volume of sequence data thus produced at relatively low cost and to progress in the bioinformatic analysis of these sequence data, high throughput sequencing (HTS), also referred to as NGS, deep sequencing or large scale sequencing has largely superseded in the past 15 years all previously used virus discovery approaches (Maree *et al.*, 2018; Massart *et al.*, 2014; Rott *et al.*, 2017; Villamor *et al.*, 2019).

Application of high throughput sequencing (HTS) in plant virology

In the plant virology research field, HTS has been used to unravel the etiology of some disease through the discovery of both known and unknown viruses, for the study of viral intraspecific genetic diversity, for the analysis of plant response to infection and in epidemiology and virus ecology studies (Massart *et al.*, 2014; Villamor *et al.*, 2019).

First of all, HTS has been used in etiology efforts for viral pathogen discovery, largely displacing the previously used, biossay-based approaches. One well-known and particularly illustrative example concerns the discovery of Grapevine red blotch virus (GRBV). Initially grapevine red blotch disease, with symptoms resembling those of grapevine leafroll disease, was first observed in 2008 in a *Vitis vinifer*a ‘Cabernet Sauvignon’ vineyard in California and a ‘Cabernet franc’ vineyard in New York (Sudarshana *et al.*, 2015). However, extensive conventional testing showed

negative results for all known leafroll-associated viruses. The identification of GRBV was first achieved in 2011 using Illumina HTS on cDNAs prepared from purified double stranded RNA (dsRNA) (Rwahnih *et al.*, 2013). Meanwhile other similar studies using other target nucleic acids populations for HTS also revealed the existence of this novel ssDNA Gemini-like virus in the infected samples (Krenz *et al.*, 2012; Poojari *et al.*, 2013). Since then, many symptomatic, leafroll-negative vines have been shown to be positive for GRBV (Krenz *et al.*, 2014). Subsequently the infectivity of GRBV and its involvement in the disease have been validated by the construction of an infectious clone which reproduced typical disease symptoms upon inoculation in grapevine (Cieniewicz *et al.*, 2018; Yepes *et al.*, 2018). In another example, the causal agent of mulberry mosaic dwarf disease, which has reduced mulberry production in China for a century was finally identified when the results of small RNA sequencing on the Illumina platform indicated the existence of a new geminivirus seemingly responsible for this disease (Lu *et al.*, 2015; Ma *et al.*, 2015). A highly significant correlation between virus presence and disease was then observed, and the virus correspondingly named mulberry mosaic dwarf-associated virus (MMDaV) and the causal role later demonstrated by bioassays (Ma *et al.*, 2015; Yang *et al.*, 2018).

High throughput sequencing has similarly been applied in plant materials undergoing quarantine or post-quarantine screening, revealing viral infections that had escaped more classical tests. For example, Candresse *et al.* (2014) discovered by HTS of virus-derived small interfering RNAs (siRNAs) and of virion-associated nucleic acids (VANA) a novel mastrevirus named sugarcane white streak virus (SWSV) in two quarantined sugarcane plants. In another example, Bag *et al.* (2015) detected using HTS a new luteovirus from nectarine plants after post-entry quarantine. Through this study, they suggested the importance of including HTS analysis as an essential tool to assess the plant health status of traded propagation plant materials, as a supplement to the traditionally used biological indexing process (Bag *et al.*, 2015). These studies highlight the "blind spot" that exists in most classical approaches through their reliance on prior knowledge of the viruses that are to be detected. For its part, biological indexing does not rely on such prior knowledge but is only effective if specific symptoms are observed in the indicator plant(s) used. The current quarantine-testing protocols, that are therefore largely limited to the testing of known pathogens incompletely address the risk of invasion of new viral pathogens, explaining the interest to complement them with HTS-based approaches that have the potential to detect any viral agent present, irrespective of the existence of any prior knowledge (Bag *et al.*, 2015; Candresse *et al.*, 2014; Maliogka *et al.*, 2018; Maree *et al.*, 2018; Villamor *et al.*, 2019).

High throughput sequencing has also been applied in the study of intra-host or intra-specific diversity of viruses. As we know, viruses have very high mutation rates (and particularly RNA viruses), short generation times and large population sizes, all factors that may result in high degree of genetic diversity of virus populations in an infected host (Duffy *et al.,* 2008). Because of their diversity, intra-host virus populations are often referred to as mutant clouds, swarms, or viral quasi-species (Beerenwinkel *et al.*, 2012; Lauring and Andino, 2010). Based on HTS and bioinformatic analyses, Jo *et al.* (2017) have revealed viral populations and identified their quasi-species in susceptible and resistant pepper plants. HTS approaches also enable the complete assembly of viral genomes and reveal single nucleotide variations at a pangenomic level (Jo *et al.,* 2015; Jo *et al.*, 2016). High intra-host and intra-species diversity of plum bark necrosis stem pitting-associated virus (PBNSPaV) (Marais *et al.*, 2013) and little cherry virus 1 (LChV1)(Katsiani *et al.*, 2018) has been accurately explored through HTS indicating new era of highly accurate virus diagnosis. It has also been reported that mathematical models coupled with HTS can accurately describe both selection and genetic drift shaping the evolutionary dynamics of viruses within or between hosts (Fabre *et al.*, 2012). HTS has been used to reveal the intra- and inter-hosts genetic diversity of zucchini yellow mosaic viruses under natural and greenhouse conditions (Simmons *et al.*, 2012). The obtained high resolution sequencing data suggested that some mutations persisted during inter-host transmission as well as within individual hosts, an indication that vector-imposed transmission bottleneck and systemic bottleneck may not be as strong as initially thought (Simmons *et al.*, 2012). The development and application of the next wave of very long reads HTS approaches such as the Minion, may bring further improvements to our ability to analyze viral diversity at a viral pangenomic scale (Filloux *et al.,* 2018b)

In addition, HTS-based metagenomics has been used for the non-targeted discovery and description of viral communities in a wide range of environments or holobionts by analyzing viral nucleic acids (virome). Virome studies have been conducted on various plant individuals, cultivars and species. For example, the virome in a single grapevine has been characterized by HTS of double-stranded RNAs, revealing the presence of a substantial set of mycoviruses among the detected viruses (Al Rwahnih *et al.*, 2011). Still for grapevine, the grapevine population in a vineyard has been screened by HTS of dsRNA and a census virome built (Coetzee *et al.*, 2010). The viromes in peach cultivars (Jo *et al.,* 2018), in sweet potato cultivars (Gu *et al.,* 2014), in watermelon pools have also been studied (Luria *et al.,* 2019). These studies generally revealed the presence of both known and novel viruses but they most often represent small scale surveys mainly addressing issues such as the prevalence of specific viruses or monitoring epidemies. Beyond

studies of the virome of individual plants or plant species, large and complex plant populations representative of the flora of a given environment or sampling point have also sometimes been analyzed with the aim to gain information on the biotic or abiotic factors shaping the diversity and composition of viromes (Bernardo *et al.*, 2018; Susi *et al.*, 2019; Thapa *et al.*, 2015). In an early work, Thapa *et al.* (2015) has investigated over four years 400 plants from six wild plant species collected from twenty sites in the Tallgrass Prairie Preserve in northeastern Oklahoma. At species, spatial or temporal scales, the viromes were described and compared and the results indicated that host identity has a significantly stronger effect on virome composition than other factors such as location and sampling year (Thapa *et al.*, 2015). An investigation of the viromes from 1,725 geo-referenced plant samples collected over two years in two biodiversity hotspots (Western Cape region of South Africa and Rhône river delta region of France) suggested that agriculture substantially influences plant virus distributions at a landscape scale (Bernardo *et al.*, 2018). Different from the relatively small scale metagenomic studies focusing on a crop species, these large scale metagenomic studies focused on wild plant species and tried to classify the enormous amount of viral sequences into Operational Taxonomy Units (OTUs) for a better assessment of virus diversity (Bernardo *et al.*, 2018; Thapa *et al.*, 2015). These pioneering studies have started to fill the knowledge gap on virus diversity in wild plant species, and have provided strategies for the implementation of plant virus metagenomics with an ecology perspective, making pioneering efforts towards the precise assessment of viromes diversity.

Critical methodological points for the implementation of the "wet lab" part of HTS-based plant virus metagenomics

1. Sampling

As the earliest step for HTS-based virome study, the choice of the sampling strategy has a major importance, in particular in relation to the question(s) being addressed by the study. As we briefly described above, the analyzed samples can be individual plants, groups of plants of the same species or complex pools containing different plant species and reflecting the sample plant population at a sampling point or in an environment. When the objective is to describe the whole virome of a multispecies plant population, the two main strategies used are the so-called "Metagenomics" and "Ecogenomics" approaches as defined by Roossinck (2012). The "Metagenomics" strategy, also called the "lawnmover" method generally pools all the above ground parts of plants and combines them into a single (highly) composite sample (Roossinck *et al.*, 2015) representative of the sampling site flora and of the associated virome. It thus provides a

global vision of the virome but does not preserve the information about the host(s) of the individual detected viruses. The other strategy, "Ecogenomics", allows to trace back each output viral sequence to host plant(s), thus retaining the host(s) information, but may require the analysis of many more sample in order to get a global virome vision for complex plant populations (Bernardo *et al.*, 2018; Roossinck *et al.*, 2015).

2. Nucleic acids preparation

Plant viruses, like many other viruses, are characterized by two properties: (i) they are highly variable and as a group do not share universally conserved sequences that might be used for barcoding approaches and (ii) they are rarely accessible outside of their hosts or vectors (presence in surface waters of some plant viruses would be a counter example to this second property; (Mehle *et al.*, 2014; Mehle *et al.*, 2018; Ravnikar *et al.*, 2018)). A practical consequence of these features is that HTS-based metagenomics studies of plant viruses generally use very complex nucleic acids mixtures that contain both hosts and viral nucleic acids. A range of potential nucleic acids populations can be targeted and have in practice been used in virus discovery efforts. These include total RNA (totRNA) with or without ribosomal RNA depletion, polyadenylated RNA (poly(A)RNA), double stranded RNA (dsRNA), virion-associated nucleic acids (VANA), virus-derived small interfering RNA (sRNA) and RNA after subtractive hybridization with healthy plant RNA (Adams and Fox, 2016; Roossinck *et al.*, 2015; Wu *et al.*, 2015). These methods differ in their efficiency at capturing viruses with different genome types (as listed in the Baltimore classification) and in the enrichment in viral sequences they offer. Their advantages and disadvantages have been reviewed in detail (Adams and Fox, 2016; Roossinck *et al.*, 2015). A brief summary of these approaches is provided here.

Total RNA: one of the most direct approaches, it does not enrich in viral sequences but can detect a large spectrum of RNA viruses, DNA viruses and viroids; its main disadvantage is that large amounts of non-viral sequences are generated, including for host ribosomal RNAs. As a consequence, high sequencing depth is needed, in particular for low titer viruses, making this approach more costly and more intensive in the bioinformatics analysis phase.

Ribo-depleted total RNA: a modification of the total RNA approach in which the plant ribosomal RNAs are removed from the total RNA before sequencing, resulting in a ca. ten-fold enrichment in viral sequences (Adams and Fox, 2016). Similar to total RNA, it allows the detection of all types of viral agents. The cost of this approach remains significant because of the extra cost imposed by the ribo-depletion step.

Poly(A) RNA: similar to ribo-depletion, the purification of messenger RNAs through the selection of poly-adenylated molecules counter-selects the host ribosomal RNAs (and other noncoding RNAs), allowing some level of enrichment of viral sequences. However, viruses with genomes that do not contain a polyA are also counter-selected (Wu *et al.*, 2015).

Small interfering RNA: This approach focuses on the small 21-24 nucleotides (nt) RNAs which are produced by cleavage of viral RNAs by the host Dicer enzymes as a consequence of the antiviral silencing defense reaction (Hamilton and Baulcombe, 1999; Lu *et al.*, 2003). The advantage of this approach is the generality of the silencing defense and therefore, the ability to detect RNA viruses, DNA viruses and viroids (Pooggin, 2018). As for total RNA, a lot of host-derived sequences are generated in parallel with viral ones and the proportion of viral reads may be quite low, in particular in woody species (Massart *et al.*, 2018). In addition, assembly of viral genomes from the small siRNA reads is often not as efficient and straightforward as for the long reads produced by other approaches (Massart *et al.*, 2018).

Double stranded (ds)RNA: this approach is based on the purification of double-stranded RNAs from the analyzed plant sample (Marais *et al.*, 2018). This particular type of nucleic acids is generally absent from non-infected hosts and is produced during their replication by all types of RNA viruses (Weber *et al.*, 2006). Double-stranded RNAs are also sometimes observed for some DNA viruses (possibly as a consequence of incomplete bi-directional transcription termination) but this is not a general feature so that DNA viruses are largely counter-selected by this approach (Roossinck *et al.*, 2015). This dsRNA-based approach has been also used for the discovery of fungal viruses (Roossinck, 2015). Double-stranded RNA purification may provide a high level of enrichment of viral sequences (Roossinck *et al.*, 2010), thus reducing the sequencing power (and associated cost) needed as compared to no/low enrichment approaches.

Virion-associated nucleic acids (VANA): this is undoubtedly the most widely used technique in viral metagenomics (Bernardo *et al.*, 2018; Filloux *et al.*, 2018a; Thapa *et al.*, 2015), in part because it is particularly well suited to analyze viruses present in environmental water samples (Rosario *et al.*, 2009). It is somewhat less direct when host samples are to be used. It relies on the (semi)purification of viral particles by differential centrifugation (Filloux *et al.*, 2015). Non encapsidated nucleic acids are then removed by a nuclease digestion step, before protected viral nucleic acids are finally recovered following the disruption of viral particles. It effectively enriches viral nucleic acids of encapsidated viruses but requires rather complex sample

processing. In addition, the way in which its performance might be affected for viruses with unstable particles or by hosts rich in purification-interfering components remains a question.

Nucleic acids selected by subtractive hybridization: it is possible to enrich viral sequences by first performing a substractive hybridization step against healthy host(s) nucleic acids. This approach requires an access to healthy host(s) and involves time-consuming and complex processing; It is therefore considered not well suited in high throughput diagnostic settings but can be useful for etiology studies (Adams *et al.*, 2009).

Sequence-independent sequencing: pointing to.

Amplicon sequencing: it is also possible to sequence amplification products. These can come in the form of rolling circle amplification (RCA) products that have proved useful for the detection or characterization of DNA viruses with circular genomes such as Geminiviridae, Nanoviridae or of viruses with pseudocircular genomes such as Caulimoviridae (Idris *et al.*, 2014; Jeske, 2018; Ng *et al.*, 2011; Rosario *et al.*, 2013). They can also be PCR products obtained using polyvalent, genus or family-specific primers targeting conserved genomic regions. This approach is then very close to the barcoding approaches used in fungal or bacterial metagenomics but with a narrower taxonomic breadth. Given the upstream PCR amplification, this strategy offers higher resolution for the parallel detection of both high and low titer viruses. The amplicon sequencing strategy can also be tuned to study viral intra-specific diversity such as in a study of the diversity of prunus necrotic ringspot virus (PNRSV) in *Prunus* trees (Kinoti *et al.*, 2017).

Overall, the main difference and advantages/disadvantages of the above approaches mainly concern the spectrum of detectable viruses and the enrichment achieved (with consequences for sequencing depth and cost). There are also some potential considerations on applicability to a wide range of host species. As a consequence, the choice of approach may vary depending of the study objective(s), on the number and complexity of the samples to be analyzed or on the available budget. Given that there have been so far few side-by-side comparisons, it may not be easy to determine the best choice or even whether there exists such a best choice. To gain a clearer vision and reason the choice of target nucleic acids population in a certain context, more comparative analyses are needed. A few such comparisons have so far been performed. For example, a comparison of virus-derived small interfering RNAs (siRNA) and virion-associated nucleic acids (VANA) for a new DNA virus discovery was reported by Candresse *et al.* (2014). In this case, higher genome coverage and longer contigs were generated using VANA than siRNAs. To test whether the same representation of within-host viral population structure could be obtained,

siRNA and VANA-RNA have been compared by Kutnjak *et al.* (2015). The results revealed that both approaches provided highly similar viral mutational landscapes but also indicated that VANA-derived sequences performed better in complete viral genome reconstruction and allowed to more readily detect recombinant genomes (Kutnjak *et al.*, 2015). The comparison of siRNA and ribosomal RNA depleted total RNA for citrus tristeza virus [(+)ssRNA, *Closteroviridae*] and citrus dwarfing viroid (*Pospiviroidae*) characterization in grapefruit showed that rRNA-depleted total RNA is superior to sRNA in *de novo* genome assembly and coverage for the closterovirus but not for the viroid (Visser *et al.*, 2016). For the detection of viroids and of plant viruses with different genome types in nine different plant samples, the performance of these two approaches was virus-dependent, but longer contigs and higher genome coverage were generated using rRNA-depleted total RNA (Pecman *et al.*, 2017). In the sole study to date that incidentally compared dsRNA and VANA for wide scale metagenomics to describe viral diversity in six native plant species from the Nature Conservancy's Tallgrass Prairie Preserve in northeastern Oklahoma, the results showed that more operational viral taxonomic units (OTUs) were discovered by the dsRNA approach (29 against seven for VANA). In addition, 86% of VANA-OTUs were also detected by dsRNA. The two approaches also showed different performance when analyzing the effects of sites on virome compositions (Thapa *et al.*, 2015). Overall, while all approaches have proven feasible and yielded interesting results in virus discovery studies in which a limited number of simple samples are generally analyzed, two of them, dsRNA and VANA have been consistently chosen for wider scale metagenomics studies because the enrichment of viral sequences they offer directly translate in lower sequencing costs when a larger number of samples or more complex samples need to be analyzed. However, while these two approaches have been shown to perform well in a range of plants and for a range of viruses, there is still very limited information allowing to reason such a methodological choice in plant virus metagenomics studies.

3. Sequencing platforms

The first HTS platform, Roche 454 was originally released in 2005. This platform captures a template molecule in a bead that is further loaded on a well of a picotiter plate for amplification using emulsion PCR and finally sequenced using pyrosequencing (Rothberg and Leamon, 2008). The Illumina sequencer, which largely displaced it, is based on sequencing by synthesis using fluorescently labeled dye-terminators and the process of bridge amplification of adaptor-ligated DNA fragments on the glass surface of flow cell (Bentley *et al.*, 2008). The Illumina platform has been and still is the most widely used technology as it provides the highest throughput, lowest

error rate and is the most cost effective among currently available HTS platforms (Villamor *et al.*, 2019). SOLiD is a system that utilizes a sequence by ligation method using a DNA ligase (Valouev *et al.*, 2008): it provides the second highest throughput after Illumina but only accommodates 75 bp (100 bp for paired-end read) as the longest read length. The Ion Torrent platform can produce 400 bp read length, however the throughput is still lower than that of the Illumina and SOLiD systems (Rothberg *et al.*, 2011), while the error rate is higher and comparable to the of the 454 pyrosequencing. Different from the above mentioned second-generation technologies, the third-generation sequencing platforms require no template amplification prior to sequencing since individual RNA/DNA molecules are used as templates (Rhoads and Au, 2015; Wang *et al.*, 2015). For example, PacBio-Illumina is the most popular third-generation platform, and uses hairpin adaptors to form a closed ssDNA template called SMRTbell (Rhoads and Au, 2015). This platform can generate very long reads (20 kilobases (kb) and more) but has a high error rate. The other third-generation sequencing platform, proposed by Oxford Nanopore generates similar very long reads but higher error rate output but a lower throughput. On the other hand, it has the advantage of being highly portable in its MinION format (Deamer *et al.*, 2016; Jain *et al.*, 2016). Despite the high error rate, >99% accuracy of consensus sequence has been achieved with the MinION and given the low set up cost and portability of this platform, it has already generated interest in the plant virus field for example for the detection of maize streak virus, maize yellow mosaic virus and maize totivirus in maize plants (Adams *et al.*, 2017), of plum pox virus in plum plants (Bronzato Badial *et al.*, 2018) or of viruses affecting water yam plant (Filloux *et al.*, 2018b). The latter study also compared the performance of the Illumina and MinION platforms for the quality of the genomic sequences obtained, demonstrating that high quality sequences (>99.8% accuracy), very close to Illumina ones can be obtained with the MinION despite its high error rate (Filloux *et al.*, 2018b). Since this technology may provide excellent genome reconstruction together with high consensus sequence accuracy, it might represent the future for viral metagenomics because could solve the problems linked to the short read length, such as incomplete, or chimeric genome assemblies (Filloux *et al.*, 2018b).

Critical methodological points for the implementation of the "dry lab" bioinformatics part of HTS-based plant viral metagenomics

1. Reads demultiplexing, cleaning, assembly and annotation

Generally, during the library preparation step, individual "barcode" sequences are added to each DNA fragment, which are called Multiplex Identifiers (MIDs) and allow many libraries to be

pooled and sequenced simultaneously in a multiplexed format during a single run. While it effectively reduces the cost of HTS, multiplexing however introduces some other problems for the downstream analysis such as mistagging (Esling *et al.*, 2015) or index-hoping (Illumina, 2017; van der Valk *et al.*, 2019) which may results in a low background of inter-sample cross-talk.

A typical HTS dataset is original stored in a proprietary format or as FASTQ files and sequence quality can be evaluated by FASTQC program (Andrews, 2010). The generated reports can be used for the subsequent trimming of low quality reads. The trimmed sequences will be demultiplexed using available softwares (Blawid *et al.*, 2017). After this pre-processing, the most widespread approach is *de novo* assembly into contigs using a range of pipelines (Villamor *et al.*, 2019) or commercial softwares such as CLC Genomics Workbench (https://www.qiagenbioinformatics.com/products/clc-main-workbench/). This assembly step is in particular known to improve the efficiency of identification of viral sequences and to reduce the volume of the unannotated "dark matter" (Francois *et al.*, 2018). The annotation of sequences and the search for viral ones are conventionally performed by homology searches using Blast (Altschul *et al.*, 1990) or similar programs. An alternative option is to rely on the targeted search of specific conserved motifs using RPS-Blast (Reverse Position-Specific BLAST; (Marchler-Bauer *et al.*, 2009)) for comparison with motifs databases such as PFAM (El-Gebali *et al.*, 2018; Punta *et al.*, 2011), NCBIfams (Haft *et al.*, 2018) or SMART (Letunic and Bork, 2017). On the other hand, if the identification of known viruses is the objective, the pre-processed reads can be direct mapped on reference viral genomes using a range of available tools (Fonseca *et al.*, 2012).

2. Difficulties in linking metagenomics sequence annotation with the ICTV taxonomy

Enormous amounts of viral sequences, including long scaffolds and short fragments, have been identified and annotated, generally by Blast-based approaches, potentially down to virus species level with corresponding identity percentages and e-values. However, these seemingly annotated nucleotide/translated amino acid sequences are not only associated with little/no biological data but also provide quite limited taxonomic reference points. As mentioned above, a large fraction of viruses identified in metagenomics studies are novel and have therefore no counterpart in Genbank or other similar databases. A consequence is that Blast will only identify the closest virus present in the database, even if the relationship to the annotated sequence is only very distant. The species- and genus-level annotation are frequently considered unreliable for novel viruses so that a conservative family-level annotation is frequently used, even if its reliability is incomplete (Massart *et al.*, 2014; Roossinck *et al.*, 2010). This problem is compounded with other ones

reviewed in detail by Simmonds (2015). The frequent incomplete coverage and assembly of virus genomes induced by the random whole genome amplification process or by the mere complexity of the viromes analyzed, the possible assembly problems leading to recombinant contigs and the fact that different genomic regions many have different evolutionary origins or be under very different selection pressures all cause difficulties when trying to assign sequences to the original virus entity. The same applies to the variability in the criteria used to define taxons in different viral families. It should also be considered that HTS can generally not establish links between the genomic segments of multipartite viruses. Lastly, the currently used homology-based annotation approaches may not be able to detect highly novel and divergent viruses which have no counterpart in the databases used for comparison purposes so that other approaches may be needed (Soueidan *et al.*, 2015).

As mentioned above, the Blast-based annotations of novel agents are rarely acceptable at genus-level and have to be considered even at family level (Bernardo *et al.*, 2018; Thapa *et al.*, 2015). For example, in the virome study of wild plant species in Nature Conservancy's Tallgrass Prairie Preserve in northeastern Oklahoma, Thapa *et al.* (2015) combined viruses into relatively broad taxonomic categories at family and genus level, a consistent process performed for all samples. In another metagenomics study, virus prevalence and diversity were evaluated only at family level and, on the basis of pairwise sequence similarity related virus sequences were grouped into operational taxonomic units (OTUs) at family level for the deeper analysis (Bernardo *et al.*, 2018).

The notion of defining OTUs for viruses was raised as early as 2013, when an OTU-like approach was proposed to analyze virus diversity in an Australian Hypersaline lake sample (Emerson *et al.*, 2013). This was done by first predicting functional domains of query sequences using the InterProScan tool (Quevillon *et al.*, 2005). The predicted protein sequences were clustered at 40% amino acid identity with 0.3% mis-clustered sequences and then seven "universal" marker genes (viral gene signatures) to group the sequences into OTUs in a way to maximize the included virus populations (Emerson *et al.*, 2013). The use of such broadly but not universally conserved protein signature sequences to regroup viral sequences in OTUs was also proposed by Klingenberg *et al.* (2013) and theorized by Simmonds (2015). This potentially wide ranging solution, has three basic steps (i) the identification and alignment of informative conserved genome regions of contigs and of reference viral genomes, (ii) a phylogenetic analysis of the alignment and the regrouping of related sequences through a clustering approach to define OTUs and, if needed (iii) the annotation of these OTUs by Blast analysis of representative sequences. This last step potentially allows the

naming of the new viral entities, indicating the clear analogy with existing virus genera and species (Simmonds, 2015).

The number and diversity of viral sequences that are identified in metagenomic data far exceeds that of experimentally characterized virus isolates. Generally, the approval of a new species by the ICTV depends on the availability of biological data for the corresponding virus, which has always limited the number of recognized virus species catalogued in the master species list (MSL). Unfortunately, viral sequences discovered by metagenomics are very frequently incomplete and may not be associated with any biological information. Given the importance of metagenomics data in revealing vast, previously unknown parts of the virosphere, the ICTV has recently changed its perspective and rules about the elements needed to describe a novel virus species (Simmonds *et al.*, 2017). It thus decided that, with appropriate quality control, viruses that are known only from metagenomic data can, and should be, incorporated into its official classification scheme. A minimal requirement of many ICTV Study Groups for such a change is today the availability of complete genome information (as opposed to the complete genome, since genomes ends are frequently missing from metagenomics data).

General features of plant-associated viromes

Despite of the limited number of metagenomics studies performed to date, some general features of plant associated viromes can be tentatively identified. Among these is the observation that the richness in DNA viruses tends to be lower than that of RNA viruses, respecting in that the balance between these two groups in the currently recognized ICTV taxonomy (Bernardo *et al.*, 2018). A second observation is that diverse dsRNA viruses, a very large proportion of which appears to be novel, usually dominate the plant-associated viromes (Bernardo *et al.*, 2018; Thapa *et al.*, 2015). In addition, when enriching specifically for viral nucleic acids by VANA or dsRNA purification, there almost always remains a significant proportion of reads that are not detectably homologous to any known agents (Roossinck, 2015). This fraction has been sometimes referred-to as "dark matter" and its viral nature remains an open question.

As indicated above, dsRNA viruses belonging to such families as *Partitiviridae*, *Amalgaviridae*, *Endornaviridae* (moved to ssRNA virus recently), *Totiviridae* and *Chrysoviridae* appear to represent a large fraction of plant-associated viromes with a 70% prevalence (Roossinck *et al.*, 2010). A similar observation of a high frequency of dsRNA viruses belonging to these families was observed in single grapevine virome (Al Rwahnih *et al.*, 2011). While members of the family *Amalgaviridae* appear so far to be restricted to plants, viruses in the *Partitiviridae*, *Endornaviridae*

and possibly *Totiviridae* families (Cox *et al.*, 2000; Fermin *et al.*, 2018) have either plant (as persistent plant viruses) or fungal (mycoviruses) hosts. *Chrysoviridae* members were only reported so far to infect fungi (Roossinck, 2015). A recurrent question is therefore whether these frequent dsRNA viruses are really plant viruses or whether they infect fungi that are associated with the analyzed plant samples (Roossinck, 2012). Indeed, during the characterization of the virome of a single grapevine by dsRNAs HTS, twenty-six putative fungal virus groups were identified, from which 19% (5/26) were found to infect fungal cultures isolated from the analyzed plant, suggesting that at least part of these agents are indeed mycoviruses (Al Rwahnih *et al.*, 2011). The analysis of the mycovirome of mycelial cultures isolated from plant samples certainly constitutes a way to address this question, even if it is well-known that only a minor fraction of the fungi present can be cultured (Blackwell, 2011; Feldman *et al.*, 2012).

Different Virus infection patterns in crops and in wild plants?

It has been suggested that virus families characterized from crops may be quite different from those from wild plants (Roossinck, 2012) and the results of Bernardo *et al.* (2018) suggested that particular viral families may be more frequently associated with agricultural contexts while others would be more frequently associated with native vegetation. Such results or observations raise many questions that, at a more fundamental level, can be regrouped by the general question of the identity of the evolutionary forces and drivers directing the assembly of viral communities. Under this “umbrella” question are many of the problems currently being addressed or raised in viral ecology and metagenomics, such as the contribution of the diversity of plant populations to virome diversity (Malmstrom *et al.*, 2011; Shates *et al.*, 2018), the contribution of changes in plant populations on viral pathogenicity or emergence (Elena *et al.*, 2014; Pagán *et al.*, 2012; Rodelo-Urrego *et al.*, 2013; Rodríguez-Nevado *et al.*, 2017; Sacristán *et al.*, 2004), the contribution of viruses to the functioning of plant populations (Malmstrom and Alexander, 2016), the dynamics of plant-associated viromes and the intensity and directionality of fluxes of viruses between crops and wild plants or whether viruses contribute significantly to the extended phenotype of plant holobionts (Shates *et al.*, 2018).

SCIENTIFIC QUESTIONS ADDRESSED IN THE PRESENT THESIS

Several aspects related to methodological aspects or to pending questions in plant virus metagenomic research are addressed in the different chapters of this thesis. Each chapter is presented as a manuscript in the process of being submitted. Throughout this work, I have used various approaches to characterize the virome of different plant or fungal samples but a constant logic has been to try as much as possible to do so not only at family level but also using an approach that is based on the definition of OTUs. This has been made possible by the development of an annotation pipeline (virAnnot) which integrates a routine for the automated definition and annotation of OTUs. During this PhD, I contributed to the definition and validation of a clustering cut-off value that allows to define in many families OTUs that are an acceptable proxy to viral species. A submitted Resource Announcement describing this pipeline and to which I am associated is provided in an annex.

- **Chapter I.** ***Crop and wild plants/weed species-associated viromes in a horticultural context: diversity, prevalence and stability over a two-year period***

  The INRA laboratory had initialed a multi-year study in 2010 with the objective to characterize the virome associated with crops, weeds and wild plant species in a horticultural setting and to analyze its stability over time by repeatedly sampling the plant populations in spring and fall over a two-year period. The sampling, sequencing and initial analysis of the results had been performed before my arrival, I continued these efforts, performing a more detailed, OTU-based analysis together with statistical analyses and drafting of the manuscript.

  The results obtained show that virus infection was common in the sampled species and identify a rich viral diversity of 245 OTUs representing 18 viral families and confirmed the dominance of novel viruses with dsRNA genomes. A key finding it that virome structure was relatively stable over time when considering the ratio of ssRNA *versus* dsRNA viruses and number of detected viral families but that it proved remarkably dynamic at the OTU level with a very minor proportion of OTUs consistently detected over the two-year period.

- **Chapter Ⅱ.** ***Comparison of double-stranded RNAs (dsRNAs) and Virion-associated nucleic acids (VANA) based metagenomic approaches for phytovirome analyses***

  As indicated above, a number of methodological questions relevant for phytovirome analyses are yet unexplored. One such aspect concerns the choice of the target nucleic acid population to be submitted to HTS. I directly compared the performance of the HTS analysis of highly purified dsRNAs and of VANA for phytovirome description in six cultivated or unmanaged sampling sites.

  The results obtained show that the dsRNA-based approach consistently revealed a broader and more comprehensive diversity for RNA viruses than the VANA approach, whatever the assessment criterion. This study also illustrated the power of the OTU-based approach to virus richness estimation. It allows to escape empirical choices by reasoning the methodological choices in phytovirome studies and, likely, in the study of other viromes.

- **Chapter Ⅲ.** ***Viromes and phyllosphere mycobiomes in complex plant samples and in complex fungal populations cultured from these plants***

  As discussed above, dsRNA viruses are diverse and abundant in plant viromes. However, the origin of these dsRNA viruses is still a matter of debate as they may represent either plant-infecting viruses or viruses infecting fungi associated with the sampled plants. In an effort to bring some further data in this debate, I analyzed in parallel the fungal (mycobiome) and viral populations associated with complex plant samples and with complex fungal pools cultivated from these plant samples.

  The results obtained showed that both plant-associated mycobiome and virome composition showed a strong site specificity. Diversity comparisons indicate that the mycobiome was more diverse in unmanaged sites while the plant-associated virome showed a higher family-level richness in cultivated sites, suggesting that mycobiome and virome are under the influence of different driving forces. Fungal culturomics captured ca. 10% of the fungal diversity but there was virtually no correlation between the virome directly obtained from plant samples and the mycovirome from fungal cultures.

- **Chapter Ⅳ.** ***Metagenomic analysis of virome cross-talk between cultivated Solanum lycopersicum and wild Solanum nigrum***

  There is still limited information on the extent and directionality of transfer of viruses between crops and weeds growing in close association with them. Using a metagenomics approach I explored virus diversity in a cultivated crop plant, tomato, and in a common botanically related weed, *Solanum nigrum* (european black nightshade). I also contrasted the virome of *S. nigrum* growing in close proximity to tomatoes with that of *S. nigrum* growing away from tomato, among unrelated crop species.

  The results obtained, while preliminary, show that a large variability in virome richness was observed but without a clear ability to link this to a particular host or to local conditions. While only 17.9% of OTUs were shared between tomato and nightshade, the assembly of very long contigs allowed a detailed population analysis for several viruses. In the case of potato virus Y (PVY), the results support a model of infection in nightshade resulting from virus spillover from tomato crops. Highly diverse broad wilt bean virus 1 (BBWV1) populations with potential genome reassortments were only detected from nightshade, suggesting the existence of barriers to the transfer of BBWV1 to tomato. A new ilarvirus infecting both plant species was characterized and tentatively named solanum nigrum ilarvirus 1. The results obtained provide information on the circulation of several viruses between these two *Solanum* species and enrich our knowledge of the tomato virome.

CHAPTER I

Crop and wild plants/weed species-associated viromes in a horticultural context: diversity, prevalence and stability over a two-year period

Crop and wild plants/weed species-associated viromes in a horticultural context: diversity, prevalence and stability over a two-year period

Yuxin Ma, Armelle Marais, Sébastien Theil[$], Marie Lefebvre, Laurence Svanella-Dumas, Chantal Faure, Bernard Bergey & Thierry Candresse*

UMR 1332 BFP, INRA, Univ. Bordeaux, CS20032, 33882 Villenave d'Ornon cedex, France

Abstract: 245 words

Text: 4994 words (Introduction + M&M + Results + Discussion)

* Corresponding author thierry.candresse@inra.fr

$ current address: Unité Mixte de Recherche sur le Fromage, 20 côte de Reyne, 15000 Aurillac, France

Cleaned virome sequence reads have been deposited on the INRA National Data Portal under the identifier https://doi.org/10.15454/5BLYMJ

Running title: Dynamics of horticultural phytoviromes

ABSTRACT

Using purified double-stranded RNAs (dsRNAs) and high-throughput sequencing (HTS) we analyzed the metavirome associated with crops and surrounding weeds/wild plants in a horticultural context in southwestern France. Temporal virome variations were analyzed by repeatedly sampling of the plant populations. In total, 126 species-specific composite samples representing 48 unique plant species were collected and analyzed over a two-year period. The obtained HTS sequence were annotated by Blast-based methods and classified into Operational Taxonomic Units (OTUs) representing a proxy to viral species. A rich viral diversity of 231 OTUs representing 18 viral families was identified. The largest group of viruses detected corresponds to novel viruses with dsRNA genomes. For ssRNA viruses, the proportion of novel viruses was only 48.1%. Virus infection was common, with 86.5% of the composite samples with at least one viral OTU. The number of unique OTUs increased linearly with the number of samples for a given plant species, indicating that the overall virome is likely to be much larger and that uncovering it may necessitate the analysis of hundreds of plants per species. Virome structure was relatively stable over time when considering the ratio of ssRNA versus dsRNA viruses and number of detected viral families. However, virome composition proved remarkably dynamic at the OTU level, with 68.8% of OTUs detected from a single sample and only 6 (2.6%) OTUs consistently detected over the 2-year period. The sharing of viral OTUs between crops and weeds was also analyzed, showing an over-representation of ssRNA viruses.

Key words: virome, metagenomics, richness, OTU, plant virus

INTRODUCTION

Large-scale environmental metagenomics studies, enabled by high-throughput sequencing (HTS) technologies, have revealed that viruses infecting prokaryotes are the most abundant biological entities on earth with an estimated number of $10^{30\text{-}31}$ species (Breitbart and Rohwer, 2005). Yet most of what we know about viral populations comes from marine environments and the human microbiome (Brum *et al.*, 2015; Reyes *et al.*, 2010). Recently a ground-breaking large-scale meta-transcriptomics study revealed the unprecedented diversity of invertebrate RNA viruses (Shi *et al.*, 2016). They discovered 1445 RNA viruses from 220 invertebrate species, some of which are sufficiently divergent to comprise new families, thus redefining the invertebrate virosphere. The same team recently also explored the vast diversity of vertebrate RNA viruses and proposed an evolutionary history for most concerned virus groups (Shi *et al.*, 2018).

Plant virus communities have been generally understudied as compared to animal viruses, and their diversity must correspondingly be largely underestimated. In fact, as early as a dozen years ago it was realized that our understanding of plant virus diversity is both limited and biased (Wren *et al.*, 2006) in particular because efforts in plant virology have been largely focused on disease-causing viruses in economically important cultivated crops. This lead to an over-representation of crop-infecting viruses, with over 77% of plant viruses then listed in the Viral Identification Data Exchange (VIDE) Database having been initially identified from cultivated crops (Wren *et al.*, 2006). Yet, cultivated plants account for only a minute fraction of all plant species, and virus infections are not always acute and associated with visible symptoms (Cooper and Jones, 2006).

After decades of studying plant viruses (Fargette *et al.*, 2006), it is now apparent that the emergence of new diseases following changes in viral host ranges is driven by adaptive viral evolution in response to novel ecological conditions (Jones, 2009; Lefeuvre *et al.*, 2019). In agroecosystems, crop plants often grow side by side with bordering wild plants and weeds. These wild plants may in turn constitute “reservoirs” for viruses that may be transferred to

cultivated plants by a wide array of vectors, leading to epidemies or to the emergence of novel viruses (Anderson *et al.*, 2004; Elena *et al.*, 2014; Pagán *et al.*, 2012; Power, 2008; Roossinck and García-Arenal, 2015). Conversely, epidemies developing in crops may spill over and impact wild plant populations. A much wider knowledge of virus richness, prevalence and dynamics in wild plant populations is therefore desirable in order to better understand virus epidemiology and virus emergence in crops.

Viruses in wild plant species have been reported to be diverse and often asymptomatic (Prendeville *et al.*, 2012; Roossinck, 2012). They potentially play important ecological roles in wild plant communities (Malmstrom *et al.*, 2011). The number of studies addressing viruses in wild plant populations is however still limited (Bernardo *et al.*, 2018; Fraile *et al.*, 2017; García-Arenal and Zerbini, 2019; Moreno-Pérez *et al.*, 2014; Susi *et al.*, 2019; Thapa *et al.*, 2015). These studies have generally shown that virus occurrence is quite common in wild plants, independently of the presence of symptoms, with as high as 70% prevalence and most identified viruses novel to science (Roossinck *et al*., 2010). Besides the huge viral variability thus revealed, these studies have also revealed variability in viral prevalence or in viral communities. For example, Susi *et al.* discovered variations not only in virus prevalence in 12 *Plantago lanceolata* populations but also in the virus communities present in these host populations (Susi *et al.*, 2019). These studies also highlighted the importance to accumulate information on virus communities in a wider range of wild plant species and to begin to understand the drivers shaping viral communities in crops and wild plants (Bernardo *et al.*, 2018; Fraile and Garcia-Arenal, 2016).

While the above mentioned studies have started to bring information about the spatial variation of viral communities there is to date very limited information on their temporal variation. In one of the few studies to date, Thapa *et al.* investigated over four years six wild plant species comprising 400 specimens collected from twenty sites in the Tallgrass Prairie Preserve in northeastern Oklahoma. The viromes were described and compared at host species, spatial or temporal scales,

and the results indicated that host identity has a significantly stronger effect on the virome composition other than location and sampling year (Thapa *et al.*, 2015).

Another area of virome studies that has to date received incomplete attention concerns the methods used to access viral richness at a refined taxonomic level, closer to viral species, which would allow to precisely compare the virus communities from different environments. Unlike bacteria and fungi, viruses lack universal gene markers to facilitate community surveys, therefore random whole genome amplification (WGA) after the enrichment of virus-associated nucleic acids has so far been the method of choice in virus metagenomics studies (Marais *et al.*, 2018; Roossinck *et al.*, 2010). The taxonomic assignation of the viral contigs identified among the HTS data face many challenges that have been reviewed in detail by (Simmonds, 2015). Besides the fact that genomes are frequently incompletely assembled, different genomic regions of viral genomes may have different evolutionary origins while species distance discrimination criteria vary between viral families. Given the large number of novel agents uncovered in metagenomics studies, the most widely used approach, Blast-based annotation, generally provides unreliable results at the species and genus level and still has weaknesses at family level (Roossinck *et al.*, 2010; Simmonds, 2015). A consequence is that most studies to date have either addressed viral richness at family level. For plant viruses for example, Thapa *et al.* (2015) combined viruses into relatively broad taxonomic categories at family and genus level using a consistent process for all samples (Thapa *et al.*, 2015). Bernardo *et al.* (2018) assessed virus prevalence and diversity at family level but, to avoid over-counting for some analyses, grouped virus-like sequences into Operational Taxonomic Units (OTUs) on the basis of pairwise sequence similarity. To potentially improve this situation, Simmonds (2015) proposed a strategy based on the use of conserved, informative genome regions for the clustering of viral sequences and the definition of OTUs that can be designed to mimic taxonomic levels below the family taxon level.

In the present study we explored virus diversity in cultivated plants and in neighboring wild

plants/weeds using HTS-based metagenomics. To improve the taxonomic assignation of viral sequences, we combined the classic blast-based annotation approach and an OTU-based approach following the strategy followed by Simmonds (2015) and implemented in an automated pipeline (Lefebvre *et al.* submitted for publication). The results obtained provide information on virus prevalence in the sampled plant populations and allow to describe some general virome properties. The repeated sampling over a two-year period showed a surprisingly dynamic virome at the OTU level.

MATERIALS AND METHODS

Study sites, plant samples and pooling strategy

The two study sites are two horticultural plots and their immediate surroundings in Villenave d'Ornon (VO), and Bergerac (BE) in Southwest France. The VO site has mixed horticultural productions, including lettuce, radish, spinach etc, while the BE site has only tobacco with some neighboring corn. The majority of samples collected are dicotyledonous wild plants and weeds but samples of the crop plants were also collected (Table S1). The identification of plants samples was performed down to species level (or in a few cases, only to genus level) by a scientist with experience in botany. Besides the crops, in each site and at each time point, samples of the dominant plant species were collected. Monocotyledonous species were not sampled and no specific efforts were made to collect symptomatic plants. However, plants with obvious necrosis of with insect colonization were not collected. For each sampled plant species 15 individual plants (100 mg each) were collected (in a few cases, less than 15 plants could be collected, Table S1). Pools of leaf tissues were thus constituted for each sampled species and dried over anhydrous $CaCl_2$. In parallel, samples of the individual plants were similarly conserved. The VO site was sampled twice a year (spring and autumn) over a 2-year period (2010-2011), while the BE site was only sampled twice, in spring and fall of 2010.

Double-stranded RNA (dsRNA) purification, amplification and pyrosequencing

dsRNAs were purified by two rounds of CF11 chromatography according to the protocol described in Marais *et al.* (2018) from each of the species-specific pools described above. For library preparation, 3 µl of purified dsRNAs were denatured during 5 min at 99°C and submitted to a reverse transcription initiated by a mixture of primers consisting of 1 µM dT18 and 2 µM PcDNA12 (5' TGTGTTGGGTGTGTTTGGN$_{12}$ 3') using the SuperscriptII Reverse Transcriptase according to the manufacturer's instructions (Invitrogen). Complementary DNAs (cDNAs) were used as templates for a whole-genome amplification (WGA) procedure (Marais *et al.*, 2018), allowing at the same time their conversion to double-stranded cDNAs and their tagging with multiplex identifier (MID) adaptors. The 10-bp MID tags used can tolerate up to two sequencing errors and still allow reliable demultiplexing of samples. PCR products were purified using the MinElute PCR Purification Kit (Qiagen, Courtaboeuf, France) and their concentration determined spectrophotometrically before being pooled and analyzed in a multiplexed format on a Roche 454 GS FLX Titanium sequencer at the GenoToul platform (INRA Toulouse, France).

Reads cleaning, contigs assembly and annotation, OTU classification

For each library, sequencing reads were first demultiplexed in order to assign individual reads to the relevant plant sample. The adaptors containing the MID tags were removed from the reads, which were then trimmed on quality and length using the virAnnot pipeline (Lefebvre *et al.*, submitted for publication). Clean reads were assembled *de novo* using Newbler (http://454.com/contact-us/software-request.asp) with default parameter settings. The annotation of contigs and singletons was performed using BlastN and BlastX with an e-value cut-off of 10^{-4}. Viral contigs were assigned to a viral family on the basis of the first Blast hit.

Viral contigs and singletons were also classified into Operational Taxonomical Units (OTUs) using the virAnnot pipeline (Lefebvre *et al.*, submitted for publication). Briefly, RPS-Blast (Marchler-Bauer *et al.*, 2002) against the Pfam database (Punta *et al.*, 2012) was used to detect

sequences encoding conserved viral protein motifs, in particular those corresponding to the families of RNA-dependent RNA polymerases (RdRp1, 2, 3 and 4; Koonin, 1991). The identified motifs were then aligned and a clustering analysis allowed to group together in the same OTU sequences that share more than 90% amino acid identity (Figure S1). By comparing virAnnot OTUs with ICTV taxonomy, this 10% threshold value has been validated as providing OTUs that approximate, in different families, the ICTV species level, allowing to use such OTUs as a proxy to taxonomic species (Lefebvre *et al*., submitted for publication). The annotation of OTUs at viral family level was performed on the basis of first Blast hit of a representative contig. OTUs were named using the following scheme: first four letters of family name plus OTU number (ex. BROM_001, ALPH_001…).

Removal of false positives due to inter-sample cross-talk according to RT-PCR validation assays

At the end of OTU classification, an OTU table giving the number of reads for each OTU in the different samples analyzed was obtained. The possibility of low level false positives in the OTU table due to the creation of hybrid reads and other sequencing artifacts such as Index hoping (Illumina, 2017; van der Valk *et al.*, 2019) was experimentally addressed by performing reverse-transcriptase PCR (RT-PCR) validation for a total of 82 samples for 8 OTUs. The primers were designed according to the 454 sequences and, when available, to reference sequences in the NCBI database. The RT-PCR results showed that confirmation of an OTU presence in a sample was generally not achieved (only 33% of tests) for read counts of 2 or less but that it was almost systematically obtained (97.7% of tests) for reads counts of 3 or more. A threshold of at least 3 OTU-related reads was therefore used to remove the potential false positives in the original OTU table (Figure S1) when an OTU had been detected in multiple samples (and with therefore a risk of inter-sample cross-talk). After filtering with this threshold, 370 (43%) of 863 positive

OTU/sample combination were discarded. In total, 231 RdRp OTUs with 392 occurrences for the VO site and 81 OTU with 101 occurrences for the BE site were thus retained.

Phylogenetic analyses

Multiple alignments of nucleotide or amino acid sequences were performed using the ClustalX program (Thompson *et al.*, 1997) as implemented in Mega 7.0 (Kumar *et al.*, 2016). Pairwise strict nucleotide and amino acid distances were computed using Mega 7.0 and phylogenetic trees were reconstructed using the neighbor joining method in Mega 7.0.

RESULTS

Overview of the datasets

For the VO sampling site, which was sampled in spring and autumn of 2010 and 2011, a total of 126 composite plant samples (1806 individual plants) was collected, corresponding to a total of 48 unique dicotyledonous plant species, representing 22 families in 13 orders (Apiales, Asterales, Brassicales, Caryophyllales, Cucurbitales, Fabales, Geraniales, Lamiales, Malpighiales, Malvales, Ranunculales, Rosales and Solanales) (Table S1). In the BE sampling site a total of 31 composite samples was collected in 2010, representing 13 plant families in 10 orders (data not shown). The majority of sampled wild/weed species have not previously been screened for viruses.

After demultiplexing and trimming, a total of about 0.5 million clean reads corresponding to 133.7 megabases were generated and for each sampling date, the number of clean reads ranged from 51,573 to 158,533 (Table 1). The minimal reads length was set at 60 nucleotides (nt). The average read length was 265 nt. Out of the total clean reads, 75,181 reads were singletons while 419,951 (84.34%) were incorporated into contigs. In total, 4,034 contigs were assembled with a length ranging from the cut-off of 100 nt to 15,503 nt (Figure S2A). The average contig size was 535 bp and the N50 682 bp. Depending on the sample, the number of reads ranged from 231 to 19,468 with an average of 3,172 reads (Figure S2B).

Diversity of RNA viruses at the two sampling sites and their phylogenetic relationships

As indicated in the Materials and Methods section, the assembled contigs and singletons were classified in Operational Taxonomic Units (OTUs) on the basis of the RNA-dependent RNA polymerase conserved domain, using the virAnnot pipeline (Lefebvre *et al*., submitted for publication). Using an OTU cut-off criterion of 10% nt or aa divergence, a total of 300 unique RdRp OTUs were defined from both VO and BE sites and assigned to known virus families using BlastN and BlastX analysis of representative sequences. Over the two-year sampling period, a variety of viruses corresponding to a total of 231 viral RdRp OTUs were identified at the VO sampling site (Table S2), representing 18 viral families including 14 ssRNA families (52 OTUs), 4 dsRNA families (171 OTUs) and some unclassified OTUs (Figure 1 and Figure 2). A phylogenetic analysis performed on representative sequences of all OTUs and on reference viruses for each viral family confirmed the quality of the annotations obtained in this study, since very generally OTUs clustered with the reference viruses representative of the family to which they had been assigned (Figure 1). This results indirectly validates the use of the 100 aa region around the RdRp conserved motif to accurately assign sequences into each viral family. Some unclassified/unassigned OTUs clearly clustered in the *Partitiviridae* and *Totiviridae* family (Figure 1) highlighting the incomplete taxonomic annotation of some likely members of these families in NCBI database.

In this analysis, the OTUs in family *Partitiviridae* were separated into 2 clades, which prompted a detail phylogenetic analysis performed with additional reference sequences and the OTUs identified in the present study (Figure S3). The results obtained show that the the vast majority of the OTUs clustered into one of the four genera identified in the family (Figure S3). The plant-specific *Deltapartitiviridae* (Nibert *et al.*, 2014) clustered the largest number of OTUs (n=13), followed by the fungus-specific *Gammapartitiviridae* (n=6), the plant or fungus-infecting *Alphapartitiviridae* (n=4), and the *Betapartitiviridae* (n=2). In total, 5 OTUs were found to be

nearly identical, in the sequenced region, with known *Partitiviridae* and to represent the detection of beet cryptic viruses 1 and 2, Raphanus sativus cryptic viruses 1 and 2 and spinach cryptic virus 1 (Figure S3). According to the plant- or fungi-infecting status of their closest relative in this phylogenetic analysis 21 of the 34 Partitiviridae OTUs are assumed to be plant-associated while 13 are expected to be associated with fungi (Figure S3). This tentative analysis is indirectly supported by the number of reads integrated in the various OTUs, which tends to be higher for the plant-associated *Partitiviridae* (data not shown).

dsRNA viruses: diverse and largely new

Similar to the analysis performed above for the *Partitiviridae*, it was possible to classify OTUs between known or novel ones depending on whether a viral sequence clustering with the 10% cut-off envelope could be identified for each OTU in the Genbank database (Table S2 and Figure 2). In the case of ssRNA viruses, 27 OTUs (51.9%) are thus considered to correspond to known agents and collectively belong to 10 ssRNA families or to represent one unassigned Picornavirales and 3 virus-associated RNAs (Table S3). For some of these viruses, the results of the present study represent the detection of new host plants, in particular in weeds (Table S3). On the contrary, only 11 dsRNA viruses OTUs (6.4%) correspond to known viruses in the *Partitiviridae, Totiviridae, Amalgaviridae* and *Chrysoviridae* (Figure 2 and Table S2). The proportion of known versus novel viruses is thus very different between dsRNA and ssRNA OTUs.

Virome overall structure is relatively stable over time

In the four sampling periods (June 2010, October 2010, June 2011 and October 2011) respectively 66, 90, 64, and 103 OTUs were identified, from respectively 16, 31, 38 and 41 plant samples (Table 1). At each of the four sampling dates, dsRNA OTUs accounted for the largest proportion, ranging from ca. 65.6% to 75.8% while ssRNA OTUs ranged from ca. 21.2% to 31.3% (Figure 3). The proportion of dsRNA OTUs appears therefore relatively stable over time. The more limited

sampling performed at the BE site provides a similar message: the proportions of dsRNA OTUs were respectively ca. 76.2% and 72.0% in the spring and fall sampling periods and for ssRNA OTUs respectively ca. 16.7% and 20.0%.

At the family level, with a total of 14 families ssRNA viruses systematically represented a higher diversity than dsRNA viruses (4 families) (Figure 3). Depending on the sampling period, viruses in the *Potyviridae*, *Tombusviridae*, *Endornaviridae* and *Betaflexiviridae* accounted for a relatively large proportions of ssRNA OTUs, whereas *Partitiviridae* and *Totiviridae* OTUs clearly dominated the dsRNA ones, accounting for 45.3%-62.1% of total OTUs depending on the sampling period (Figure 3). The more limited sampling at the BE site yielded parallel results, with respectively seven and eight families identified from 15 (spring) and 16 (autumn) composite plant samples. 15 ssRNA OTUs can be assigned to 5 known families (*Solemoviridae*; *Bromoviridae*; *Closteroviridae*; *Potyviridae* and *Endornaviridae*) and two virus-associated RNAs. Fifty six dsRNA OTUs can be also assigned to the same four known viral families. Family *Totiviridae* comprised the majority of OTUs with a proportion ranging from 40.5% to 48% depending on the season. These results highlight a stability in overall virome composition over time in terms of the proportion of ssRNA and dsRNA OTUs and also some prominent group of OTUs such as *Totiviridae*.

High virus prevalence in the sampled plant populations

In the VO sampling site, among the 126 composite plant samples analyzed, only 17 (13.5%) (Figure S4A) were found to be free of viruses, with no detected viral OTU. A similar value (13%) was observed at the BE site. The average number of OTUs per sample is 3.11 +/-2.88, the high standard deviation value reflecting the large diversity in the number of OTUs identified per sample (Figure S4A and Table S1). A composite *Trifolium repens* (white clover) sample collected in June 2010 had the highest number of OTUs (n=14; Table S1).

When aggregating OTU data for plant species that have been sampled multiple times (from 2 to 4 times) the average number of OTUs per plant species increases to 6.9 +/- 4.5 (Figure S4B) with the richest viral diversity found in *Malva sylvestris* (mallow) with 22 OTUs followed by *Brassica rapa* (turnip) with 17 OTUs (Table S4). No plant species repeatedly sampled was found free of viral infection (Figure S4B), indicating that repeated sampling of the same plant species allows to uncover a larger viral diversity. Indeed, when considering the seven plant species that were resampled 4 times, the aggregated number of unique OTUs increased linearly with the number of samplings (Figure S5) with no sign of reaching a plateau, indicating that the overall virome of these species is likely much larger than what was identified here.

A large fraction of viruses is only detected from a single plant species

Out of the 231 detected OTUs, 68.8% (n=159) were detected from a single sample, and 31.2% (n=72) were detected from multiple samples (Figure S4C). On average, each OTU had 1.70 +/- 1.8 host plants, a notable exception being BROM_001 (*Cucumber mosaic virus*) which was detected from 20 samples corresponding to 15 plant species (Figures S4C and S4D). From a plant species perspective, on average an OTU had 1.4 +/- 1.3 host plant species (Figure S4D). However, this overall value masks the fact that a large proportion of OTUs (80.1%, n=185) were identified from a single species while 19.9% (n=46) were found to be promiscuous and to have multiple host species.

OTU virome composition is highly dynamic over time

While only 10.4% of sampled plant species were collected only once, 72.7% of OTUs were detected only once, providing a first indication that the virome is much more unstable than the sampled plant populations (Figure 4). In parallel, only 2.6% of OTUs (n=6) were detected at all sampling dates (not necessarily from the same host) (Figure 4). These 6 OTUs are AMAL_002 (*Lactuca sativa*; *Spinacia oleracea*), CHRY_013 (*Conyza canadensis*; *Malva sylvestris*),

PART_026 (*Malva sylvestris*; *Matricaria inodora*; *Persicaria maculosa*; *Sysimbrium officinale*; *Papaver dubium*; *Solanum lycopersicum*), PART_029 (*Raphanus sativus*; *Amaranthus retroflexus*; *Beta vulgaris*; *Sysimbrium officinale*), TOTI_072 (*Malva sylvestris*; *Geranium rotundifolium*; *Capsella bursa-pastoris*; *Coronopus didymus*; *Portulaca oleracea*; *Veronica persica*; *Cerastium sp.*; *Papaver dubium*; *Petroselinum crispum*; *Convolvulus sepium*) and POTY_001 (*Raphanus raphanistrum*; *Brassica rapa*; *Urtica urens*). They correspond respectively to five novel dsRNA viruses and to *Turnip mosaic virus* (Table S2). It is also noteworthy that all of these OTUs were detected from multiple plant species, which may have contributed to their ability to persist at the sampling site over extended periods of time.

Focusing on the 14.6% of the plant species (n=7) that were sampled four times over the two-year period (Figure 4; *Lactuca sativa, Sonchus asper, Coronopus didymus, Amaranthus retroflexus, Spinacia oleracea, Malva sylvestris* and *Datura stramonium*), they also showed a highly dynamic virome composition, with the great majority of OTUs (n=45, 81.8%) detected only once. Only a single OTU (AMAL_002) was detected through all samplings in the same host, spinach.

Virome cross-talk between crops and weeds/wild plant species

During the two-year sampling period, 3, 6, 6 and 9 crop species were collected respectively in June 2010, October 2010, June 2011 and October-2011 (Table 1). At each time point, between 5.8% to 7.8% (5~6 OTUs) of OTUs were shared between crops and surrounding weeds/wild plants (Figure 5), including both ssRNA viruses (*Closteroviridae*, *Bromoviridae*, *Potyviridae* and *Betaflexiviridae*) and dsRNA viruses (*Partitiviridae* and *Totiviridae*). The majority of the shared ssRNA OTUs correspond to known viruses such as cucumber mosaic virus, beet yellows virus, Apium virus Y and potato virus Y. Similar results were obtained from the BE site, in which only 3 OTUs (6%) were shared between 2 crop species and 14 wild plant/weed species, including cucumber mosaic virus and potato virus Y.

DISCUSSION

The temporal variation of phytoviromes has so far been rarely studied. Here, we investigated the viromes of both crop and bordering wild plants/weed species in two horticultural sampling sites over a two-year period. The virus prevalence in 126 composite samples was 86.5%, which is relatively close to the 70% value recorded in samples from Tall Grass Prairie Preserve and from the Conservacion Guanacaste area in Costa Rica (Roossinck *et al.*, 2010) but much higher than the 25.8% to 35.7% reported from high biodiversity areas in France and South Africa (Bernardo *et al.*, 2018). This difference very likely reflects differences in sampling strategy and in prevalence calculation method since most samples analyzed here pooled 15 individual plants when Bernardo *et al*. used samples composed of one or a few individual plants when and considered only plant-associated viruses, eliminating *Totiviridae* and *Chrysoviridae* members in their assessment (Bernardo *et al*., 2018). It is noteworthy that even when discounting *Totiviridae* and *Chrysoviridae* members, high prevalence values of respectively 77% and 67.7% are observed for the VO and BE sites, suggesting that the size of the sampled plant populations is likely the most important parameter, an hypothesis supported by the observation that repeated sampling of the same species lead to a linear augmentation of the aggregated virome (Figure S5).

This observation suggests that the virome of individual plant species is likely much larger than the viromes documented here, and that uncovering it may necessitate the analysis of hundreds if not thousands of individual plants in a variety of sites and over different time intervals. Indeed, even with the pooling strategy used here, only viruses with a relatively high prevalence in the sampled plant populations, on the order of 10%, would have relatively high probability of being captured. For example, a virus with 5% prevalence would only have close to a one in two chances of being detected, this value falling to only 14% for a virus with a 1% prevalence. The conclusions drawn here, as in most other previous studies are therefore possibly only valid for high prevalence viruses, a limitation that is rarely commented. It should also be noted that the extremely high correlation

between the number of OTUs detected and the number of sampled plants is not paralleled by a correlation between sequencing depth and the number of detected OTUs (data not shown). A consequence is that overcoming limitations in virome completeness can be best addressed by increased sampling rather than by increasing sequencing depth.

Similar to other phytovirome studies (Bernardo *et al.*, 2018; Thapa *et al.*, 2015), it was possible to identify already known viruses through their high similarity Blast scores. Remarkably, the ratio of known to novel proved very different between dsRNA and ssRNA viruses. This likely reflects their different lifestyles, with dsRNA viruses being very frequently persistent, symptomless viruses (Prendeville *et al.*, 2012, Roossinck, 2012; Roossinck, 2015) and having been neglected as compared to more frequently pathogenic, acute/chronic ssRNA viruses.

In the present study, viral diversity was characterized on the basis of Operational Taxonomic Units defined on the basis of the conserved RdRp motifs that are shared by all RNA viruses (Koonin, 1991). The virAnnot pipeline (Lefebvre *et al*., submitted for publication) allows the automated definition of OTUs and with suitable data normalization, precise comparisons of viral diversity between samples (alpha and beta diversity). In the present study, close to 70% of OTUs were only detected from a single composite sample. This is parallel to the results of Thapa *et al.*, who showed that only six out of 30 genus- or family-level OTUs had average incidences of 5 percent in host plants (Thapa *et al.*, 2015) and suggests that a large fraction of plant viruses may be specialists with narrow host ranges. This specialization may minimize competition between the different viral species and may result in highly polymorphic viral populations in complex environments (Lefeuvre *et al.*, 2019; Stroud and Losos, 2016). However, a few generalist viruses were also identified, among which cucumber mosaic virus which was detected from over 15 plant species and is known to have one of the widest host range among plant viruses (Jacquemond, 2012; Scholthof *et al.*, 2011). Such generalist viruses were reported to have access to a larger array of resources but compete with other viruses which is expected to results into low-diversity

populations dominated by one or a few of the best adapted viral genotypes (Stroud and Losos, 2016).

The virus cross-talk between crops and bordering wild plants/weeds was relatively stable between different time points, ranging from 5.8 to 7.8% of total OTUs. ssRNA OTUs over-represented in this shared fraction of the virome. Once again this may reflect broad differences in lifestyles between these two groups, and the fact that acute/chronic ssRNA viruses frequently have vectors (Roossinck, 2015) that would facilitate their movement between host plants. Indeed, in this study 36.5% (19/52) ssRNA OTUs were observed to have multiple host species and 52.6% (10/19) of these ssRNA generalist OTUs are known viruses, values much higher than the 15.2% (26/171) observed for dsRNA viruses OTUs of which 19.2% (5/26) are known viruses. These contrasted results indicate an over-representation of ssRNA viruses among generalist OTUs.

The proportion of ssRNA to dsRNA viruses OTUs proved relatively stable over time and between the two sampling sites. This observation also parallels the results of other studies (Thapa *et al*., 2015; Bernardo *et al*., 2018), suggesting that this may be a more general feature of phytoviromes. Whether this is truly general or applies only under specific circumstances remains however to be further investigated. However, the virome composition remarkably varied over time at the OTU level despite its relative stability when considering only viral families. This dynamic situation has also been documented in other virome studies and could be caused by landscape heterogeneity due to the human activities, which by affecting the host plant populations may indirectly affect virome composition (Power, 2008; Rodelo-Urrego *et al.*, 2013). Given the incompleteness of the viromes documented here, it is not possible to know whether the dynamic changes reflected here represent presence/absence changes or mere changes in prevalence. Larger scale efforts will clearly be needed to resolve that important issue.

ACKNOWLEDGEMENTS

The authors would like to thank Carole Couture, Sandy Contreras and Yec'han Laizet for sampling and sequencing data preprocessing, A. Raoult, and F. Villeneuve (Centre Technique Interprofessionnel des Fruits et Légumes, Lanxade), and F. Dorlhac de Borne (Imperial Tobacco, Bergerac) for access to sampling sites. We also thank the Genotoul Platform (INRA, Toulouse, France) for the Illumina sequencing, and China Scholarship Council for YM support during her PhD.

REFERENCES

Anderson PK, Cunningham AA, Patel NG, Morales FJ, Epstein PR, Daszak P (2004). Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. Trends Ecol Evol 19: 535-544.

Breitbart M, Rohwer F (2005). Here a virus, there a virus, everywhere the same virus? Trends Microbiol 13: 278-284.

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

Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S, Cruaud C, De Vargas C, Gasol JM (2015). Patterns and ecological drivers of ocean viral communities. Science 348: 1261498.

Cooper I, Jones RA (2006). Wild plants and viruses: under-investigated ecosystems. Adv Virus Res 67: 1-47.

Elena SF, Fraile A, Garcia-Arenal F (2014). Evolution and emergence of plant viruses. Adv Virus Res 88: 161-191.

Fargette D, Konate G, Fauquet C, Muller E, Peterschmitt M, Thresh JM (2006). Molecular ecology and emergence of tropical plant viruses. Annu Rev Phytopathol 44: 235-260.

Fraile A, Garcia-Arenal F (2016). Environment and evolution modulate plant virus pathogenesis. Curr Opin Virol 17: 50-56.

Fraile A, McLeish MJ, Pagán I, González-Jara P, Piñero D, García-Arenal F (2017). Environmental heterogeneity and the evolution of plant-virus interactions: Viruses in wild pepper populations. Virus Res 241: 68-76.

García-Arenal F, Zerbini FM (2019). Life on the Edge: Geminiviruses at the Interface Between Crops and Wild Plant Hosts. Annual Review of Virology 10.1146/annurev-virology-092818-015536.

Illumina (2017). Effects of index misassignment on multiplexing and downstream analysis https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf.

Jacquemond M (2012). Cucumber mosaic virus. Advances in virus research. Elsevier. pp 439-504.

Jones RA (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: 113-130.

Kumar S, Stecher G, Tamura K (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33: 1870-1874.

Lefeuvre P, Martin DP, Elena SF, Shepherd DN, Roumagnac P, Varsani A (2019). Evolution and ecology of plant viruses. Nature Reviews Microbiology 10.1038/s41579-019-0232-3.

Malmstrom CM, Melcher U, Bosque-Pérez NA (2011). The expanding field of plant virus ecology: Historical foundations, knowledge gaps, and research directions. Virus Res 159: 84-94.

Marais A, Faure C, Bergey B, Candresse T (2018). Viral Double-Stranded RNAs (dsRNAs) from Plants: Alternative Nucleic Acid Substrates for High-Throughput Sequencing. Methods Mol Biol 1746: 45-53.

Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 30: 281-283.

Moreno-Pérez MG, Pagán I, Aragón-Caballero L, Cáceres F, Fraile A, García-Arenal F (2014). Ecological and genetic determinants of Pepino Mosaic Virus emergence. J Virol 88: 3359-3368.

Nibert ML, Ghabrial SA, Maiss E, Lesker T, Vainio EJ, Jiang D, Suzuki N (2014). Taxonomic reorganization of family Partitiviridae and other recent progress in partitivirus research. Virus Res 188: 128-141.

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: e1002796.

Power AG (2008). Community Ecology of Plant Viruses. In: Roossinck MJ (ed). Plant Virus Evolution. Springer Berlin Heidelberg: Berlin, Heidelberg. pp 15-26.

Prendeville HR, Ye X, Morris TJ, Pilson D (2012). Virus infections in wild plant populations are both frequent and often unapparent. Am J Bot 99: 1033-1042.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K et al (2012). The Pfam protein families database. Nucleic Acids Res 40: D290-301.

Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466: 334.

Rodelo‐Urrego M, Pagán I, González‐Jara P, Betancourt M, Moreno‐Letelier A, Ayllón M, Fraile A, Piñero D, García‐Arenal F (2013). Landscape heterogeneity shapes host‐parasite interactions and results in apparent plant–virus codivergence. Mol Ecol 22: 2325-2340.

Roossinck MJ, Saha P, Wiley GB, Quan J, White JD, Lai H, Chavarria F, Shen GA, Roe BA (2010). Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol Ecol 19: 81-88.

Roossinck MJ (2011). The big unknown: plant virus biodiversity. Curr Opin Virol 1: 63-67.

Roossinck MJ (2012). Plant Virus Metagenomics: Biodiversity and Ecology. Annu Rev Genet 46: 359-369.

Roossinck MJ (2015a). Metagenomics of plant and fungal viruses reveals an abundance of persistent lifestyles. Front Microbiol 5.

Roossinck MJ (2015b). Plants, viruses and the environment: Ecology and mutualism. Virology 479–480: 271-277.

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

Scholthof KBG, Adkins S, Czosnek H, Palukaitis P, Jacquot E, Hohn T, Hohn B, Saunders K, Candresse T, Ahlquist P (2011). Top 10 plant viruses in molecular plant pathology. Mol Plant Pathol 12: 938-954.

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

Shi M, Lin X-D, Chen X, Tian J-H, Chen L-J, Li K, Wang W, Eden J-S, Shen J-J, Liu L, Holmes EC, Zhang Y-Z (2018). The evolutionary history of vertebrate RNA viruses. Nature 556: 197-202.

Simmonds P (2015). Methods for virus classification and the challenge of incorporating metagenomic sequence data. J Gen Virol 96: 1193-1206.

Stroud JT, Losos JB (2016). Ecological opportunity and adaptive radiation. Annu Rev Ecol Evol Syst 47.

Susi H, Filloux D, Frilander MJ, Roumagnac P, Laine A-L (2019). Diverse and variable virus communities in wild plant populations revealed by metagenomic tools. PeerJ 7: e6140.

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

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25: 4876-4882.

van der Valk T, Vezzi F, Ormestad M, Dalén L, Guschanski K (2019). Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. Mol Ecol Resour https://doi.org/10.1111/1755-0998.13009.

Wren JD, Roossinck MJ, Nelson RS, Scheets K, Palmer MW, Melcher U (2006). Plant Virus Biodiversity and Ecology. PLoS Biol 4: e80.

Table 1. Summary of plant samples analyzed at the two sampling sites and time points and viral metagenome characteristics.

	VO site				BE site	
	JUN 2010	OCT 2010	JUN 2011	OCT 2011	JUN 2010	OCT 2010
Cultivated crops sampled	3	6	6	9	1	2
Wild plants/weeds sampled	13	25	32	32	14	14
Total no. of plant samples	16	31	38	41	15	16
Total no. of individual plants*	239	382	570	615	174	204
No. of clean reads	82195	51573	158533	124753	60955	19917
No. of OTU-associated reads	6999	5865	6235	8381	1687	2725
No. of OTUs	66	90	64	103	42	50
No. of viral families	12	12	12	12	7	8
No. of infected samples$	15	25	32	37	13	14
% of infected samples	93.7%	80.6%	84.2%	90.2%	86.7%	87.5%

* A few collected composite samples involved less than 15 individual plants.
$ Infected samples correspond to plant samples from which at least one viral OTU was identified.

Legend to the Figures

Figure 1. Phylogenetic tree reconstructed using the amino acid sequences of a 100 amino acid long region surrounding "GDD" conserved RdRp motif for tentative OTUs and for reference viral sequences obtained from Genbank. Tree was constructed using the neighbor-joining method and a strict identity distance. The statistical significance of branches was evaluated by bootstrap analysis (1,000 replicates). Only bootstrap values above 50% are indicated. The scale bar represents 10% amino acid divergence. Solid circles represent OTUs belonging to ssRNA families, solid squares represent OTUs belonging to dsRNA families, grey triangles unclassified/unassigned OTUs. Black triangles indicate reference sequences obtained from GenBank.

Figure 2. Proportion of OTUs corresponding to know viruses and to putatively novel ones. The inner circle indicates the genome type of OTUs and the corresponding number: dsRNA OTUs (171), ssRNA OTUs (52), and unclassified OTUs (8). The number of OTUs corresponding to known or novel viruses in each category is shown on the outer doughnut. Green color indicates OTUs corresponding to known viruses, grey color OTUs of novel viruses (>10% aa divergence in the conserved RdRp region with known viruses).

Figure 3. Proportion of OTUs belonging to different viral families identified during the four sampling periods at the VO sampling site. The legends in blue correspond to ssRNA virus families, those in red to dsRNA virus families and grey to unclassified/unassigned viruses for which no genome type information is available.

Figure 4. Bar chart showing the proportion of plant species and of viral OTUs that were sampled or detected once, two time, three times or four times during the 2-year sampling period.

Figure 5. Venn diagrams showing for each sampling time at the VO site the number of unique OTUs identified in crop species, in wild plants/weed species or shared between crops and wild plants/weed species. The names of the shared OTUs are indicated.

Figure 1

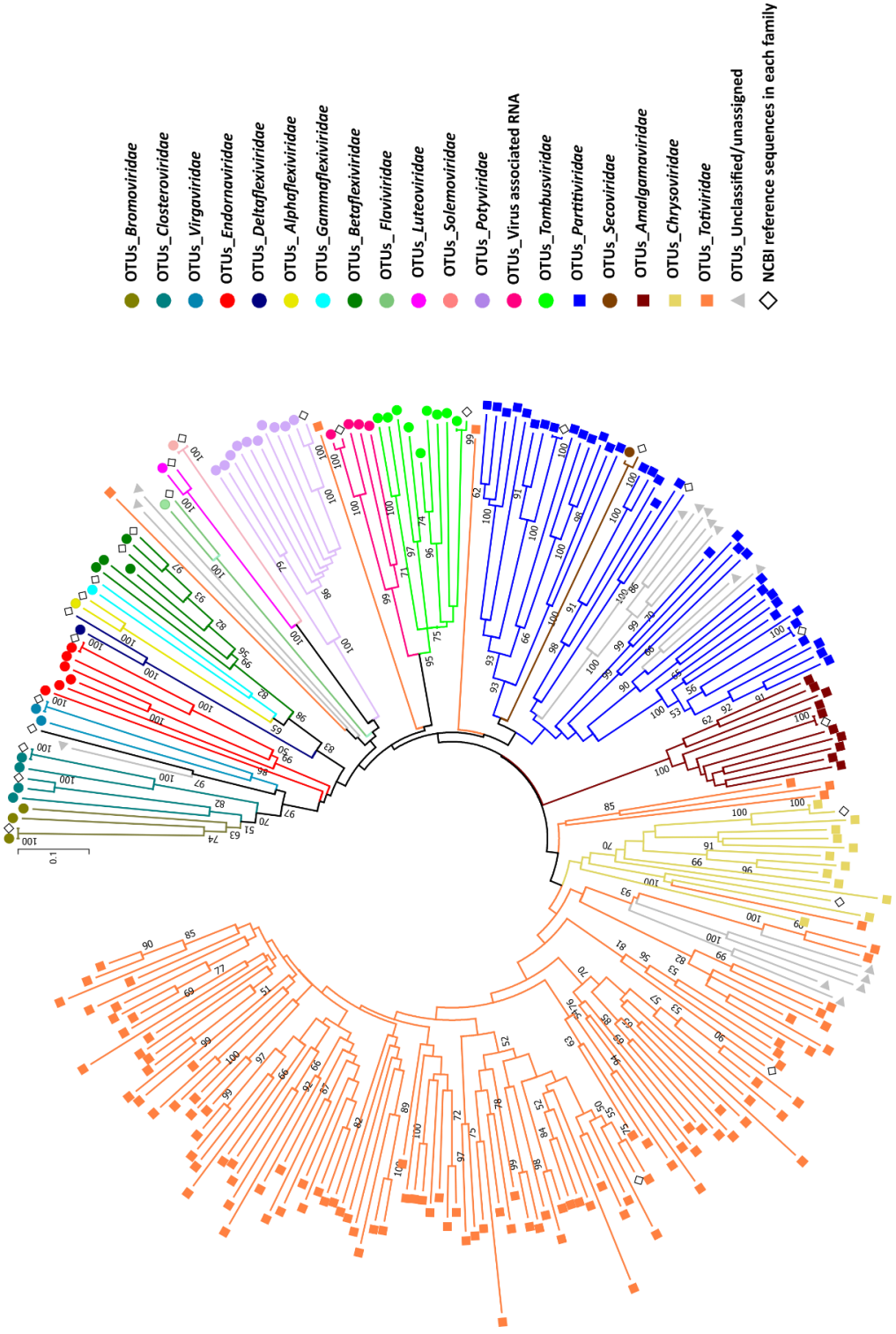

OTUs_Bromoviridae
OTUs_Closteroviridae
OTUs_Virgaviridae
OTUs_Endornaviridae
OTUs_Deltaflexiviridae
OTUs_Alphaflexiviridae
OTUs_Gammaflexiviridae
OTUs_Betaflexiviridae
OTUs_Flaviviridae
OTUs_Luteoviridae
OTUs_Solemoviridae
OTUs_Potyviridae
OTUs_Virus associated RNA
OTUs_Tombusviridae
OTUs_Partitiviridae
OTUs_Secoviridae
OTUs_Amalgamaviridae
OTUs_Chrysoviridae
OTUs_Totiviridae
OTUs_Unclassified/unassigned
NCBI reference sequences in each family

Figure 2

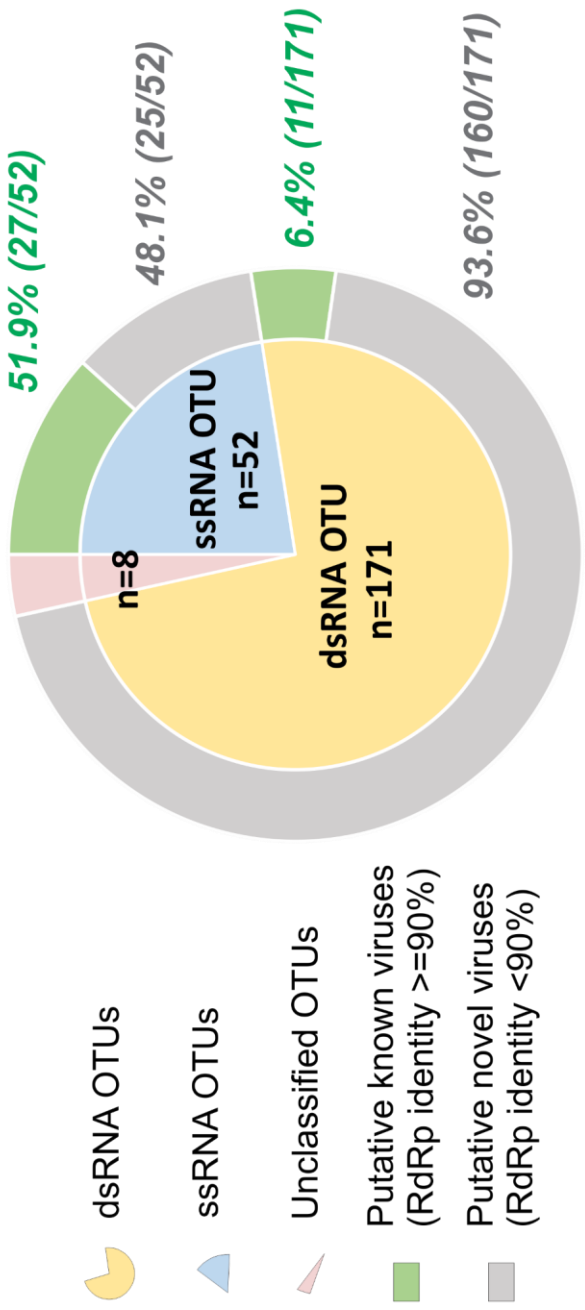

51.9% (27/52)
48.1% (25/52)
6.4% (11/171)
93.6% (160/171)
n=8
ssRNA OTU
n=52
dsRNA OTU
n=171
dsRNA OTUs
ssRNA OTUs
Unclassified OTUs
Putative known viruses
(RdRp identity >=90%)
Putative novel viruses
(RdRp identity <90%)

Figure 3

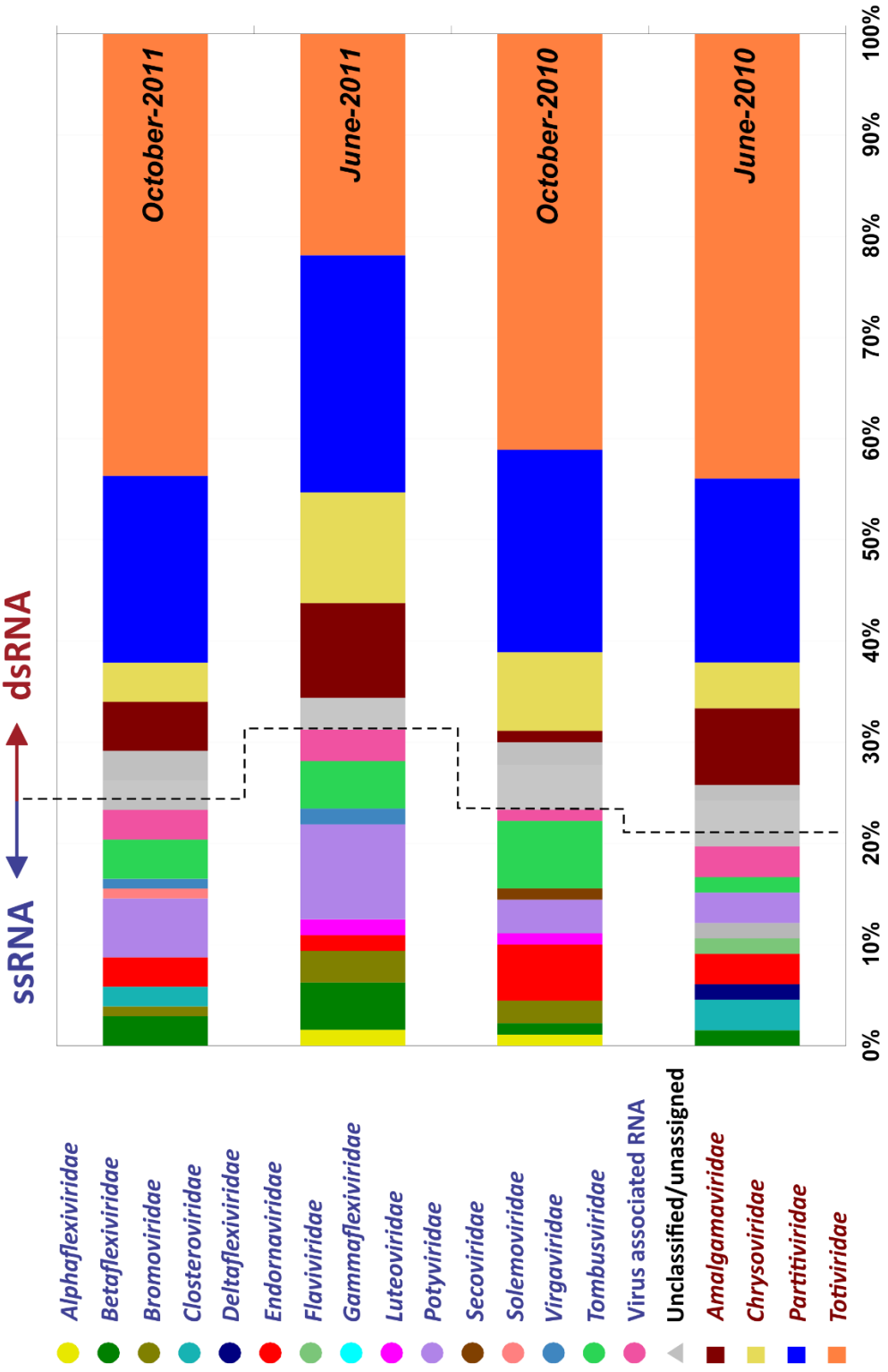

October-2011
June-2011
October-2010
June-2010
dsRNA
ssRNA
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Alphaflexiviridae
Betaflexiviridae
Bromoviridae
Closteroviridae
Deltaflexiviridae
Endornaviridae
Flaviviridae
Gammaflexiviridae
Luteoviridae
Potyviridae
Secoviridae
Solemoviridae
Virgaviridae
Tombusviridae
Virus associated RNA
Unclassified/unassigned
Amalgamaviridae
Chrysoviridae
Partitiviridae
Totiviridae

Figure 4

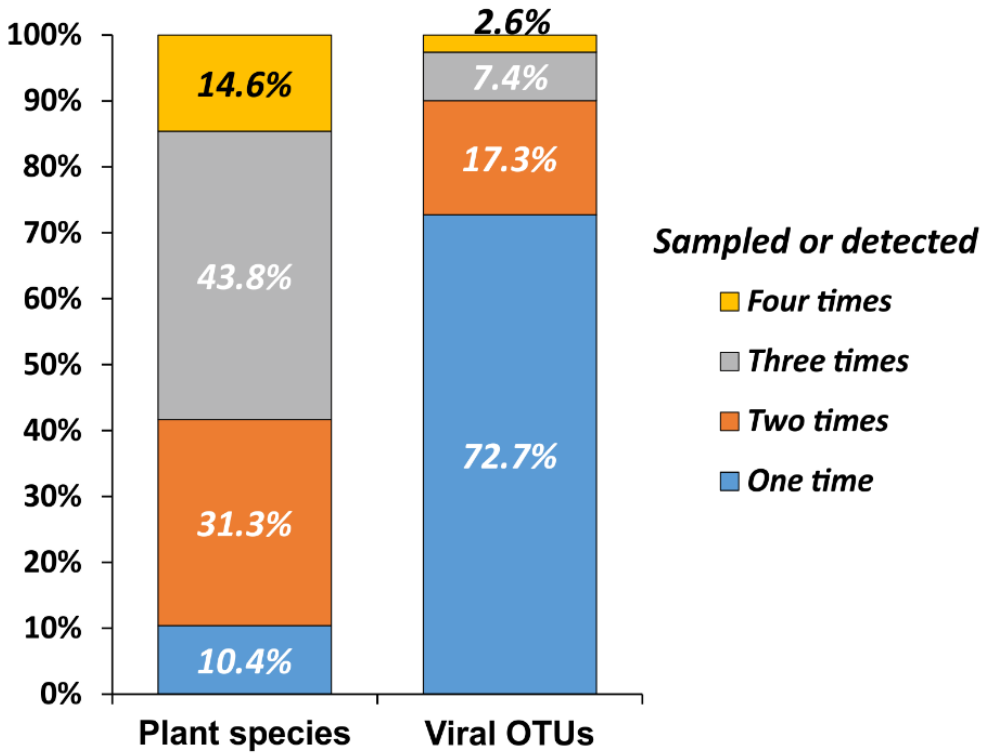
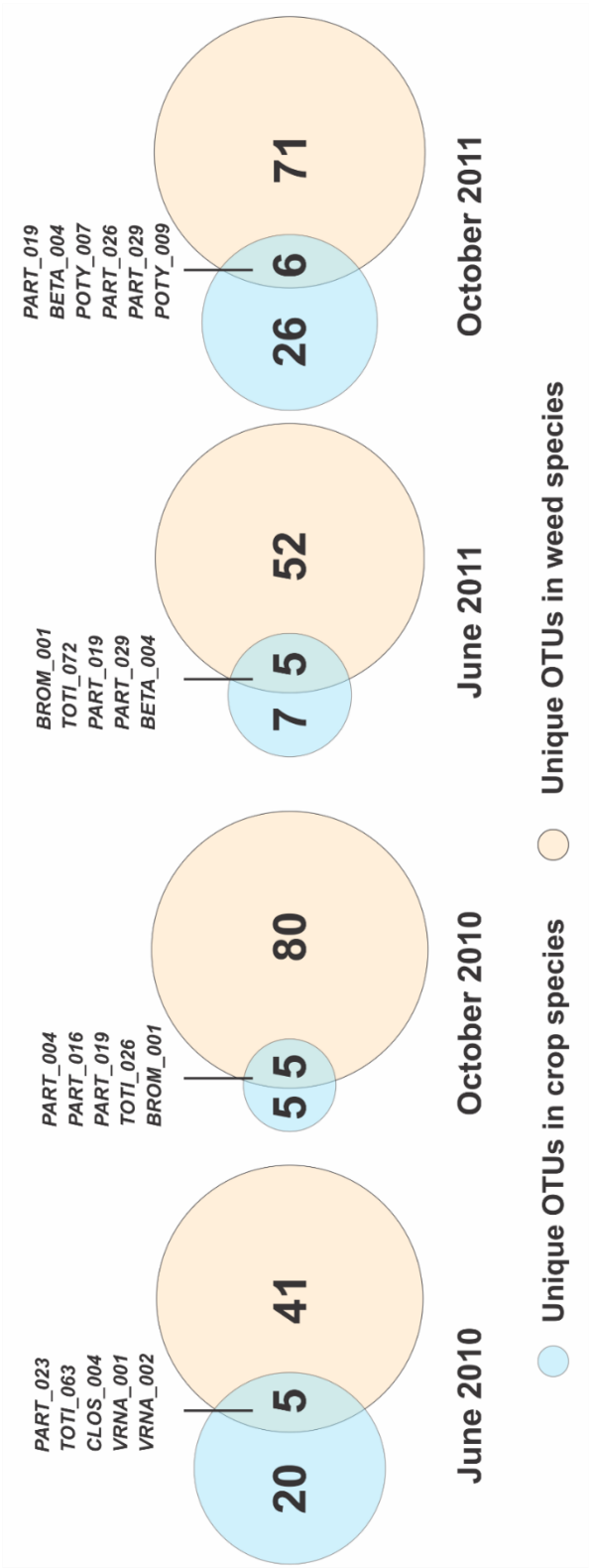

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
14.6%
43.8%
31.3%
10.4%
2.6%
7.4%
17.3%
72.7%
Plant species
Viral OTUs
Sampled or detected
Four times
Three times
Two times
One time

Figure 5


PART_023
TOTI_063
CLOS_004
VRNA_001
VRNA_002
20
5
41
June 2010
PART_004
PART_016
PART_019
TOTI_026
BROM_001
5
5
80
October 2010
BROM_001
TOTI_072
PART_019
PART_029
BETA_004
7
5
52
June 2011
PART_019
BETA_004
POTY_007
PART_026
PART_029
POTY_009
26
6
71
October 2011
Unique OTUs in crop species
Unique OTUs in weed species

Supplementary Materials [Supplementary Tables do not fit easily in an A4 format and are available at https://doi.org/10.15454/JRB3P4]

Table S1. Plant sample identity, plant biology information, number of reads and number and identity of detected OTUs.

Table S2. Taxonomic information provided by the BlastX analysis of OTU-representative sequences and host species information of OTUs. DOI:

Table S3. Known ssRNA viruses discovered in Villenave d'Ornon site and their host plants.

Table S4. Table S4. OTUs identified from each plant species.

Figure S1. Schematic representation of the pipeline for OTUs classification and for removal inter-sample cross-talk (see Material and Methods for details)**.**

Figure S2. Distribution of the length of contigs (A) and of the number of reads per composite plant sample (B).

Figure S3. Phylogenetic tree reconstructed using in the 100 amino acid region surrounding the conserved GDD RdRp motif for OTUs and reference sequences belonging to the family *Partitiviridae.* Tree was constructed by the neighbor-joining method and a strict identity distance. The statistical significance of branches was evaluated by bootstrap analysis (1,000 replicates) and only bootstrap values above 50% are indicated. The scale bar represents 10% amino acid divergence. The genera to which the reference species belong are indicated with dashed lines on the right of the figure. Branches in orange indicate fungi-infecting viruses, green branches plant-infecting viruses and light blue protest-infecting viruses. Cucumber mosaic virus and pepino mosaic virus are used as outgroups.

Figure S4. Distribution of the number of OTUs (n=231) detected per (A) composite plant sample (n=126) and (B) plant species (n=48). Distribution of the number of OTUs based on the number of plant samples (C) or of plant species they infect (D). The corresponding statistic values were shown under the histograms.

Figure S5. Aggregated number of unique OTUs detected for the 7 plant species analyzed at all four sampling times. The points illustrating the number of unique OTUs in these seven plant species wholy **(A)** and seperately **(B).** The coefficient of determination R^2 as well as the linear best fit equation are given.

Figure S1

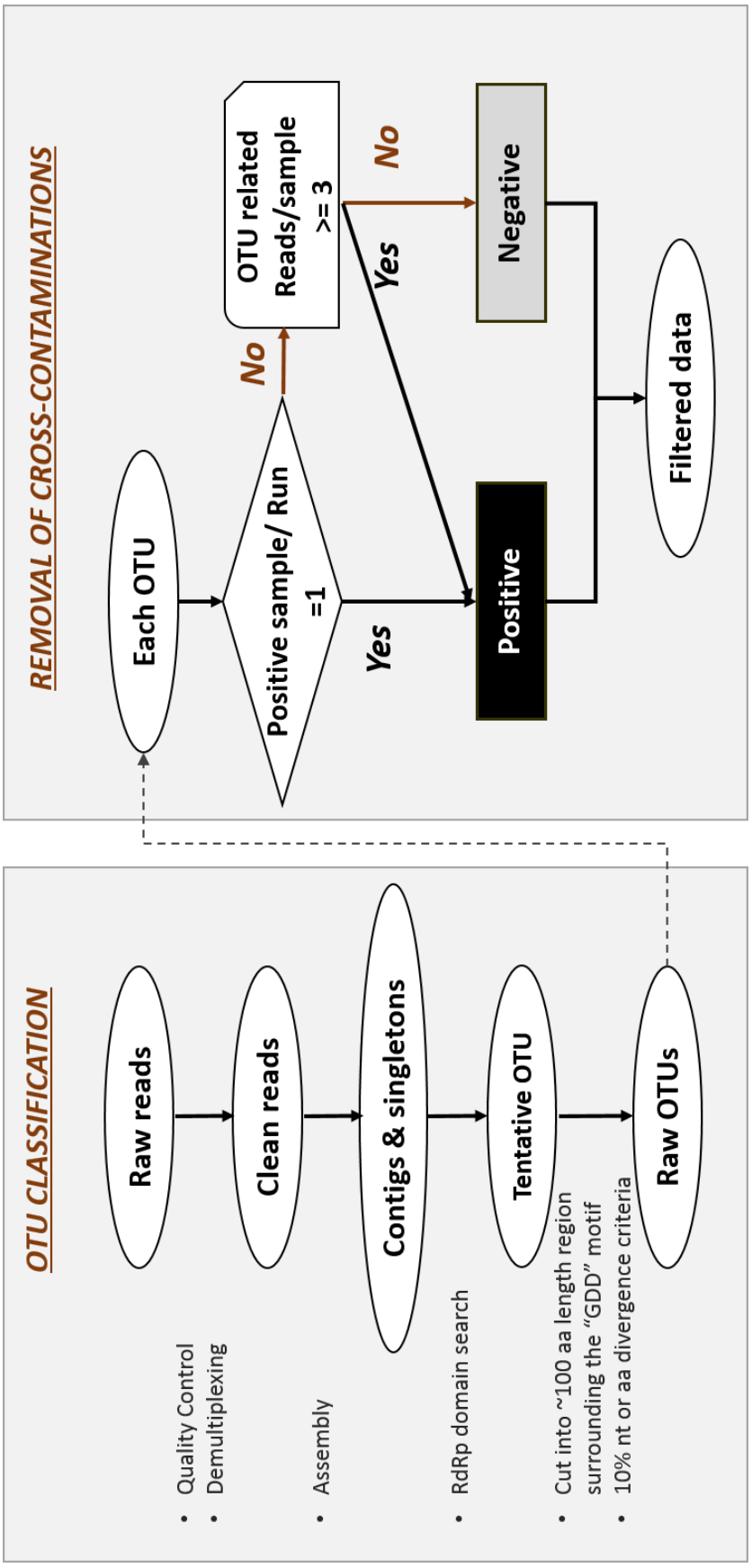

OTU CLASSIFICATION
Raw reads
Clean reads
Contigs & singletons
Tentative OTU
Raw OTUs
• Quality Control
• Demultiplexing
• Assembly
• RdRp domain search
• Cut into ~100 aa length region surrounding the "GDD" motif
• 10% nt or aa divergence criteria
REMOVAL OF CROSS-CONTAMINATIONS
Each OTU
Positive sample/ Run =1
No
OTU related Reads/sample >= 3
No
Yes
Yes
Negative
Positive
Filtered data

Figure S2

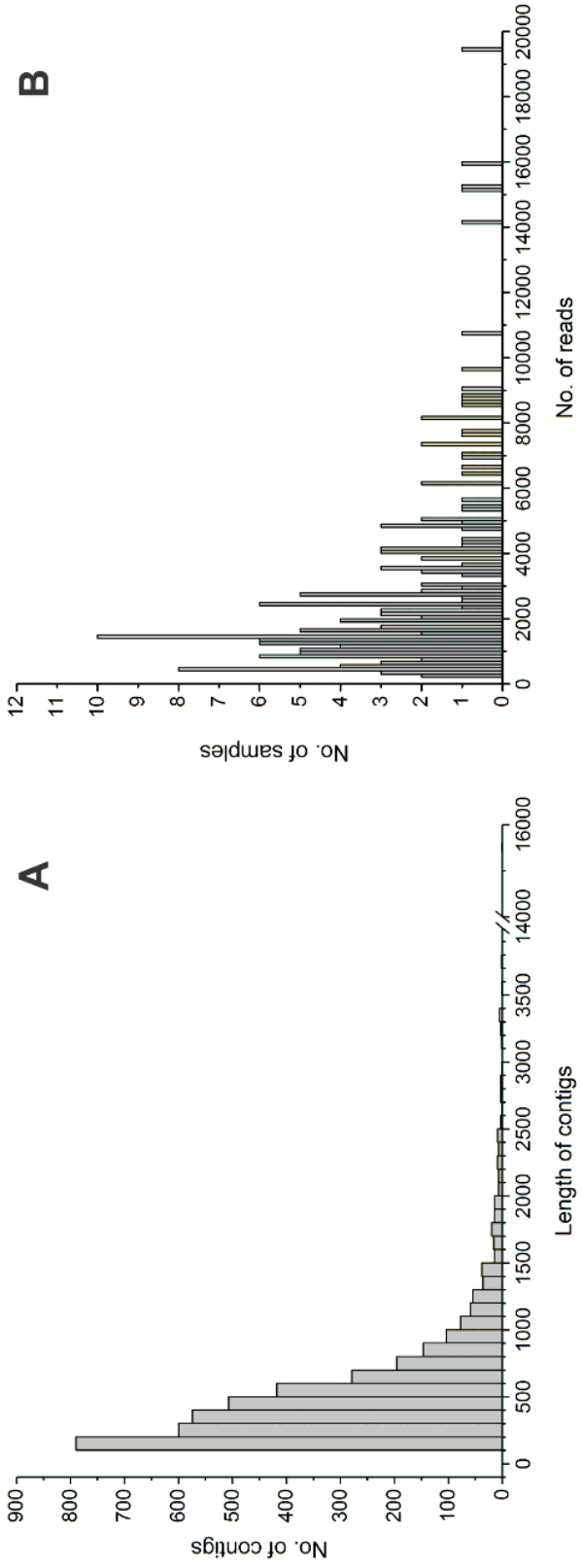

B
No. of samples
0
1
2
3
4
5
6
7
8
9
10
11
12
No. of reads
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
A
No. of contigs
0
100
200
300
400
500
600
700
800
900
Length of contigs
0
500
1000
1500
2000
2500
3000
3500
14000
16000

Figure S3

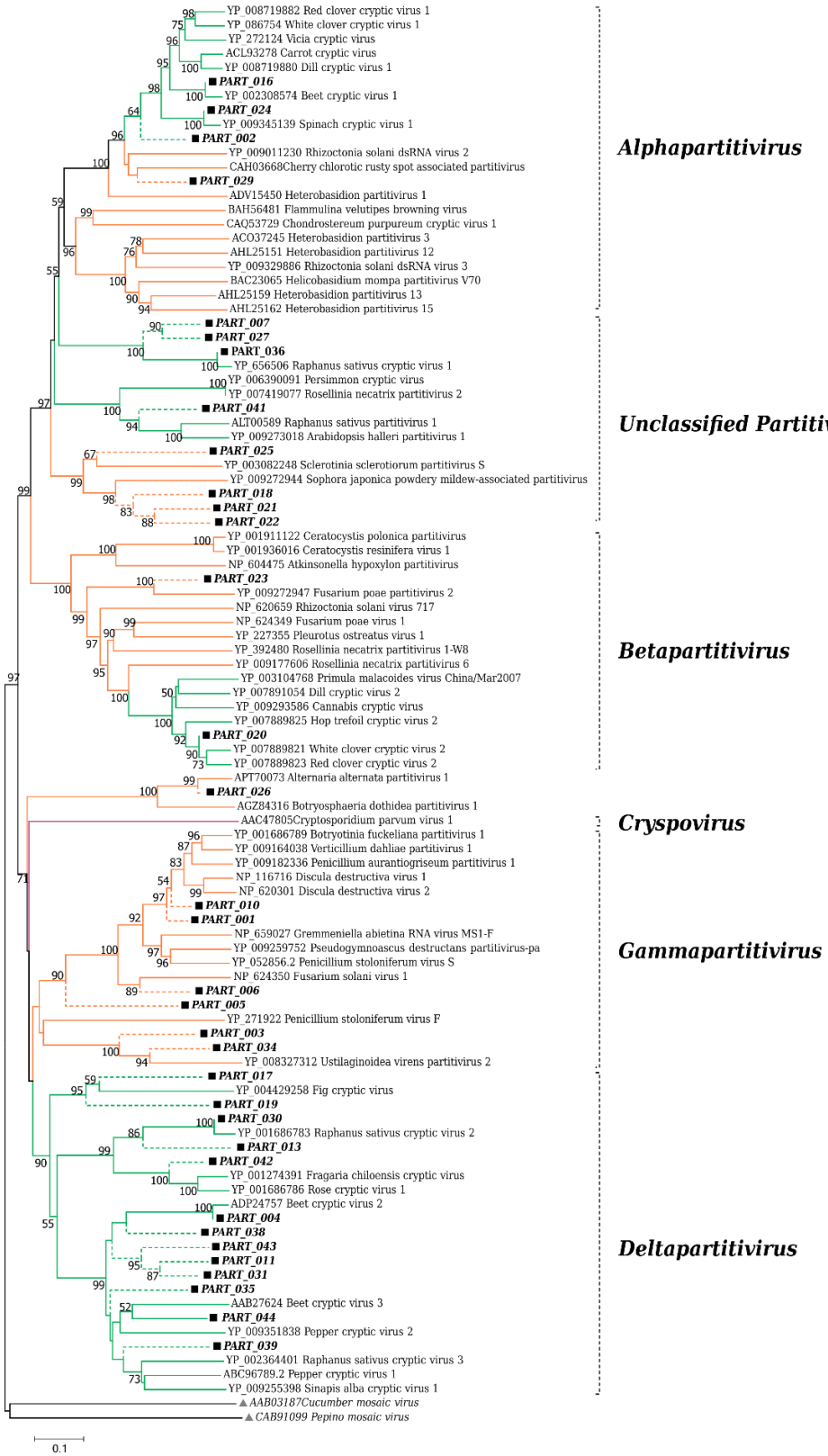
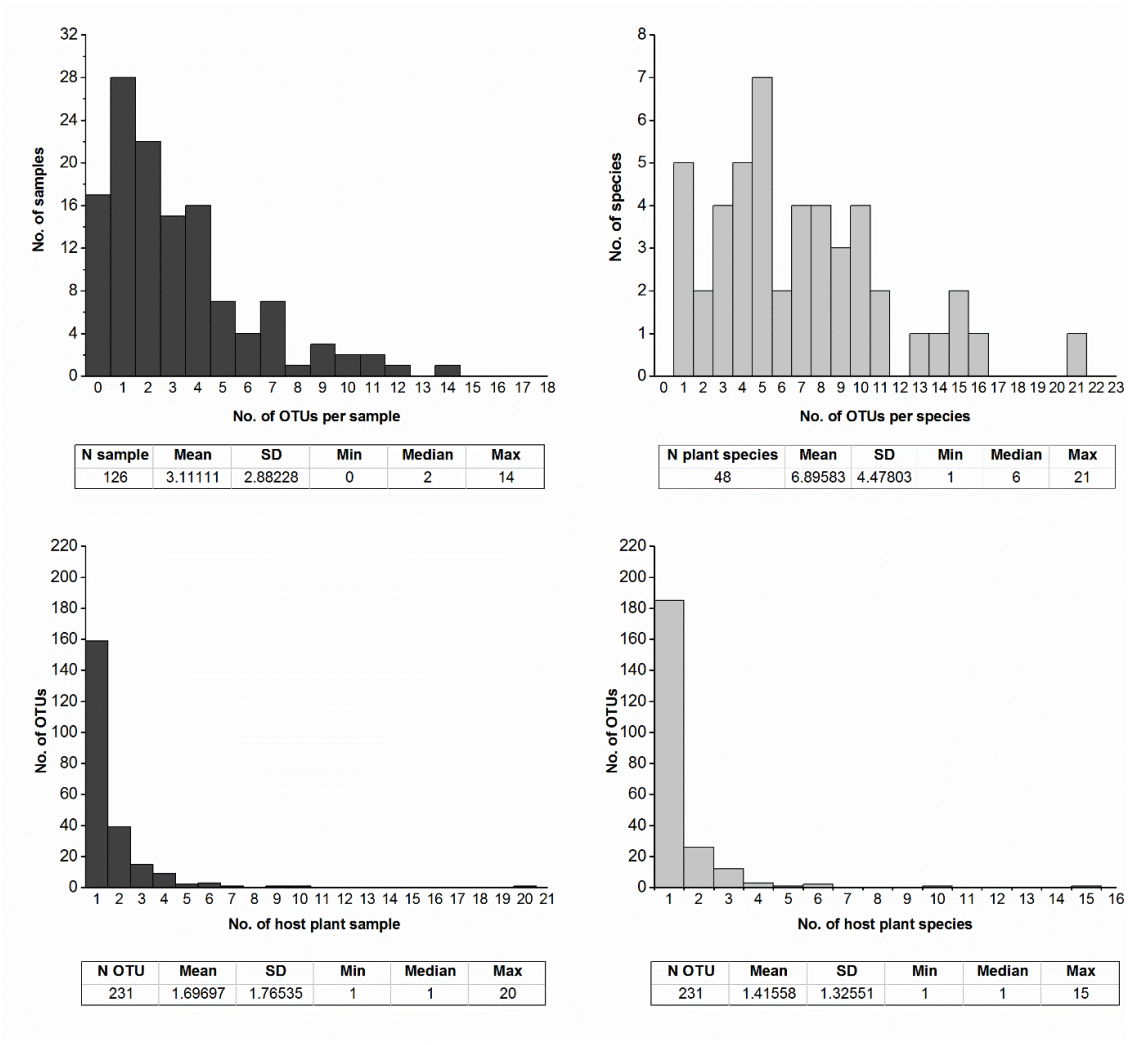

YP_008719882 Red clover cryptic virus 1
YP_086754 White clover cryptic virus 1
YP_272124 Vicia cryptic virus
ACL93278 Carrot cryptic virus
YP_008719880 Dill cryptic virus 1
PART_016
YP_002308574 Beet cryptic virus 1
PART_024
YP_009345139 Spinach cryptic virus 1
PART_002
YP_009011230 Rhizoctonia solani dsRNA virus 2
CAH03668Cherry chlorotic rusty spot associated partitivirus
PART_029
ADV15450 Heterobasidion partitivirus 1
BAH56481 Flammulina velutipes browning virus
CAQ53729 Chondrostereum purpureum cryptic virus 1
ACO37245 Heterobasidion partitivirus 3
AHL25151 Heterobasidion partitivirus 12
YP_009329886 Rhizoctonia solani dsRNA virus 3
BAC23065 Helicobasidium mompa partitivirus V70
AHL25159 Heterobasidion partitivirus 13
AHL25162 Heterobasidion partitivirus 15
PART_007
PART_027
PART_036
YP_656506 Raphanus sativus cryptic virus 1
YP_006390091 Persimmon cryptic virus
YP_007419077 Rosellinia necatrix partitivirus 2
PART_041
ALT00589 Raphanus sativus partitivirus 1
YP_009273018 Arabidopsis halleri partitivirus 1
PART_025
YP_003082248 Sclerotinia sclerotiorum partitivirus S
YP_009272944 Sophora japonica powdery mildew-associated partitivirus
PART_018
PART_021
PART_022
YP_001911122 Ceratocystis polonica partitivirus
YP_001936016 Ceratocystis resinifera virus 1
NP_604475 Atkinsonella hypoxylon partitivirus
PART_023
YP_009272947 Fusarium poae partitivirus 2
NP_620659 Rhizoctonia solani virus 717
NP_624349 Fusarium poae virus 1
YP_227355 Pleurotus ostreatus virus 1
YP_392480 Rosellinia necatrix partitivirus 1-W8
YP_009177606 Rosellinia necatrix partitivirus 6
YP_003104768 Primula malacoides virus China/Mar2007
YP_007891054 Dill cryptic virus 2
YP_009293586 Cannabis cryptic virus
YP_007889825 Hop trefoil cryptic virus 2
PART_020
YP_007889821 White clover cryptic virus 2
YP_007889823 Red clover cryptic virus 2
APT70073 Alternaria alternata partitivirus 1
PART_026
AGZ84316 Botryosphaeria dothidea partitivirus 1
AAC47805Cryptosporidium parvum virus 1
YP_001686789 Botryotinia fuckeliana partitivirus 1
YP_009164038 Verticillium dahliae partitivirus 1
YP_009182336 Penicillium aurantiogriseum partitivirus 1
NP_116716 Discula destructiva virus 1
NP_620301 Discula destructiva virus 2
PART_010
PART_001
NP_659027 Gremmeniella abietina RNA virus MS1-F
YP_009259752 Pseudogymnoascus destructans partitivirus-pa
YP_052856.2 Penicillium stoloniferum virus S
NP_624350 Fusarium solani virus 1
PART_006
PART_005
YP_271922 Penicillium stoloniferum virus F
PART_003
PART_034
YP_008327312 Ustilaginoidea virens partitivirus 2
PART_017
YP_004429258 Fig cryptic virus
PART_019
PART_030
YP_001686783 Raphanus sativus cryptic virus 2
PART_013
PART_042
YP_001274391 Fragaria chiloensis cryptic virus
YP_001686786 Rose cryptic virus 1
ADP24757 Beet cryptic virus 2
PART_004
PART_038
PART_043
PART_011
PART_031
PART_035
AAB27624 Beet cryptic virus 3
PART_044
YP_009351838 Pepper cryptic virus 2
PART_039
YP_002364401 Raphanus sativus cryptic virus 3
ABC96789.2 Pepper cryptic virus 1
YP_009255398 Sinapis alba cryptic virus 1
AAB03187Cucumber mosaic virus
CAB91099 Pepino mosaic virus
0.1
Alphapartitivirus
Unclassified Partitiviridae
Betapartitivirus
Cryspovirus
Gammapartitivirus
Deltapartitivirus

Figure S4



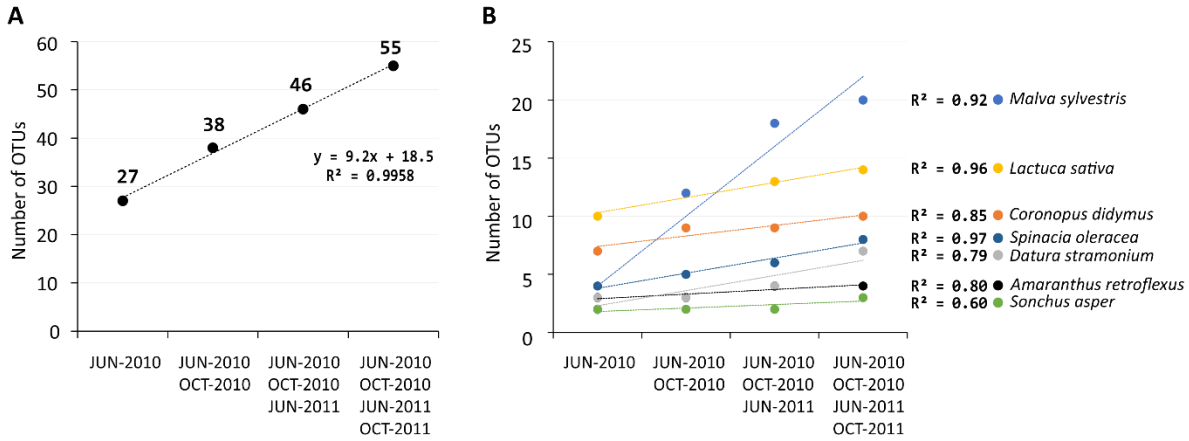
N sample	Mean	SD	Min	Median	Max
126	3.11111	2.88228	0	2	14

N plant species	Mean	SD	Min	Median	Max
48	6.89583	4.47803	1	6	21

N OTU	Mean	SD	Min	Median	Max
231	1.69697	1.76535	1	1	20

N OTU	Mean	SD	Min	Median	Max
231	1.41558	1.32551	1	1	15

Figure S5



CHAPTER Ⅱ

Phytovirome analysis of wild plant populations: comparison of double-stranded RNA (dsRNA) and Virion-associated nucleic acids (VANA) metagenomic approaches

Phytovirome analysis of wild plant populations: comparison of double-stranded RNA (dsRNA) and Virion-associated nucleic acids (VANA) metagenomic approaches

Yuxin Ma, Armelle Marais, Marie Lefebvre, Sébastien Theil[1], Laurence Svanella-Dumas, Chantal Faure and Thierry Candresse*

UMR 1332 BFP, INRA, Univ. Bordeaux, CS20032, 33882 Villenave d'Ornon cedex, France

Abstract word count: Abstract: 250 words; Importance: 136 words

Text word count: 5924 words (Introduction-Results-Discussion-Materials and Methods)

* Corresponding author thierry.candresse@inra.fr

1: current address INRA UMRF, 20, côte de Reyne, 15000 Aurillac, France

Cleaned virome HTS reads have been deposited on the INRA National Data Portal under the identifier https://doi.org/10.15454/TVWBCQ

Running title: dsRNA vs VANA for phytovirome description

ABSTRACT

Metagenomic studies have indicated that the diversity of plant viruses was until recently far underestimated. As important components of ecosystems, there is a need to explore the diversity and richness of the viruses associated with plant populations and to understand the drivers shaping their diversity in space and time. Two viral sequence enrichment approaches, double-stranded RNA (dsRNA) and Virion-associated nucleic acids (VANA), have been used and compared here for the description of the virome of complex plant pools representative of the most prevalent plant species in unmanaged and cultivated ecosystems. A novel bioinformatics strategy was used to assess viral richness not only at family level but also by determining Operational Taxonomic Units (OTU) following the clustering of conserved viral domains. A large viral diversity, dominated by novel dsRNA viruses was detected in all sites while a large between sites variability limited the ability to draw clear conclusion on the impact of cultivation. A trend for a higher diversity of dsRNA viruses was nevertheless detected in unmanaged sites (118 vs 77 unique OTUs). The dsRNA-based approach consistently revealed a broader and more comprehensive diversity for RNA viruses than the VANA approach, whatever the assessment criterion. In addition, dissimilarity analyses indicated both approaches to be largely reproducible, but not necessarily convergent. These findings illustrate features of phytoviromes in various ecosystems and a novel strategy for precise virus richness estimation. These results allow to reason methodological choices in phytovirome studies and, likely in other viromes study where RNA viruses are the focal taxa.

IMPORTANCE

There are today significant knowledge gaps on phytovirus populations and on the drivers impacting them, but also on the comparative performance methodological approaches for their study. We used and compared two viral sequences enrichment approaches, double-stranded RNAs (dsRNA) and virion-associated nucleic acids (VANA) for phytovirome description in complex pools representative of the most prevalent plant species in unmanaged and cultivated ecosystems. Viral richness was assessed by determining Operational Taxonomic Units (OTU) following the clustering of conserved viral domains. There is some limited evidence of an impact of cultivation on viral populations. These results provide data allowing to reason the methodological choices in virome studies. For researchers primarily interested in RNA viruses, the dsRNA approach is advised because it consistently provided a more comprehensive description of the analysed phytoviromes, but it understandably underrepresented DNA viruses and bacteriophages.

KEYWORDS: metagenomics, double stranded RNA (dsRNA), virion-associated nucleic acids (VANA), virome, OTU, viral diversity, phytovirome

INTRODUCTION

Until recently, plant virology has been largely focused on important crops and on destructive viruses impacting agricultural production, limiting our understanding of plant virus diversity (1). In particular, viruses infecting wild plants have been largely neglected, even if they represent reservoirs for both known viruses and for novel, emerging agents. The few metagenomic studies to date have shown that viruses are common in wild plants, even in the absence of symptoms, with a prevalence as high as 70% and a majority of novel agents (2-5). These studies also showed that in wild plants a majority of the detected agents are persistent viruses which are mostly asymptomatic and transmitted vertically through host cell division and sexual reproduction (6, 7). Building on these pioneering efforts, metagenomics and virus ecology are now trying to answer some fundamental questions centered on the identity and diversity of plant-associated viruses, the evolutionary drivers influencing the assembly in space and time of these viral communities and their contribution to the functioning of plant communities (8-10).

High-throughput sequencing (HTS) coupled with bioinformatic analyses are at the core of metagenomics but have also largely superseded all previously used approaches in virus discovery and etiology because of the ability to efficiently detect known and novel viruses without any *a priori* information (11-16). Moreover, metagenomics studies (17-19) have greatly contributed to a redefinition of the RNA virosphere of invertebrates and to a reshaping of our understanding of the origin and evolution of RNA viruses. HTS has been successfully used for a large range of plants (20-23), demonstrating its wide applicability. These efforts also show that a variety of nucleic acids populations can be used, with consequences for the range of identified viruses (23). So far, the main approaches have targeted double-stranded RNA (dsRNA) (7, 22), virus-derived small

interfering RNA (siRNA) (24, 25), virion-associated nucleic acids (VANA) (26-28), total RNA with or without ribosomal RNA depletion (29, 30) and polyadenylated RNA (31). Their respective advantages and disadvantages for virus discovery and etiology efforts involving single plant samples or samples of low complexity have been discussed in detail (21, 23, 32).

As compared to these efforts, the study of viromes associated with plant populations requires the analysis of a large number of plants. Two strategies have essentially been used, the so-called "ecogenomics" (23) or "geometagenomics" (33), which involve the analysis of single plants or of low complexity pools in a highly multiplexed format and the "metagenomics" or "lawnmower" approach which relies on the direct analysis of complex plant pools (30). While the first strategy retains information on the host(s) of each agent, the second allows a more direct virome characterization for multiple environmental points. However, with the currently used multiplexing strategies, a low level index-hopping may ultimately have a negative impact on data quality (34, 35). Given budgetary constraints, both approaches have so far relied almost exclusively on the two strategies providing an enrichment of viral sequences, dsRNA and VANA (7, 33, 36).

Unfortunately, there is little information on the comparative performance of these two approaches for virome description. Some elements can be gleaned, however, from virus discovery efforts. Candresse *et al*. (37) used both siRNA and VANA and showed that higher genome coverage and longer contigs were obtained using VANA for a DNA mastrevirus. Another study compared siRNA and VANA to test if the same representation of within-host viral population structure could be obtained (38). Both approaches provided similar viral mutational landscapes but VANA performed

better in complete viral genome reconstruction and allowed to more readily detect recombinants (38).

In 2016, a comparison of siRNA and ribosomal RNA-depleted total RNA for *Citrus tristeza virus* (CTV) [(+)ssRNA, *Closteroviridae*] and *Citrus dwarfing viroid* (*Pospiviroidae*) characterization on grapefruit indicated that ribosomal RNA depleted total RNA is superior to siRNA in *de novo* assembly and genome coverage for CTV but not for the viroid (30). The same approaches have also been compared for the detection of various viruses and viroids in different plants (29). The main conclusion was that the performance of these two approaches was virus-dependent but that consistent with (34), longer contigs and higher genome coverage were generated using ribosomal RNA depleted total RNA. Additionally, a *Cytorhabdovirus* was detected only from ribosomal RNA depleted total RNA (29).

In the sole study to date that compared dsRNA and VANA for wide scale metagenomics, Thapa *et al* (36) used the two approaches to describe viral diversity in six native plant species from the Tallgrass Prairie Preserve in Oklahoma and test the effects of host identity, location, and sampling year on the virome composition. More operational viral taxonomic units (OTUs) were discovered by the dsRNA approach (29 against 7 for VANA). In addition, 86% of VANA-OTUs were also detected by dsRNA. The two approaches also showed different performance when analyzing the effects of sampling site on virome composition (36). It should be pointed out that similar to that of Thapa *et al.*, most studies to date have used a quite broad definition for OTUs, considered as relatively wide taxonomic categories such as families or genera (33, 36).

Overall, while the available approaches have proven effective in a virus detection context in a range of plant/virus combinations, there is still limited information allowing to reason methodological choices in plant virus metagenomics. Here, we directly compared the performance of dsRNA and VANA for virome description using complex metagenomics plant pools from unmanaged and cultivated environments. The results uncovered rich viromes and suggest that the dsRNA approach should be preferred when analyzing such complex plant pools since it consistently provided a more comprehensive description of the analysed phytoviromes, with the exception of the DNA viruses.

- **RESULTS**

Summary of HTS datasets and sequencing depth normalization

The phytoviromes of 6 different study sites were analysed using pools of equal complexity composed of 200 plants assembled using 5 individual plants of each of the 40 most prevalent species. Following double-stranded RNAs (dsRNAs) or Virion-associated nucleic acids (VANA) extractions, target nucleic acids populations were converted to cDNA and submitted to random whole genome amplification (WGA) before Illumina sequencing. In order to evaluate the reproducibility of the WGA, all samples were amplified in duplicates involving different multiplex identifier (MID) tags. This situation is reflected in the name of the various libraries which indicates the name of the target nucleic acids (ds or VANA) followed by the study site and PCR1 or PCR2 to indicate the WGA replicate. A total of 20 million reads (paired-end and singletons) were generated from the 12 dsRNA libraries, 16 million reads from the 12 VANA libraries and 1 million reads from one negative control blank reagent-only library (Table S1). Following demultiplexing

and quality trimming, each library provided 0.5 to 3 million reads but, in order to limit inter-sample cross-talk only both pair members for which the expected MID tag was identified in both pair members (122 to 997 thousands pairs, depending on the library) were used for subsequent steps. To allow meaningful comparisons between approaches and sites, a normalization was performed by randomly subsampling all libraries to a depth of 122,259 pairs, corresponding to the library with the fewest reads, VANA-IT-PCR2 (Table S1). All further analyses were performed on these normalized datasets. The average read length for the dsRNA libraries is 120.9 +/- 1.3 nucleotides (nt), which is not significantly different from the 121.2 +/- 1.5 nt obtained for VANA libraries (*p* value=0.6075) (Table 1 and Table S1).

Comparisons of contigs assembly and annotation

Overall, a significantly higher proportion of reads from dsRNA libraries were assembled into contigs (average 80.4% +/- 4.3%) as compared to the VANA ones (average 63.4% +/- 12.4%) (36% reads in the blank control) (Table 1 and Table S1). Statistically significant differences between the two approaches were observed in all parameters describing contig length [*total length of contigs*, *mean*, *median*, *N50* and *N90* (Table 1 and Table S1)], with dsRNA libraries assemblies providing, on average, longer contigs than VANA ones. Taken together, these results would suggest a higher complexity, leading to a decreased assembly efficiency for the VANA libraries.

Contigs were then annotated using BlastN and BlastX analyses against the GenBank database and an e-value cut-off of 10^{-4} (39). For those contigs annotated as viruses, taxonomic assignation was retained at the family level, if available, since in many cases our own and others observations have shown that assignation at lower taxonomic level (genus or species) are frequently unreliable (36, 40). Those viruses with no family information were either kept as unclassified viruses or, if

genome-type information was available, were annotated as single-stranded RNA or double-stranded RNA unassigned viruses. The proportions of plant, virus or unknown contigs proved highly variable between libraries. In addition, the VANA libraries obtained from two sampling sites showed significant presence of contigs of bacterial origin (average of 47.3% and 50.3% of contigs for the VO and SP sites, respectively) (Table S1). On average, dsRNA libraries yielded 33.3% of viral contigs (standard deviation 7.3%, extremes 23.6-44.8%) as compared to 20.7% +/- 10.3% for VANA libraries (extremes 8.2-34.5%) (Table 1 and Table S1), a statistically significant difference (*p* value=1.3e-06). When taking into consideration the reads integrated in the different contigs, 49.9% +/- 14.3% of dsRNA reads were annotated as viral, as compared to 40.5% +/- 16.6% for VANA reads. However, this difference was not found statistically significant (*p* value=0.2193) (Table 1). Unsurprisingly, 94.3% of contigs in the blank control library were annotated as being of bacterial origin and no viral contigs were identified (Table S1).

Family-level viral diversity as reflected by contig annotation

The results of the Blast annotation show that at family level, the dsRNA-based approach consistently detected more viral families per study site (extremes 11-16 families, average 13.3 +/- 1.7 families) than VANA (extremes 6-15 families, average 9.3 +/- 2.6 families) (Table 1 and Table S1). Therefore, on average dsRNAs allowed the identification of 3.9 +/- 2.3 more viral families per study site than VANA. However, when considering all samples together, VANA allowed the identification of a total of 24 viral families, as compared to 21 for dsRNA. This difference is largely due to the infrequent detection of DNA viral or phages families not detected by dsRNA (*Metaviridae*, *Siphoviridae*, *Podoviridae*, *Genomoviridae*, *Geminiviridae* and *Circoviridae*). Conversely, dsRNA allowed the detection of RNA viral families not detected by VANA

(*Reoviridae, Cystoviridae, Rhabdoviridae* and *Narnaviridae*) (Table S2). Overall, phages represented only a very minor fraction of the detected viral contigs (16 contigs or 1.2% of viral ontigs for VANA and 11 contigs or 0.5% of viral contigs for dsRNA, respectively).

While most DNA virus families were only detected by VANA from a few sites, many dsRNA or ssRNA families such as the *Amalgaviridae*, *Chrysoviridae*, *Closteroviridae*, *Benyviridae*, *Luteoviridae and Secoviridae* were detected from significantly more study sites using dsRNA than VANA (Table S2). This situation is particularly striking for the *Chrysoviridae* (6 sites vs 1) or the *Closteroviridae* (6 sites vs 3). On the other hand, as judged by reads number the ssDNA *Nanoviridae* family was very poorly detected by the dsRNA approach as compared with the VANA one (Table S2).

Representation of viral families as estimated by read number

The number of reads for each viral family varied significantly between study sites and, for a given site, between the two virome sequencing approaches (Fig. 1). The most represented viral family in the dsRNA approach is the *Endornaviridae*, and overall it accounts for nearly 4 times the reads observed with the VANA approach (410884 vs 107904). Intuitively, *Endornaviridae* reads may have saturated the dsRNA libraries of the SP site, reaching 66-70% of the viral reads (Table S2). For other dsRNA viral families, the same general trend of a higher representation in dsRNA libraries is also observed. This is particularly clear for the *Chrysoviridae* but generally applies to all dsRNA viruses. There are however some exceptions as for example for *Partitiviridae* at the BP study site or for *Totiviridae* at the INRA one (Table S2). Conversely the expected better representation of DNA viral families in the VANA approach is observed but these families were only detected in a minority of the study sites (Fig. 1). For ssRNA viral families, the picture is more

complex. However, it seems noteworthy that viral families showing the strongest over-representation in the VANA libraries, with up to 3 times more reads as compared to the dsRNA ones, tend to have particularly stable particles, such as for the *Virgaviridae* or *Solemoviridae* (41, 42). On the other hand, the *Closteroviridae*, which have unstable and hard to purify particles or the low titer *Luteoviridae* showed, with over two times more reads, a tendency to be more represented in the dsRNA libraries (Fig. 1 and Table S2).

OTU-based assessment of viral richness with the dsRNA and VANA approaches

For a variety of reasons, including the absence of universally conserved genomic elements and the frequently incomplete genome coverages, the in depth characterization of viromes at a level close to taxonomic species has remained largely elusive. However a possible strategy to circumvent these difficulties has been proposed, involving the clustering of contigs encoding proteins sharing conserved motifs (40). We have developed a pipeline which sequentially identifies such contigs for a range of conserved viral motifs using RPS-Blast against the Pfam database (Table S3), aligns the contigs and finally performs a clustering, allowing the definition operational taxonomic units (OTUs) on the basis of a defined identity cut-off value (43).

The dsRNA and VANA datasets were analysed using this strategy and a 10% cut-off value, which reasonably approximates in many families the envelope of viral species variability (43). RPS-Blast of all contigs identified contigs encoding 47 different viral conserved protein motifs, including those corresponding to well-known signature sequences such as RNA-dependent RNA polymerases (RdRp) and viral helicases (Table S3). For example, the matches for the different RdRp signatures (RdRp_1, 2, 3 and 4) collectively consist of sequences covering a very wide range of plant or fungal RNA virus families. Contigs corresponding to motifs with a much more restricted

taxonomic signature were also identified, such as pfam01787, a protein family specific of the coat protein of Ilarviruses in the family *Bromoviridae* (Table S3).

In order to avoid counting multiple times the same contig if it contained multiple signature sequences, the analysis was focused on the four RdRp protein families: RdRp_1, 2, 3, 4 which are specific for RNA viruses and cover the broadest diversity of these agents. This has however the side effect of focusing the analysis on RNA viruses, so that a detailed analysis of DNA viruses would require, in addition, to also consider some DNA viruses-specific motifs. Using a 10% identity clustering cut-off value, a total of 239 RdRp OTUs were identified when taking into consideration all dsRNA and VANA datasets. Annotation of contigs representative of each OTU by BlastX allowed the identification of 16 RNA virus families (Table S4), to be compared with the 18 RNA virus families detected by the direct annotation of contigs (Fig. 1). This difference might be explained by families for which a low coverage has resulted in incomplete genome assemblies in which the conserved viral RdRp motif is missing. The two families detected by direct Blast annotation of contigs but not by OTUs annotation were *Reoviridae* and *Rhabdoviridae.*

When comparing the VANA and dsRNA approaches, VANA detected 14 of the 16 RNA virus families detected using dsRNA, missing only two families, the *Amalgaviridae* and *Benyviridae* (Fig. 2D). As indicated above, this is likely due to the low read numbers for these families in the VANA approach (Fig. 1), resulting in incomplete genome coverage and in an absence of contigs covering the RdRp conserved domain for the viruses in these families. This result confirms that when considering the viral families detected, the performance of the VANA and dsRNA approaches are significantly but not widely different.

The dsRNA strategy detected a total of 228 OTUs, while VANA only detected 80 OTUs, of which 69 were detected by both strategies (Fig. 2A). A large number of dsRNA OTUs (n=159) were not detected by VANA (Fig. 2A). Sixty percent of these 159 OTUs were annotated as corresponding to *Totiviridae* members while the remaining 40% come from other families (Table S4). This difference is also observed if different, lower (3%) or higher (20%) cut-off thresholds are used in the clustering phase for the definition of OTUs (Fig. 2B and C).

If defining as novel the OTUs for which there are no sequences in GenBank that share less than the 10% clustering cut-off criterion, the majority of the VANA (81.2%) and dsRNA (89.5%) OTUs correspond to novel agents (Fig. 2E and F). In both approaches the putative novel OTUs group integrates almost all the dsRNA OTUs while only around half of the ssRNA OTUs (48%-54%) appear to correspond to novel agents (Fig. 2E and F).

Comparison of the dsRNA and VANA approaches at the level of individual plant populations

When analyzing independently each sampling site, the same pattern emerged and significantly more OTUs were identified using the dsRNA as compared to the VANA one (Fig. 3). The virome compositions at family level were also more diverse (Fig. 3). On average 9.8 +/-1.3 families were identified using the dsRNA approach per sampling site as compared to only 6.2 +/- 1.9 for VANA (*p* value= 0.0007), with the SP site showing the lowest viral richness (5 OTUs and 4 viral families, Fig. 3). In most sites, *Totiviridae* was the most represented family by OTU, with OTUs making up on average 49.2% +/- 12.0% of the virome for the dsRNA approach, as compared to 33.5% +/- 23.6% for VANA (*p* value=0.2132).

At the individual OTU level also the dsRNA approach revealed a significantly higher diversity,

with an average of 51.5 +/- 17.0 OTUs per sampling site compared to 17.2 +/- 8.9 OTUs for VANA (*p* value=0.003032). In addition, a large proportion of the OTUs identified using VANA (73% +/- 20%, extremes 40%-95%) were also discovered using the dsRNA approach, while a large majority of dsRNA OTUs are not detected by VANA (73% +/- 15%, extremes 56%-96%).

Reproducibility of the VANA and dsRNA approaches

Since two random amplifications and ensuing libraries sequencing were performed for each complex pool, it is possible to evaluate the reproducibility of the viromes obtained from the two whole genome amplification (WGA) replicates but also virome composition specificity in the different study sites. For most variables, there were no statistically significant differences between the two libraries obtained from each sample, including for variables such as number of assemble reads, number of contigs, N50, number of viral contigs, of viral families and of OTUs identified (Table S5).

Besides, the reproducibility of the viromes from either different WGA replicates or different enrichment strategy (dsRNA or VANA) were further evaluated based on OTUs presence/absence data (Fig. 4). The results of hierarchical clustering analyses based on these data show that even if some variability is observed between replicates, the distance between replicates is systematically much lower than the distances between samples (Fig. 4A and B). In addition when comparing the results obtained with the dsRNA and VANA approaches, it is clear that the replicates for each site/technique combination end up very close (Fig. 4C). As shown above there is a very significantly clustering of libraries corresponding to a given site (ANOSIM analysis, R=0.87, *p* value<0.001) (Fig. 4 C), also illustrating the fact that each virome showed strong site specificity with 41%-71% of site-specific OTUs (Table 2). The ecosystem type (cultivated or unmanaged)

had only limited impact on virome composition (ANOSIM test: R=0.2, *p* value=0.002).

Impact of management practices on virome richness and composition

There were no statistically significant differences in the number of OTUs or of viral families between the cultivated and unmanaged sites (Table 2). Similarly, although a small trend could be seen in the average values (92.3% +/-7.8% novel OTUs for unmanaged sites vs 82.7% +/- 7.3% for cultivated ones) the difference in the proportion of novel OTUs was not statistically significant (Table 2). Conversely, OTUs corresponding to already known viruses proved more frequent in cultivated sites than in unmanaged ones (18 vs 12 OTUs).

While the large variability seen at the level of individual sites limited the ability to draw clear conclusions, comparison of aggregated OTU numbers for viral families or viral groups supported the notion of a higher dsRNA viruses diversity in unmanaged sites (118 vs 77 unique OTUs). For ssRNA viruses the trend was reversed, with a marginally higher diversity (31 vs 28 unique OTUs) in managed sites. This trend was particularly clear for *Closteroviridae* (7 vs 2 OTUs) and *Secoviridae* (4 vs 1 OTUs). Conversely, persistent viruses showed an overall higher richness in unmanaged sites, in particular *Totiviridae* (84 vs 56 OTUs), *Chrysoviridae* (8 vs 3 OTUs) and *Endornaviridae* (10 vs 6 OTUs).

- **DISCUSSION**

In this study we compared the effectiveness for phytovirome description of the two most widely used nucleic acid enrichment approaches: double stranded RNA (dsRNA) and virion associated nucleic acids (VANA). The richness of the analysed viromes was assessed with two strategies: direct BlastN or BlastX-based taxonomic annotation of assembled contigs, providing a virome

richness estimate at family level and the identification of viral OTUs based on a clustering of contigs encoding viral RdRp conserved motifs (43). The Blast-based annotation of contigs representative of each OTU also allows a richness estimate at family level.

The OTU-based analysis is expected to provide a lower-bound richness estimate, because agents for which the RdRp-encoding region is not covered cannot be identified as an OTU. This may explain why direct contig annotation identified on average a slightly higher RNA viruses family-based richness than the OTU approach (paired *t*-test, *p* value = 0.0001) (Table S1). For example, for dsRNA libraries an average of 13.3 +/- 1.7 RNA virus families were identified by direct Blast annotation as compared to 11.2 +/- 1.5 families by OTUs clustering (paired *t*-test, *p* value = 0.008). Similarly significant differences were observed using the VANA approach (paired *t*-test, *p* value = 0.01). A possible strategy to increase the completeness of the OTU-based approach would be to also take into account the OTUs defined by other conserved viral motifs such as viral helicases or viral coat proteins (Table S3). A virus for which the RdRp region has no coverage could then be taken into account if its helicase is among the sequence data. This has the potential advantage of improving the ability to detect viral contigs. Indeed of the 1393 contigs identified by RPS-Blast as containing at least one virus-specific motif, 337 (24.2%) were not annotated as viral by the Blast initial analysis. However, this strategy would likely provide an over-representation of the true viral richness since a fully sequenced virus would then give rise to as many OTUs as it has conserved motifs. It is interesting to notice that the low frequency of phage sequences identified by the Blast-based annotation is confirmed by the RPS-Blast search for encoded protein motifs since overall a single VANA contig could be identified as encoding a phage-specific motif.

It should be stressed that the family-level annotation of contigs or OTUs performed here is based

on the first Blast hit and therefore does not guarantee that the agents indeed belong to the identified family. Phylogenetic analyses performed with the contigs representative of OTUs have however shown in other experiments a good general fit between the Blast-assigned family and phylogenetic affinities. Metagenomics studies (17-19) have, for example, greatly contributed to a redefinition of the RNA virosphere of invertebrates. While a wealth of novel OTUs were identified here, our results do not point to the existence of a large number of novel higher order viral taxa (family and above) associated with the sampled plant populations.

Broadly speaking, when taking into account all datasets, the dsRNA and VANA approaches recovered largely the same viral families with only a few viral families not recovered by one or the other approach. Interestingly, *Endornaviridae* members without a true capsid or particle but that produce host derived vesicles containing their nucleic acids were abundantly found from several VANA libraries, confirming similar observations in other studies (33, 44) and indicating that the VANA approach is not limited to virion-producing agents. As expected, for the dsRNA approach DNA viruses were not efficiently recovered, even if some *Nanoviridae* were identified. Indeed, the detection of DNA viruses using dsRNA has been reported in the literature (20, 45, 46). For the VANA approach, a low efficiency of detection was observed for viruses or families with low titer and/or less stable particles, although *Closteroviridae* which are known to have quite labile particles have been detected here and elsewhere (33, 47). It should be mentioned that the excess of reads annotated as having a bacterial origin detected in two sites by VANA may in fact represent the detection of phages since many integrated phages, which can make up to 10-20% of bacterial genomes (48, 49), have been sequenced and annotated as part of bacterial genomes. Overall, the results obtained would however point to a limited presence of phage in the analysed plant-

associated viromes. One possible explanation could be that the concentration in phage particles could be low in the analysed samples and that they could have been outcompeted during the sequencing phase by more frequent phyto- or mycoviruses. In any case, the search for phage-specific motifs using the VirAnnot pipeline allows to specifically search for evidence of phage presence so it will be possible in the future to confirm or infirm the results reported here.

Comparable to other studies, the characterized viromes were dominated by novel dsRNA viruses, while a significant fraction of the less abundant ssRNA viruses proved to correspond to already known agents. Although some tentative trends were observed, no statistically supported differences could be identified between cultivated and unmanaged sites, raising the question of the impact of cultivation practices on the virome of wild plants and weeds growing nearby. Among the strongest trends was to finding of a higher diversity of dsRNA viruses which largely have persistent lifestyles in unmanaged environments. This might reflect an indirect impact of the fungicide treatments applied to crops (see below) or have other causes yet to be established.

Whatever the viral richness evaluation strategy and the sample analysed, the dsRNA approach provided a more complete, richer virome representation. This statistically significant difference was observed both at the family and OTU levels (Table 1, Fig. 2 and 3) and is also observed if different, lower (3%) or higher (20%) cut-off thresholds are used in the clustering phase for the definition of OTUs (Fig. 2B and C). The reasons for this differential performance is unclear. One possibility is that the dsRNA purification protocol used allows for a greater enrichment of viral sequences. This could in turn lead to an ability to assemble longer, more efficiently annotated contigs (Table 1). An alternative hypothesis would involve the possible existence, in the case of VANA, of stronger competition effects between viruses in the complex pools analysed. In this

scenario, highly concentrated and stable viruses could outcompete less stable and/or concentrated ones during the amplification of VANA targets, resulting in a less complete representation. Under both hypotheses, the use of less complex pools and/or deeper sequencing are likely to improve VANA and dsRNA performances.

Both approaches proved to have a good (but not perfect) reproducibility. Indeed, while the libraries prepared by independent amplifications of the same target pool always showed tight clustering in NMDS (Fig. 4C), the corresponding viromes frequently show a differential detection of a small fraction of the OTUs (Fig. 4). A careful analysis shows that most of the differential OTUs are represented by low reads numbers so that small variations in representation in the dataset may strongly affect the ability to assemble contigs for them and, ultimately, their identification. However, a few OTUs with significant coverage were also observed to be differentially detected between duplicate amplification libraries, which might point to other artifactual effects.

A rich diversity identified for mycovirus-like viruses from the *Totiviridae* and, to a lower extent, *Chrysoviridae* families was at all study sites. Given that the plant holobionts were used for sampling this raises the possibility that a proportion of these agents might infect endophytic, epiphytic or parasitic fungi associated with the sampled plants. Indeed a lower richness is observed overall for these families from cultivated sites, a possible consequence of fungicides applications on overall fungal diversity (50, 51). At the same time, many typical fungal virus families such as *Hypoviridae*, *Narnaviridae*, *Fusariviridae* and *Birnaviridae* were not detected here, further complicating the issue.

Overall, unless DNA viruses are of particular interest in metagenomics efforts involving the analysis of complex sample pools by the “lawnmover” strategy (3), the results presented here

suggest that a preference should be given to the dsRNA-based approach since it consistently provides a more comprehensive vision of the virome. It should however be stressed that this recommendation may not apply when analyzing less complex samples such as individual plants or pools of plants of a single species such as in ecogenomics or geometagenomics (23, 33) since VANA has been shown to perform efficiently in virus discovery and etiology studies (21, 23, 27).

- **MATERIALS AND METHODS**

Study sites and plant samples

To analyse plant virus richness in different cultivated or unmanaged environments, six different sites were selected in southwest France (Table S6). The VO site near Bordeaux, is a cultivated horticultural agrosystem, in which the main crops are vegetables such as tomato and lettuce. The nearby unmanaged site (INRA), corresponds to a prairie and adjoining path borders within the INRA research center. Near the town of Bergerac, two cultivated agrosystems (CT and IT), with respectively carrot and tobacco crops were selected, together with two unmanaged areas (SP and BP) corresponding respectively to a dry prairie and to a deciduous forest border.

For each site, a total of 200 individual plants were collected in spring 2016 (5 individual plants of each of the 40 locally most abundant species; Table S6). In the agrosystems, the cropped species were not collected. No specific efforts were made to select symptomatic plants and plants with obvious fungal attack, insect colonization or necrotized parts were excluded. All collected plants were identified to species level or, when not possible, to genus level by a trained researcher.

Samples processing and plant pools preparation

For each sampling site, 4 different bulked samples (50 plants each, 10 different species) were

prepared for dsRNA extractions while 8 different bulked samples (25 plants each, 5 different species) were used for VANA extractions. In each case, the pools were composed of 0.1 g of fresh tissue of each sampled plant, yielding a total of 5 g of plant material for dsRNA pools and 2.5 g for VANA pools.

Viral nucleic acids enrichment, library preparation and Illumina HiSeq sequencing

Double-stranded RNAs were purified from each pool by two rounds of CF11 cellulose chromatography and converted to cDNA according to the protocol described by Marais *et al.* (22). In parallel, a negative control blank was similarly prepared using only buffer. In order to evaluate the reproducibility of the whole genome amplification (WGA) procedure, duplicate WGA PCRs involving different MID tags (19) were performed on each cDNA sample. PCR products were purified using the MinElute PCR Purification Kit (Qiagen) and their concentration determined spectrophotometrically. Finally, equal DNA amounts of the identically tagged WGA PCR products obtained from the 4 separate plant pools of each study site were pooled generating a superpool corresponding to the 200 sampled plants.

Virion-associated nucleic acids (VANA) were extracted from each bulked sample following the protocol described by Candresse *et al.* (37). Synthesis and amplification of cDNAs prepared from nucleic acids extracts were performed by combining reverse-transcriptase priming as described in the dsRNA strategy and a Klenow fragment polymerization step so as to allow the detection of both RNA and DNA viruses simultaneously (33). The resulting products were submitted to WGA in duplicates involving different multiplex identifier (MID) tags, purified, quantified and assembled in superpools as described for the dsRNA strategy. The various libraries were named based on the target nucleic acids (ds or VANA) followed by the study site and PCR1 or PCR2 to

indicate the WGA replicates (ex. ds-VO-PCR1).

In total, 12 libraries were thus prepared for the dsRNA approach (corresponding to duplicate WGA for each of the 6 sampling sites), and one blank pool library for all the negative controls (Table S1). WGA were also performed in duplicate for the VANA samples, again yielding a total of 12 libraries, The 25 resulting libraries, each having a different MID tag were separately used for preparation of independent sequencing libraries and sequenced in multiplexed format (2×150 bp) on an Illumina HiSeq 3000 system at the GenoToul platform (INRA Toulouse, France).

Bioinformatics analyses: Reads cleaning, normalization and contigs assembly

Following demultiplexing, adapters and MID tags were removed with *cutadapt* (52), and reads were quality trimmed (minimum quality score 20, minimum length 70 nucleotides). In order to limit inter-sample cross talk associated with index-hopping (34), only reads having identical MID tags on both pair members were retained for further analyses. Cleaned virome HTS reads have been deposited on the INRA National Data Portal under the identifier https://doi.org/10.15454/TVWBCQ. To compensate for uneven sequencing depth between libraries, libraries were normalized by random subsampling to the same depth (122,295 pairs) using the seqtk tool (https://github.com/lh3/seqtk) (Table S1). Contigs were *de novo* assembled for each library using IDBA-UD (https://academic.oup.com/bioinformatics/article/28/11/1420/266973).

Contigs annotation and Operational taxonomic units (OTU) clustering

All contigs were annotated using BlastN and BlastX against the NCBI Genbank non redundant nucleotide (nt) or protein (nr) databases with a conservative e-value cut-off of 10^{-4}. In this way,

contigs were assigned to one of the following categories: virus, eukaryote, bacteria, algae, and unknown. A heatmap illustrating the representation (absolute number of reads) of viral families (Table S2) in each library/site was prepared using the 'ComplexHeatmap' package without clustering in R (53).

A clustering approach (43) was used to define and count operational taxonomy units, as initially highlighted (36, 40). Briefly, a search of all contigs against the pfam database (54) was performed using Reversed Position Specific Blast (RPS-Blast) (55). The contigs encoding a virus-specific conserved protein motif (Table S4) were retrieved and aligned with reference sequences and distance matrices computed with the ETE3 toolkit (56) . These matrices were used to perform a clustering, allowing to regroup in a single operational taxonomic unit (OTU) all contigs differing by less than a set cut-off divergence value (57). We used of a 10% divergence cut-off value, which has been shown to generate in many viral families OTUs that are a relatively good approximation of taxonomic species (43). OTUs were thus defined and counted for each virus-specific conserved motif, allowing to generate an OTU table indicating for each approach/sampling site combination the presence/absence of each identified OTU. With the exception of the reproducibility analysis, all other analyses were performed by regrouping the data of the duplicate normalized libraries corresponding to the two separate PCR amplifications performed for each approach/sampling site combination.

Dissimilarity analyses between duplicate PCRs and among sampling pools/sites

The availability of two random amplifications and ensuing libraries (PCR1 and PCR2) for each approach /sampling site combination allowed to evaluate virome description reproducibility. Dissimilarity analyses were performed on OTU presence/absence binary data to generate a Jaccard

distance matrix. Based on this distance matrix, hierarchical clusterings and nonmetric multidimensional scaling (NMDS) ordination were performed using hclust with “complete” algorithm and the R ‘vegan’ package (57, 58). The significance of comparisons among different sites and between different ecosystem types (cultivated and unmanaged) were assessed using the non-parametric statistical test - ANOSIM (analysis of similarity) in R ‘vegan’ package (58-60).

ACKNOWLEDGMENTS

The authors would like to thank A. Raoult and F. Villeneuve (Centre Technique Interprofessionnel des Fruits et Légumes, Lanxade) and F. Dorlhac de Borne (Imperial Tobacco, Bergerac) for access to some sampling sites, P. Roumagnac (UMR BGPI, CIRAD) for sharing the VANA protocol, and the Genotoul Platform (INRA, Toulouse, France) for the Illumina sequencing. Y. Ma was supported by a China Scholarship Council grant.

REFERENCES

1. Wren JD, Roossinck MJ, Nelson RS, Scheets K, Palmer MW, Melcher U. 2006. Plant virus biodiversity and ecology. PLoS Biol 4:314-315.

2. Roossinck MJ. 2015. Metagenomics of plant and fungal viruses reveals an abundance of persistent lifestyles. Front Microbiol 5.

3. Roossinck MJ. 2012. Plant Virus Metagenomics: Biodiversity and Ecology. Annu Rev Genet 46:359-369.

4. Cooper I, Jones RA. 2006. Wild plants and viruses: under-investigated ecosystems. Adv Virus Res 67:1-47.

5. Roossinck MJ. 2011. The big unknown: plant virus biodiversity. Curr Opin Virol 1:63-67.

6. Roossinck MJ. 2012. Persistent Plant Viruses: Molecular Hitchhikers or Epigenetic Elements?, p 177-186. *In* Witzany G (ed), Viruses: Essential Agents of Life. Springer Netherlands, Dordrecht.

7. Roossinck MJ, Saha P, Wiley GB, Quan J, White JD, Lai H, Chavarria F, Shen GA, Roe BA. 2010. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol Ecol 19:81-88.

8. Roossinck MJ. 2013. Plant Virus Ecology. PLoS Pathog 9:e1003304.

9. Roossinck MJ. 2015. Plants, viruses and the environment: Ecology and mutualism. Virology 479-480:271-277.

10. Malmstrom CM, Melcher U, Bosque-Pérez NA. 2011. The expanding field of plant virus ecology: Historical foundations, knowledge gaps, and research directions. Virus Res 159:84-94.

11. Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333.

12. Adams IP, Glover RH, Monger WA, Mumford R, Jackeviciene E, Navalinskiene M, Samuitiene M, Boonham N. 2009. Next-generation sequencing and metagenomic analysis:

a universal diagnostic tool in plant virology. Mol Plant Pathol 10:537-45.

13. Donaire L, Wang Y, Gonzalez-Ibeas D, Mayer KF, Aranda MA, Llave C. 2009. Deep-sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes. Virology 392:203-214.

14. Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, Barker I, Simon R. 2009. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388:1-7.

15. Jones S, Baizan-Edge A, MacFarlane S, Torrance L. 2017. Viral Diagnostics in Plants Using Next Generation Sequencing: Computational Analysis in Practice. Front Plant Sci 8:1770-1770.

16. Al Rwahnih M, Daubert S, Golino D, Islas C, Rowhani A. 2015. Comparison of next-generation sequencing versus biological indexing for the optimal detection of viral pathogens in grapevine. Phytopathology 105:758-763.

17. Wolf YI, Kazlauskas D, Iranzo J, Lucia-Sanz A, Kuhn JH, Krupovic M, Dolja VV, Koonin EV. 2018. Origins and Evolution of the Global RNA Virome. MBio 9.

18. Zhang YZ, Chen YM, Wang W, Qin XC, Holmes EC. 2019. Expanding the RNA Virosphere by Unbiased Metagenomics. Annu Rev Virol doi:10.1146/annurev-virology-092818-015851.

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

20. Rott M, Xiang Y, Boyes I, Belton M, Saeed H, Kesanakurti P, Hayes S, Lawrence T, Birch C, Bhagwat B, Rast H. 2017. Application of Next Generation Sequencing for Diagnostic Testing of Tree Fruit Viruses and Viroids. Plant Dis 101:1489-1499.

21. Villamor DEV, Ho T, Al Rwahnih M, Martin RR, Tzanetakis IE. 2019. High Throughput Sequencing For Plant Virus Detection and Discovery. Phytopathology doi:10.1094/phyto-07-18-0257-rvw:Phyto07180257rvw.

22. Marais A, Faure C, Bergey B, Candresse T. 2018. Viral Double-Stranded RNAs (dsRNAs)

from Plants: Alternative Nucleic Acid Substrates for High-Throughput Sequencing. Methods Mol Biol 1746:45-53.

23. Roossinck MJ, Martin DP, Roumagnac P. 2015. Plant Virus Metagenomics: Advances in Virus Discovery. Phytopathology 105:716-27.

24. Susi H, Filloux D, Frilander MJ, Roumagnac P, Laine A-L. 2019. Diverse and variable virus communities in wild plant populations revealed by metagenomic tools. PeerJ 7:e6140.

25. Pooggin MM. 2018. Small RNA-Omics for Plant Virus Identification, Virome Reconstruction, and Antiviral Defense Characterization. Front Microbiol 9:2779-2779.

26. Filloux D, Dallot S, Delaunay A, Galzi S, Jacquot E, Roumagnac P. 2015. Metagenomics Approaches Based on Virion-Associated Nucleic Acids (VANA): An Innovative Tool for Assessing Without A Priori Viral Diversity of Plants. Methods Mol Biol 1302:249-57.

27. Filloux D, Fernandez E, Comstock JC, Mollov D, Roumagnac P, Rott P. 2018. Viral Metagenomic-Based Screening of Sugarcane from Florida Reveals Occurrence of Six Sugarcane-Infecting Viruses and High Prevalence of Sugarcane yellow leaf virus. Plant Dis 102:2317-2323.

28. Palanga E, Filloux D, Martin DP, Fernandez E, Gargani D, Ferdinand R, Zabré J, Bouda Z, Neya JB, Sawadogo M, Traore O, Peterschmitt M, Roumagnac P. 2016. Metagenomic-Based Screening and Molecular Characterization of Cowpea-Infecting Viruses in Burkina Faso. PLoS One 11:e0165188.

29. 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. Front Microbiol 8:1998-1998.

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

31. Wylie SJ, Luo H, Li H, Jones MGK. 2012. Multiple polyadenylated RNA viruses detected in pooled cultivated and wild plant samples. Arch Virol 157:271-284.

32. Boone M, De Koker A, Callewaert N. 2018. Capturing the 'ome': the expanding molecular

toolbox for RNA and DNA library construction. Nucleic Acids Res 46:2701-2721.

33. Bernardo P, Charles-Dominique T, Barakat M, Ortet P, Fernandez E, Filloux D, Hartnady P, Rebelo TA, Cousins SR, Mesleard F, Cohez D, Yavercovski N, Varsani A, Harkins GW, Peterschmitt M, Malmstrom CM, Martin DP, Roumagnac P. 2018. Geometagenomics illuminates the impact of agriculture on the distribution and prevalence of plant viruses at the ecosystem scale. ISME J 12:173-184.

34. Illumina. 2017. Effects of index misassignment on multiplexing and downstream analysis. https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf.

35. van der Valk T, Vezzi F, Ormestad M, Dalén L, Guschanski K. 2019. Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. Mol Ecol Resour 2019; 00: 1-11. doi:10.1111/1755-0998.13009.

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

37. Candresse T, Filloux D, Muhire B, Julian C, Galzi S, Fort G, Bernardo P, Daugrois J-H, Fernandez E, Martin DP, 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.

38. Kutnjak D, Rupar M, Gutierrez-Aguirre I, Curk T, Kreuze JF, Ravnikar M. 2015. Deep Sequencing of Virus-Derived Small Interfering RNAs and RNA from Viral Particles Shows Highly Similar Mutational Landscapes of a Plant Virus Population. J Virol 89:4760.

39. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. 2008. NCBI BLAST: a better web interface. Nucleic Acids Res 36:W5-9.

40. Simmonds P. 2015. Methods for virus classification and the challenge of incorporating metagenomic sequence data. J Gen Virol 96:1193-206.

41. Adams MJ, Adkins S, Bragard C, Gilmer D, Li D, MacFarlane SA, Wong S-M, Melcher U, Ratti C, Ryu KH, Consortium IR. 2017. ICTV Virus Taxonomy Profile: Virgaviridae. J Gen Virol 98:1999-2000.

42. Truve E, Fargette D. 2011. ICTV Virus Taxonomy Profile: Sobemovirus. ICTV 9th Report.

43. Lefebvre M, Theil S, Ma Y, Candresse T. 2019. The VirAnnot pipeline: a resource for automated viral diversity estimation and operational taxonomy units (OTU) assignation for virome sequencing data. Phytobiomes Journal doi:10.1094/PBIOMES-07-19-0037-A.

44. Maclot F, Candresse T, Filloux D, Roumagnac P, Massart S. 2019. Effect of species composition on virome diversity in various ecosystemic communities of Poaceae, abstr Rencontres de Virologie Végétale (RVV 2019), Aussois, France,

45. Weber F, Wagner V, Rasmussen SB, Hartmann R, Paludan SR. 2006. Double-Stranded RNA Is Produced by Positive-Strand RNA Viruses and DNA Viruses but Not in Detectable Amounts by Negative-Strand RNA Viruses. J Virol 80:5059.

46. Kesanakurti P, Belton M, Saeed H, Rast H, Boyes I, Rott M. 2016. Screening for plant viruses by next generation sequencing using a modified double strand RNA extraction protocol with an internal amplification control. J Virol Methods 236:35-40.

47. Roumagnac P, Mollov D, Daugrois J, Filloux D. 2018. Viral metagenomics and sugarcane pathogens. Achieving Sustainable Cultivation of Sugarcane P Rott, ed Burleigh Dodds Science Publishing, Cambridge:183-200.

48. Casjens S. 2003. Prophages and bacterial genomics: what have we learned so far? Mol Microbiol 49:277-300.

49. de Sousa AL, Maués D, Lobato A, Franco EF, Pinheiro K, Araújo F, Pantoja Y, da Costa da Silva AL, Morais J, Ramos RTJ. 2018. PhageWeb - Web Interface for Rapid Identification and Characterization of Prophages in Bacterial Genomes. Frontiers in genetics 9:644-644.

50. Newton A, Gravouil C, Fountaine J. 2010. Managing the ecology of foliar pathogens: ecological tolerance in crops. Ann Appl Biol 157:343-359.

51. Karlsson I, Friberg H, Steinberg C, Persson P. 2014. Fungicide effects on fungal

community composition in the wheat phyllosphere. PLoS One 9:e111786-e111786.

52. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10-12.

53. Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32:2847-9.

54. Bateman A, Smart A, Luciani A, Salazar GA, Mistry J, Richardson LJ, Qureshi M, El-Gebali S, Potter SC, Finn RD, Eddy SR, Sonnhammer EL L, Piovesan D, Paladin L, Tosatto SC E, Hirsh L. 2018. The Pfam protein families database in 2019. Nucleic Acids Res 47:D427-D432.

55. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-402.

56. Huerta-Cepas J, Bork P, Serra F. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol 33:1635-1638.

57. Murtagh F, Legendre P. 2014. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J Classif 31:274-295.

58. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin P, O'hara R, Simpson G, Solymos P, Stevens M, Wagner H. 2017. vegan: Community Ecology Package. R package version 2.3–0. 2015.

59. Clarke KR. 1993. Non-parametric multivariate analyses of changes in community structure. Aust J Ecol 18:117-143.

60. Buttigieg PL, Ramette A. 2014. A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. FEMS Microbiol Ecol 90:543-50.

TABLE 1 Comparisons of high throughout sequencing (HTS) average output and virus richness per library, based on normalized datasets obtained for dsRNA and VANA targets

Approach	*Length of reads**	*Percent reads in contigs*	*Number of contigs*	*N50*	*Percent viral contigs*	*Percent reads in viral contigs*	*Viral families identified*	*Viral RdRp OTUs[α]*
dsRNA	120.9 ± 1.3	80.4% ± 4.3%	614 ± 128	796 ± 110.0	33.3% ± 7.3%	49.9% ± 14.3%	13.3 ± 1.7	39.8 ± 13.4
VANA	121.2 ± 1.5	63.4% ± 12.4%	565 ± 121	578 ± 114.9	20.7% ± 10.3%	40.5% ± 16.6%	9.3 ± 2.6	13.3 ± 6.2
P-value[#]	0.6075	**0.0009**	0.4393	**4.185e-06**	**1.296e-06**	0.2193	**0.0001**	**1.273e-05**

Significance value was calculated using a paired *t*-test; bold text indicates a statistically significant difference at the 5% level.

***** Average length of reads was in nucleotides

[α] Operational Taxonomy Units (OTUs) were defined by clustering, using a 10% distance cut-off for contigs encoding virus-specific conserved RNA dependent RNA polymerases protein motifs.

TABLE 2 Virome characteristics in the six different study sites based on RdRp OTUs.

Site	VO	CT	IT	SP	INRA	BP
Ecosystem type	***Cultivated***	***Cultivated***	***Cultivated***	***Unmanaged***	***Unmanaged***	***Unmanaged***
Viral families	11	10	10	9	9	13
Number of OTUs	25	54	71	49	64	68
Site-specificity						
Site-specific OTUs	12	28	29	35	31	46
Percentage of site-specific OTUs	48.0%	51.9%	40.8%	71.4%	48.4%	67.6%
Novelty						
Putative novel OTUs (RdRp identity <90%)	20	41	64	49	54	63
Percentage of novel OTUs	80.0%	75.9%	90.1%	100.0%	84.4%	92.6%

FIGURES LEGENDS

FIG 1 Heatmap showing the number of reads corresponding to 28 viral families in each library, as estimated from the results of BlastN and BlastX analyses. The library names and sampling sites are indicated on the left side, viral families are indicated below. Viral families are colored-coded orange (dsRNA viral families), purple (ssRNA viral families), red (ssDNA families), green (dsDNA viral families) and blue (retro-transcribing viral family). Cells color intensity is proportional to the number of reads, following the scale on the right.

FIG 2 Virus richness and known/novel status assessed at both family and Operational Taxonomic Unit (OTU) levels using dsRNA or VANA approaches. (A), (B) and (C) Scaled Venn diagrams showing the number of OTUs discovered using dsRNA or VANA approaches and a 10% divergence criterion for OTU definition **(A)** or using 3% **(B)** and 20% **(C)** divergence criteria, respectively. **(D)** Scaled Venn diagrams showing the number and identity of OTU families discovered using dsRNA or VANA approaches. **(E) and (F)** Pie charts illustrating for the dsRNA **(E)** and VANA **(F)** approaches the proportions and known or novel RNA-dependent RNA polymerase (RdRp) OTUs for dsRNA viruses, ssRNA viruses and others (unclassified viruses and virus-associated RNAs)

FIG 3 Comparison of the viral diversity identified at each individual sampling site using the dsRNA and VANA approaches. The bar chart shows the RdRp OTU-based virome composition for the different viral families using dsRNA and VANA approaches. (**B)** Scaled Venn diagrams showing the number of RdRp OTUs discovered by either the dsRNA (light blue) or VANA (light orange) approaches or by both approaches simultaneously.

FIG 4 Dissimilarity analyses of the RdRp OTU virome composition between sites. The dissimilarity (distance) matrix was calculated using a Jaccard method on OTUs presence/absence data. Hierarchical clustering dendrograms of the 12 dsRNA libraries **(A)** and of the 12 VANA

libraries **(B)** corresponding to the 6 sampling sites were prepared using hclust and "complete" algorithm. **(C)** Non-metric multi-dimensional scaling (NMDS) of a Jaccard distance matrix generated using the presence/absence data of all dsRNA and VANA libraries. Circles represent the dsRNA approach libraries and triangles the VANA approach ones. The symbols are colored-coded according to the sampling site.

Figure 1

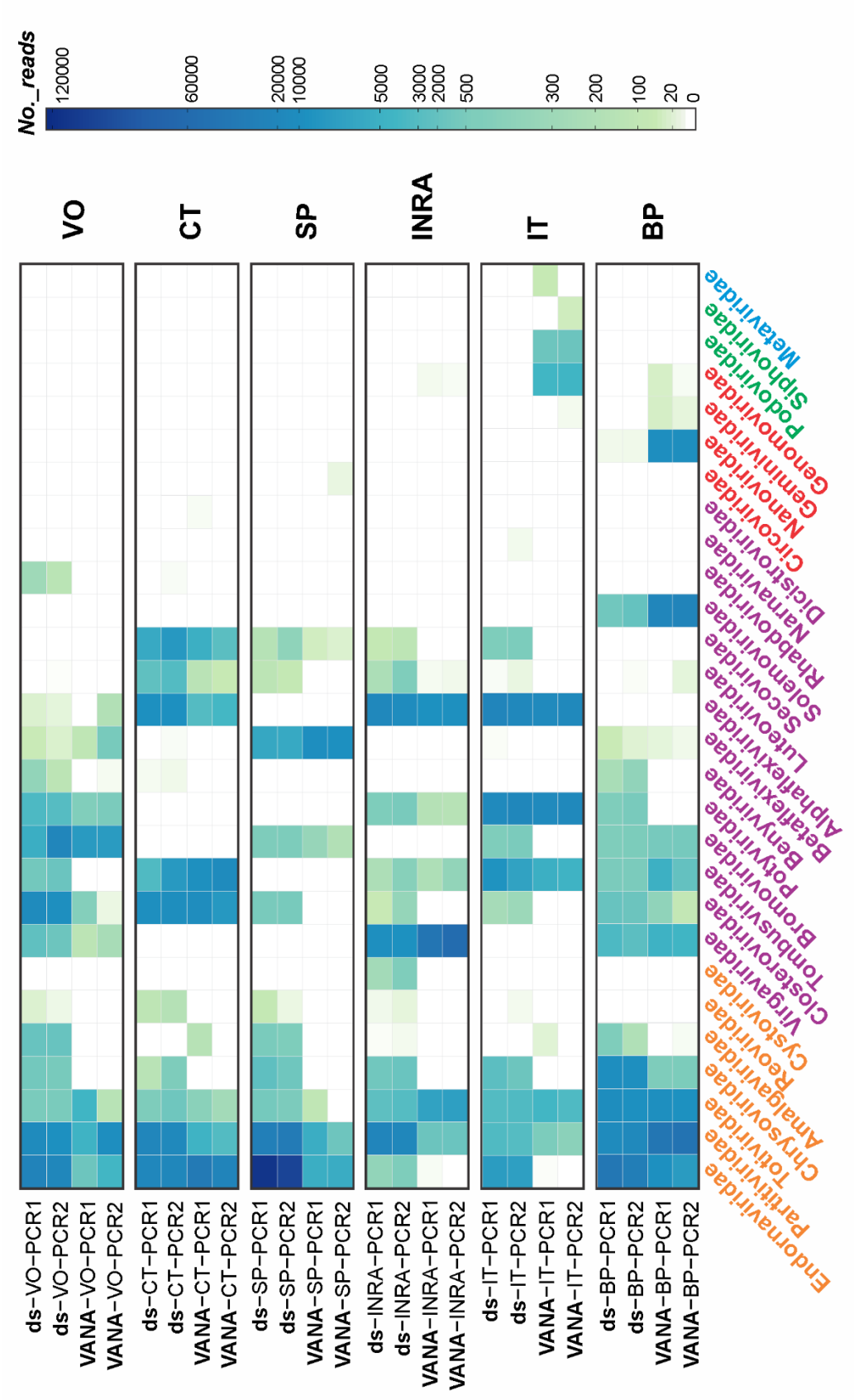

No._reads
120000
60000
20000
10000
5000
3000
2000
500
300
200
100
20
0
VO
CT
SP
INRA
IT
BP
ds−VO−PCR1
ds−VO−PCR2
VANA−VO−PCR1
VANA−VO−PCR2
ds−CT−PCR1
ds−CT−PCR2
VANA−CT−PCR1
VANA−CT−PCR2
ds−SP−PCR1
ds−SP−PCR2
VANA−SP−PCR1
VANA−SP−PCR2
ds−INRA−PCR1
ds−INRA−PCR2
VANA−INRA−PCR1
VANA−INRA−PCR2
ds−IT−PCR1
ds−IT−PCR2
VANA−IT−PCR1
VANA−IT−PCR2
ds−BP−PCR1
ds−BP−PCR2
VANA−BP−PCR1
VANA−BP−PCR2
Endornaviridae
Partitiviridae
Totiviridae
Chrysoviridae
Amalgaviridae
Reoviridae
Cystoviridae
Virgaviridae
Closteroviridae
Tombusviridae
Bromoviridae
Potyviridae
Benyviridae
Betaflexiviridae
Alphaflexiviridae
Luteoviridae
Secoviridae
Solemoviridae
Rhabdoviridae
Nertraviridae
Dicistroviridae
Circoviridae
Nanoviridae
Geminiviridae
Genomoviridae
Podoviridae
Siphoviridae
Metaviridae

Figure 2

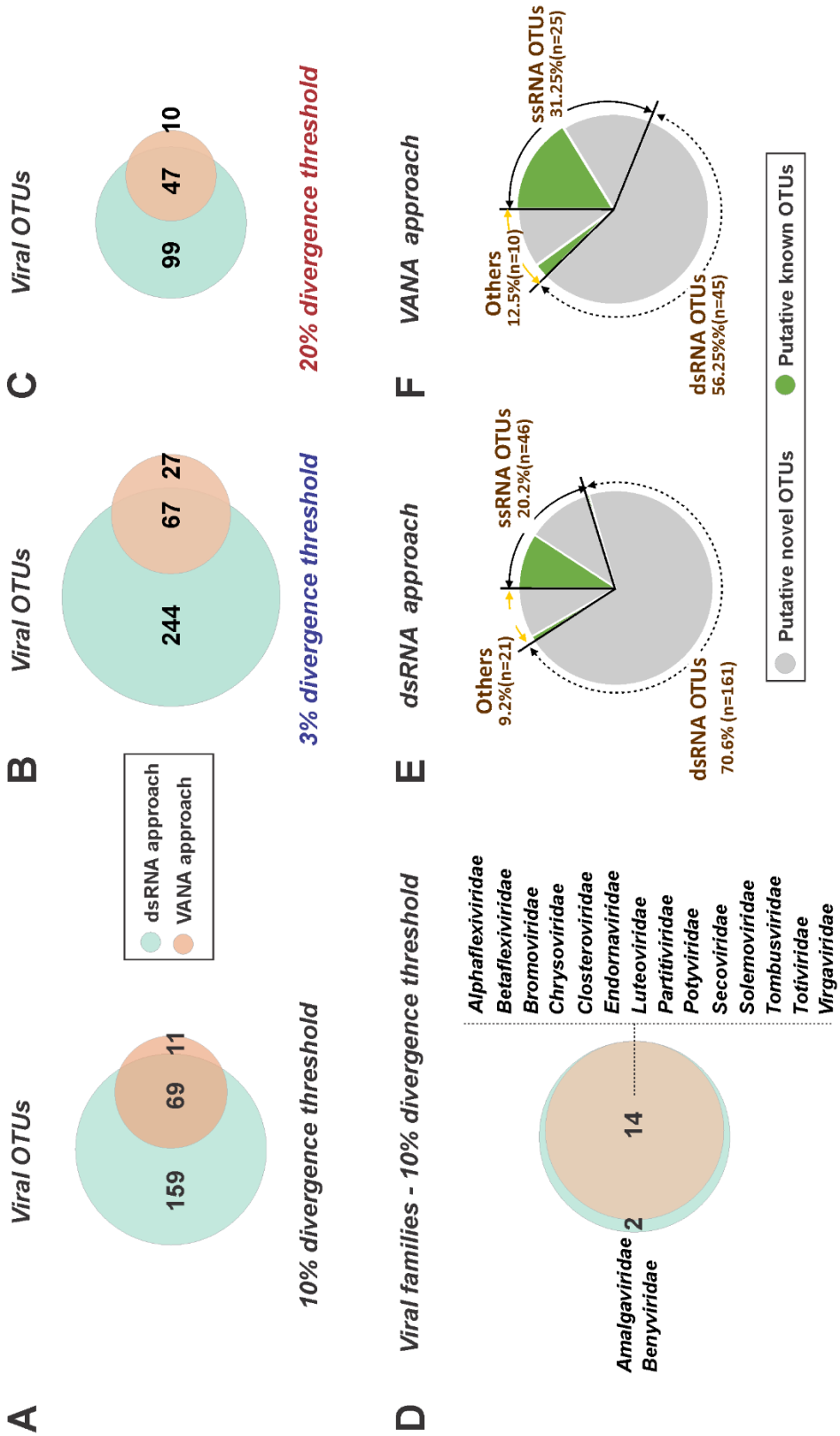

A
Viral OTUs
159
69
11
10% divergence threshold
B
Viral OTUs
244
67
27
3% divergence threshold
C
Viral OTUs
99
47
10
20% divergence threshold
dsRNA approach
VANA approach
D
Viral families - 10% divergence threshold
Alphaflexiviridae
Betaflexiviridae
Bromoviridae
Chrysoviridae
Closteroviridae
Endornaviridae
Luteoviridae
Partitiviridae
Potyviridae
Secoviridae
Solemoviridae
Tombusviridae
Totiviridae
Virgaviridae
Amalgaviridae
Benyviridae
2
14
E
dsRNA approach
Others
9.2%(n=21)
ssRNA OTUs
20.2%(n=46)
dsRNA OTUs
70.6% (n=161)
F
VANA approach
Others
12.5%(n=10)
ssRNA OTUs
31.25%(n=25)
dsRNA OTUs
56.25%%(n=45)
Putative novel OTUs
Putative known OTUs

Figure 3

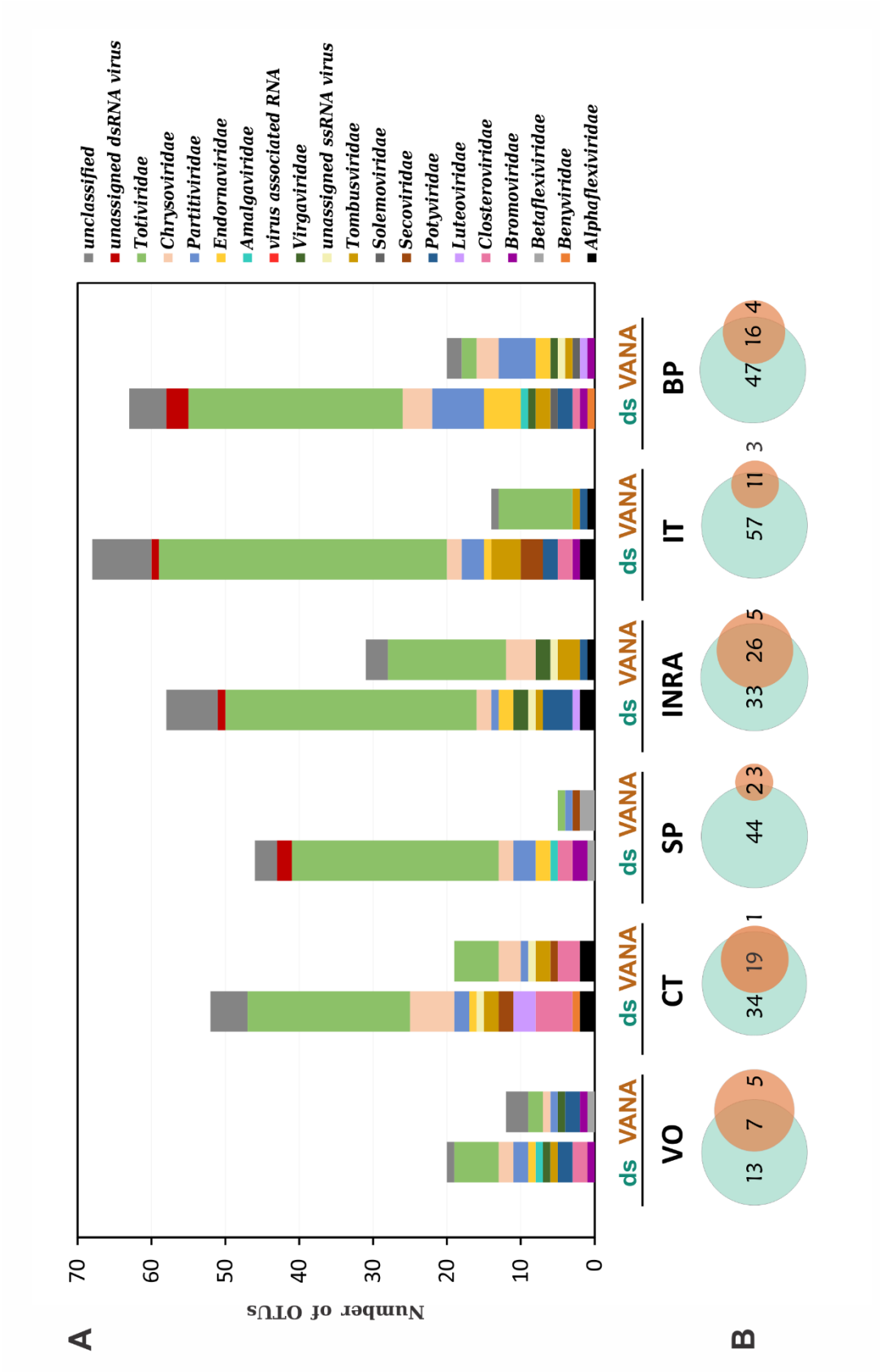
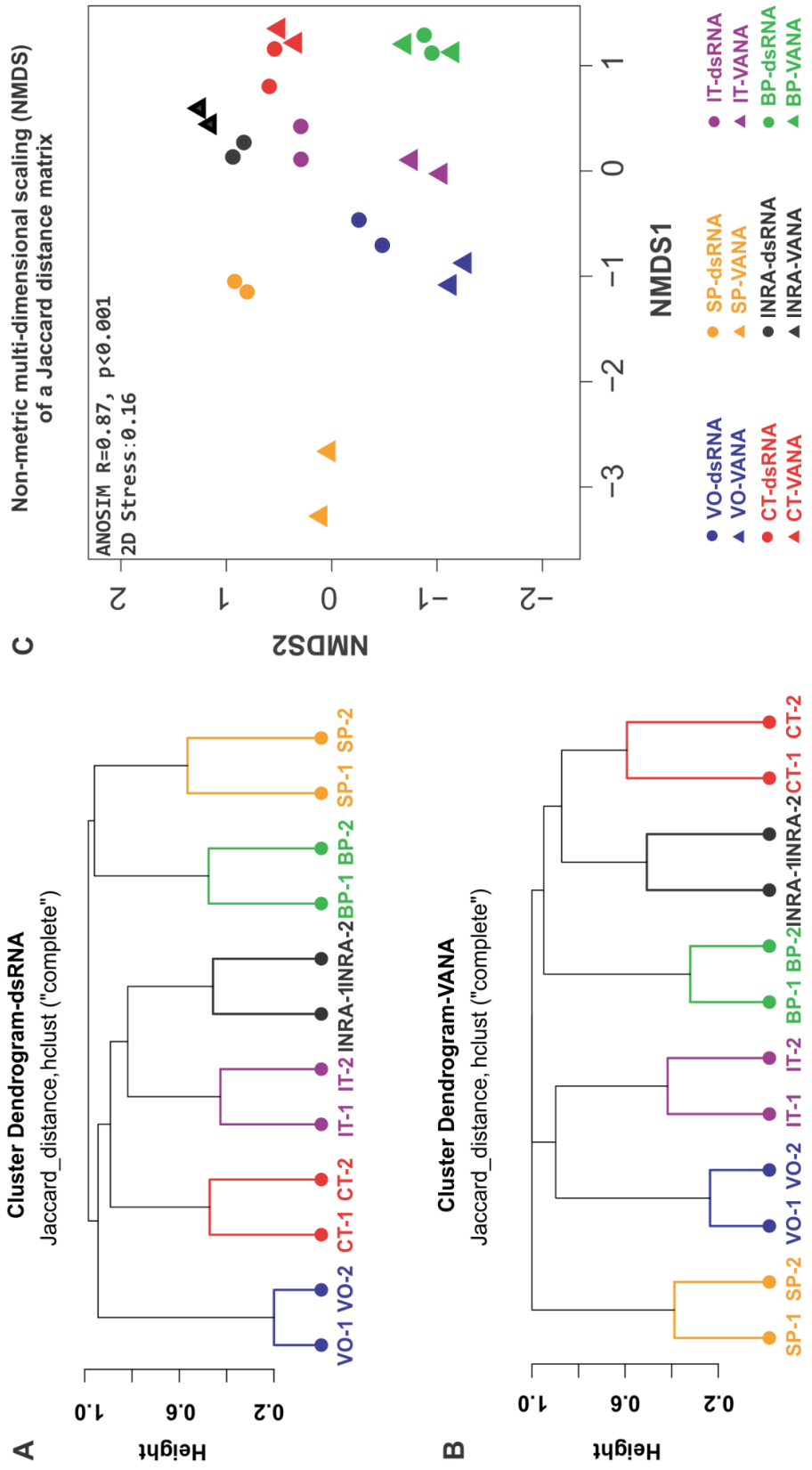

A
Number of OTUs
0
10
20
30
40
50
60
70
unclassified
unassigned dsRNA virus
Totiviridae
Chrysoviridae
Partitiviridae
Endornaviridae
Amalgaviridae
virus associated RNA
Virgaviridae
unassigned ssRNA virus
Tombusviridae
Solemoviridae
Secoviridae
Potyviridae
Luteoviridae
Closterovirinae
Bromoviridae
Betaflexiviridae
Benyviridae
Alphaflexiviridae
ds VANA
VO
CT
SP
INRA
IT
BP
B
13
7
5
34
19
1
44
23
33
26
5
57
11
3
47
16
4

Figure 4


A
Cluster Dendrogram-dsRNA
Jaccard_distance, hclust ("complete")
Height
0.2
0.6
1.0
VO-1 VO-2 CT-1 CT-2 IT-1 IT-2 INRA-1INRA-2 BP-1 BP-2 SP-1 SP-2
B
Cluster Dendrogram-VANA
Jaccard_distance, hclust ("complete")
Height
0.2
0.6
1.0
SP-1 SP-2 VO-1 VO-2 IT-1 IT-2 BP-1 BP-2INRA-1INRA-2CT-1 CT-2
C
Non-metric multi-dimensional scaling (NMDS)
of a Jaccard distance matrix
ANOSIM R=0.87, p<0.001
2D Stress:0.16
NMDS1
NMDS2
VO-dsRNA
VO-VANA
CT-dsRNA
CT-VANA
SP-dsRNA
SP-VANA
INRA-dsRNA
INRA-VANA
IT-dsRNA
IT-VANA
BP-dsRNA
BP-VANA

Supplementary Materials [Supplementary Tables donot fit easily in an A4 format and are available at https://doi.org/10.15454/OAKDVI]

TABLE S1 Characteristics of the high throughput datasets, contigs assembly and annotations results obtained for the various sites using the dsRNA and VANA approaches.

TABLE S2 Number of reads observed in each library for 20 viral families.

TABLE S3 Identified viral conserved protein pfam database motifs and number of corresponding OTUs and virus families observed.

TABLE S4 Taxonomy of identified OTUs and their prevalences among six detected sites.

TABLE S5 Statistics for the comparisons of data output from duplicate PCRs libraries for the different analyzed sampling site/approach analyzed.

TABLE S6 Identity of the 110 plant species making up the composite samples collected from the six sampling sites.

CHAPTER Ⅲ

Different patterns in leaf-associated viromes and mycobiomes of wild plant populations between cultivated and unmanaged environments

Different patterns in leaf-associated viromes and mycobiomes of wild plant populations between cultivated and unmanaged environments

Yuxin Ma[1], Tania Fort[2], Armelle Marais[1], Marie Lefebvre[1], Sébastien Theil[1,3], Corinne Vacher[2] and Thierry Candresse[1]*

[1] UMR 1332 BFP, INRA, Univ. Bordeaux, CS20032, 33882 Villenave d'Ornon cedex, France

[2] UMR BioGeCo, INRA, Univ.Bordeaux, 33615 Pessac, France

Summary: 196 words

Text: 6527 words (Introduction- Materials and Methods-Results-Discussion)

* Corresponding author thierry.candresse@inra.fr

[3] Current address INRA UMRF, 20, côte de Reyne, 15000 Aurillac, France

SUMMARY

- Plants are colonized by complex microbial communities (bacteria, fungi, viruses and other microorganisms) that affect plant growth and survival as well as ecosystem functions. However, research on leaf-associated fungal and viral communities, especially in wild plants, has been so far fairly limited.
- Using metagenomics approaches to characterize the plant core microbiome, we assessed the richness and composition of leaf-associated fungal and viral communities from complex pools of herbaceous wild plants collected in cultivated and unmanaged ecosystems.
- We identified 161 fungal families and 18 viral families comprising 249 RNA dependent RNA polymerase-based operational taxonomy units (RdRp OTUs) from the leaf samples. Fungal culturomics captured ca. 9% of the fungal diversity but there was virtually no correlation between the plant samples virome and that from fungal cultures.
- Mycobiome and, more markedly, virome composition showed a strong site specificity. The mycobiomes were more diverse in unmanaged sites while the plant associated viromes showed a higher family-level richness in cultivated sites, suggesting that mycobiome and virome are shaped by different drivers. Further efforts will be needed to confirm these trends in other settings and to begin to unravel the drivers contributing to the structuring of plant-associated fungal and viral populations.

Keywords: Mycobiome, mycovirus, phytovirus, plant, virome

INTRODUCTION

Plant leaves represent one of the largest microbial habitats on Earth (Morris, 2001) and harbor hyperdiverse microbial communities including bacteria, archaea, fungi, yeasts and viruses (Lindow & Brandl, 2003; Vorholt, 2012; Koskella, 2013; Vacher *et al.*, 2016). These microbial communities influence plant health (Arnold *et al.*, 2003; Hacquard *et al.*, 2017; Saleem *et al.*, 2017), plant nitrogen nutrition (Fürnkranz *et al.*, 2008; Moyes *et al.*, 2016; Doty, 2017), ecosystem primary productivity (Laforest-Lapointe *et al.*, 2017) and biogeochemical cycles (Osono, 2006; Morris *et al.*, 2014; Bringel & Couée, 2015). Despite their potential importance for natural and managed ecosystems, our knowledge of leaf microbial communities remains very limited to date, compared to that of root and rhizosphere communities. Experimental studies are needed to better understand the functions of leaf microbial communities (Rosado *et al.*, 2018) and predict their dynamics in response to global change (Laforest-Lapointe & Whitaker, 2019). Viruses and fungi, in particular, should be better integrated in future studies of the leaf microbiome (Laforest-Lapointe & Whitaker, 2019).

Leaf-associated viruses include viruses of plants (phytoviruses), viruses of fungi (mycoviruses), and viruses of bacteria and archea (phages). A large fraction of plant-associated viruses in wild plants are double-stranded RNA viruses infecting either plants (persistent viruses) or fungi (mycoviruses) or possibly both. It should be stressed that a recurrent question in plant-associated virome analysis, in particular when it comes to dsRNA viruses, concerns the identity of the host(s) of the identified viruses (Roossinck, 2015a), so that the proportion of phytoviruses to mycoviruses in plant-associated viromes is still unknown (Roossinck, 2015a). Phytovirus diversity has been grossly underestimated so far, as illuminated recently by high throughput sequencing (HTS)-based metagenomics studies (Roossinck *et al.*, 2010; Rosario & Breitbart, 2011; Roossinck, 2012; Shi *et al.*, 2016). Indeed, the newly released 2018b ICTV taxonomy lists a total of 4958 virus species, however only about 1337 of them (~27.0%) are plant viruses (Siddell *et al.*, 2019). This

percentage is to be compared with the ca. 5500 known mammalian species contrasted with the estimated ca. 400.000 described plant species and ca. 2.000 new plant species described each year. One of the reasons for this situation is that viruses have been traditionally thought as pathogens, which led to biased studies largely focusing on viruses causing visible symptoms in economically important crops (Wren *et al.*, 2006) which comprise only a minute fraction of all plant species. In addition, there is now evidence that only a small fraction of viruses are associated with obvious disease (Roossinck, 2005). Taken together, these biases suggest that there exists a huge gap in our understanding of plant virus diversity, evolution and ecology.

To bridge this gap some recent studies have started to analyze virus populations with a focus on wild plant populations (Thapa *et al.*, 2015; Bernardo *et al.*, 2018; Susi *et al.*, 2019), which are likely to represent reservoirs for both known viruses and novel, emerging agents (Anderson *et al.*, 2004; Cooper & Jones, 2006; Elena *et al.*, 2014; Stobbe & Roossinck, 2014; McLeish *et al.*, 2019). The few such metagenomic studies performed to date have shown that virus infection is quite common and often asymptomatic in wild plants (Roossinck, 2012). As longtime neglected plant populations, wild plants can be an important research object for the exploration of viral diversity at individual species, plant population or even ecosystem scales.

Both viromes and mycobiomes are shaped by abiotic and biotic environmental factors. In woody perennials, for example, Jumpponen and Jones's studies showed that phyllosphere fungal communities in temperate *Quercus macrocarpa* appear distinct between trees in urban and nonurban environments, possibly as a consequence of geographic distance, air pollution, human management etc. (Jumpponen & Jones, 2009; Jumpponen & Jones, 2010). Phyllosphere fungal richness of *Quercus ilex* in a mixed Mediterranean forest increased with summer season and long-term drought (Penuelas *et al.*, 2012) while in *Fagus sylvatica* mycobiome composition was significantly correlated with elevation gradient and microclimate (Cordier *et al.*, 2012). More recently, host genotype has been shown to be an important determinant of phyllosphere

mycobiomes for a wide range of cereal (Sapkota *et al.*, 2015). The study of the wheat canopy mycobiome showed geographical location as a major factor along with leaf position, growth stage and cultivar identity (Sapkota *et al.*, 2017).

Virus communities may influence microbial ecosystems by modulating microbial population size, diversity, metabolic output, and gene flow (Brum *et al.*, 2015). However, the study of phytoviral communities in different ecosystems and of the potential drivers shaping them was only initiated very recently (Roossinck, 2015b; Thapa *et al.*, 2015). The environmental factors determining plant virus community composition in both cultivated and natural ecosystems thus remain largely unknown (Malmstrom *et al.*, 2011). One recent study investigated virome composition in six native plant species from 20 sites over 4 years to test the effects of host identity, location, and sampling year (Thapa *et al.*, 2015). The results showed that only host species identity was significantly correlated with virome composition. Recent results also suggest that some viral families might be more adapted to cultivated or to unmanaged environments (Bernardo *et al.*, 2018). A few other studies focusing on specific plant or virus species have shown that the latitude, environment and temporal dynamics were correlated with variations in virome composition (Coutts & Jones, 2002; Cadle-Davidson & Bergstrom, 2004; Seabloom *et al.*, 2010).

In most cases, the results of the interaction between viruses and fungi remains unknown. There are indications that many mycoviruses induce no obvious infection phenotype in their fungal host, so that they therefore appear to be largely latent. On the other hand, some mycoviruses have been shown to impact positively or negatively (hypovirulence) the fitness and pathogenicity of their fungal host(s) (Nuss, 2008; Ghabrial *et al.*, 2015). The interactions between phytoviruses and plant-associated fungi are even more poorly described. There is however some circumstantial evidence that fungi-plant interactions might impact plant-phytovirus ones. For example, the mycorhizal fungal species *Piriformospora indica* is reported to interfere with pepino mosaic virus (PepMV) accumulation in tomato apical shoots in a light intensity-dependent fashion (Fakhro *et*

al., 2010) while tomato colonization by PepMV could also be inhibited by *Verticillium* spp. or by the oomycete *Pythium aphanidermatum* colonization (Spence *et al.*, 2006; Schwarz *et al.*, 2010). Conversely, virus occurrence and symptomatology were increased by arbuscular mycorrhizal fungi colonization in tobacco (Shaul *et al.*, 1999).

In the present study we used metagenomics-based approaches to assess the diversity of leaf-associated viruses and fungi in wild plant populations sampled in cultivated and unmanaged sites. A dissimilarity analysis of the obtained mycobiomes and viromes composition was used to evaluate the factors potentially affecting community composition. In particular, an influence of the anthropic status (cultivated vs unmanaged) of the sampling site was evidenced for both mycobiome and virome.

MATERIALS AND METHODS

Study sites and sampling design

Herbaceous wild plants and weeds were sampled in 2017 in four sites in southwest France. Two sites (VO and CT) were cultivated, horticultural agroecosystems in which vegetable crops were grown. The VO site harbored a range of crops including lettuce, spinach, pepper, turnip etc… while the CT site mostly had carrots. Two other sites (INRA and SP) were unmanaged dry grasslands (Table S1). In each site, a total of ca. 200 individual plants comprising the 29 to 40 locally most abundant species were collected in spring 2017 (Table S1). An identical number of plants was collected for each sampled species. As the focus of the study was on wild plants, crop plants were not sampled in the cultivated sites. Individual plants were selected at random but plants with necrotic tissues or with insect infestation were not collected. Plants were identified in the field by experienced researchers with botanical knowledge and subsequently stored in a cool ice chest before being brought back and processed in the laboratory.

Leaf processing for phytovirome and mycobiome analysis

For each site, 4 complex pools of ca. 50 plants were assembled for nucleic acids extraction from the ca. 200 collected plants, taking care to allocate all individual plants for a given species to a single pool (Fig. 1). For virome description, the complex pools were assembled using small fragments corresponding to 0.1 g fresh tissues of each individual plant (ca. 5 g total). For mycobiome description, the complex pools were assembled using 0.5 mg of leaf blade dessicated over anhydrous calcium chloride dry weight (ca. 25 mg).

Fungi culturing for culturome and mycovirome analysis

In order to culture out fungi from the plant samples collected at each sampling site, a modified dilution strategy (Unterseher & Schnittler, 2009) was used. For each site, the sampled 200 individual plants were divided in ca. 7-10 pools (with no shared plant species) of 20 to 30 individual leaf pieces (approximately 1 cm^2) which were added to sterile Erlenmeyer flasks containing 15 ml sterile water with 0.1% Tween 20 (Fig. 1). The flasks were then incubated at room temperature on an orbital shaker for 20 minutes before filtering with sterile gauze. Based on pilot experiments, the filtered solution was then serially diluted 10, 100 and 1000 times and 500 µl aliquots used to inoculate respectively 10 plates of malt agar (MA) and of potato dextrose agar (PDA) containing 0.025% chloramphenicol. Plates were incubated at 22°C and observed regularly for development of fungal colonies. All developing fungal colonies were isolated from the original plates and transferred to new petri dishes (4 isolates per plate) containing culture media covered by cellophane in order to facilitate the final collection of mycelia. Grown mycelia (ca. 3.5 cm in diameter) were recovered, transferred to plastic tubes and lyophilized. In this way a total of 480 to 1279 fungal colonies were obtained for each of the sampling sites. For dsRNA extraction for mycovirome characterization mycelia were assembled in pools of 480 to 640 colonies (ca. 1 mg dry weight per mycelium for a total of ca. 0.48-0.64 g dry weight per pool) while for DNA

extraction for mycobiome analysis, the mycelia were assembled in pools of ca. 250 colonies (ca. 0.1 mg dry weight per mycelium for a total of 25 mg per pool) (Fig. 1).

Double stranded RNAs extraction, whole genome amplification (WGA) and Illumina sequencing

Double stranded RNAs (dsRNAs) were extracted from each plant and fungal pool by two rounds of CF11 cellulose chromatography as described by Marais *et al.* (2018). A blank control using only reagents was prepared in parallel with every extraction. For conversion to cDNA and random amplification of dsRNAs, 3 µl of purified dsRNAs were denatured at 99°C for 5 min and submitted to a reverse transcription step initiated by a mixture of primers consisting of 1 mM dT18 and 2 mM PcDNA12 (5' TGTGTTGGGTGTGTTTGGN$_{12}$ 3') using the SuperscriptII Reverse Transcriptase according to the manufacturer's instructions (Invitrogen). A random whole genome amplification (WGA) was performed using the obtained complementary DNAs (cDNAs), allowing at the same time the tagging of pools from the same site with a specific multiplex identifier (MID) adapter (Marais *et al.*, 2018). In general, a minimum of two independent amplifications with different MID tags were performed for each pool. PCR products were purified using the MinElute PCR Purification Kit (QIAGEN). Finally, for dsRNA plant samples, equal amounts of PCR products obtained from each of the 4 plant pools corresponding to a sampling site were assembled in one library integrating the 200 sampled plants. The same pooling strategy was performed on dsRNAs of fungal cultures, therefore a library integrating all the fungi corresponding to a sampling site was prepared (Fig. 1 and Table S2). The prepared libraries, including negative controls (buffer-only libraries) for plant samples and cultured fungi dsRNA extractions (Fig. 1 and Table S2) and were sequenced in multiplexed format (2×150 bp) on an Illumina HiSeq 3000 system at the GenoToul platform (INRA Toulouse, France).

Bioinformatics for virome analysis: reads cleaning, assembly, contigs annotation and Operational Taxonomic Units (OTU) identification

Virome analysis was performed using the virAnnot pipeline (Lefebvre *et al*., 2019). More precisely, the raw sequence reads were first demultiplexed and the MID tags removed using the *cutadapt* tool (Martin, 2011). To reduce the cross-talk between samples caused by index hopping (Illumina, 2017; van der Valk *et al.*, 2019), only paired-end reads with identical MID tags identified in both members of the pair were retained for the next steps. In some cases, to compensate for uneven sequencing depth between libraries, a normalization step was performed by randomly subsampling libraries to the same depth using the seqtk tool (https://github.com/lh3/seqtk). The clean paired-end reads were *de novo* assembled into contigs using the IDBA-UD assembler (https://academic.oup.com/bioinformatics/article/28/11/1420/266973). Contigs were annotated using BlastN and BlastX against the non-redundant nucleotide (nt) and protein (nr) Genbank databases with a conservative e-value cut-off of 10^{-4}. A clustering approach (Lefebvre *et al*., 2019) was used to define operational taxonomy units, following the strategy highlighted by (Simmonds, 2015). Briefly, a search of RNA-dependent RNA polymerase (RdRp) conserved motifs was performed on all contigs and those encoding a viral RdRp were retrieved and aligned with reference sequences. Distance matrices computed with the ETE3 toolkit (Huerta-Cepas *et al.*, 2016) were used to cluster in a single OTU all contigs differing by less than 10% divergence. This 10% value has been shown to generate in many viral families, OTUs that are a relatively good approximation of taxonomic species (Lefebvre *et al*., 2019). This allowed to generate an OTU table indicating for each sampling site the presence/absence and the number of reads integrated in each identified OTU.

DNA extractions and ITS1 and ITS2 amplification for fungal metabarcoding

For comparison purposes, total DNA was extracted from each prepared pool (plants or fungi) using two different kits, the PowerSoil® DNA Isolation Kit (MO BIO) and the DNeasy Plant Mini Kit (QIAGEN), according to the instructions of the kits. The internal transcribed spacer 1 and 2 regions (ITS1 and ITS2) were amplified according to (Op De Beeck *et al.*, 2014) using respectively primer pairs ITS1F/ITS2 and ITS86F/ITS4 tailed with Illumina adaptors (https://web.uri.edu/gsc/files/16s-metagenomic-library-prep-guide-15044223-b.pdf). Libraries were prepared and paired-end sequencing performed on an Illumina MiSeq sequencer using the Nano Kit v2 at the Bordeaux Genome Transcriptome Facility (INRA – Pierroton, France).

Bioinformatics and statistical analysis for mycobiomes

- **ITS sequences processing in DADA2**

Metabarcoding datasets for ITS1 and ITS2 were similarly processed. Reads were first demultiplexed and the unique barcodes/adaptors removed (Genome Transcriptome Facility, INRA Pierroton, France). To process the sequences, the ITS primers were first removed, and the sequences were then filtered, trimmed, merged and chimeras removed using the open-source software package DADA2 ITS Pipeline Workflow (1.8) (https://benjjneb.github.io/dada2/ITS_workflow.html) running in R (Callahan *et al.*, 2016a) with parameters in detail described in supporting information Methods S1 and S2. ASV taxonomic assignments were subsequently conducted with the RDP classifier (Wang *et al.*, 2007) embedded within DADA2 and trained with the UNITE general FASTA release for Fungi-version 18.11.2018 (https://dx.doi.org/10.15156/BIO/786343) (UNITE, 2019). The ASV, taxonomy tables and sample metadata tables were integrated into one phyloseq object (Methods S3). Only fungal ASVs were retained for further analyses (McMurdie & Holmes, 2013; Callahan *et al.*, 2016b). The number of reads of ASVs found in the extraction and amplification negative controls Galan *et al.* (2016) were removed from the whole dataset.

In order to be able to compare the fungal community composition between plant pools, the ASV tables were resampled to the minimum sequencing depth observed in the datasets (corresponding to 23270 reads/sample for ITS1 and 15297 reads/sample for ITS2) using the '*rarefy_even_depth*' function in the Phyloseq package in R (McMurdie & Holmes, 2013) (Methods S4).

- **Alpha and beta diversity analyses**

Alpha diversity analyses and visualization of mycobiome community composition were performed in R using the Phyloseq package (McMurdie & Holmes, 2013) for data import and richness metrics calculation and using the ggplot2 package for visualization (Wickham, 2016) (Methods S5). Dissimilarity-based hierarchical cluster analysis (HCA) and Principal coordinates analysis (PCoA) (Ramette, 2007) were performed on distance matrices estimated with the "Jaccard-binary" (Hamers *et al.*, 1989) method in R (Methods S5). Non-parametric statistical tests ANOSIM (Analysis of similarities) (Clarke, 1993) and ADONIS (Anderson, 2001; McArdle & Anderson, 2001) were used to estimate the effects of various factors (sampling site, ecosystem type, location, extraction kit) on virome and mycobiome composition (Methods S5). Given that a better performance was obtained with the ITS1 amplicons (larger number of filtered reads, richer ASV mycobiomes), only the ITS1 information was subsequently analyzed in detail. On the other hand, a comparable but slightly better performance was obtained with PowerSoil® DNA Isolation Kit (MO BIO) (2 samples with no initial amplification using the DNeasy Plant Mini Kit) so that the reads corresponding to the ITS1 amplicons obtained from DNA extracted with PowerSoil® DNA Isolation Kit were used for the downstream analysis. Overall, for each site, four mycobiomes, one phytovirome, one culturome, and one mycovirome were sequenced and analyzed (Fig. 1).

RESULTS

Phytovirome diversity and compositional variation in different environments

Overall, more viral reads and viral contigs were detected for the CT site followed in turn by the VO, SP and INRA sites (Table 1). The ratio of dsRNA/ssRNA reads ranged from 0.7 to 1.6 except for the SP site where it reached a high value of 19.6 possibly linked to a saturation of the amplification libraries by reads from *Endornaviridae* members (81.9%) (Table S3). Few reads annotated as retro-transcribing viruses were discovered only from the VO and CT sites while virus-associated RNAs reads were only detected for the INRA site. Besides the SP site, *Endornaviridae* also accounted for a large proportion of the CT virome (30.9%), followed by *Alphaflexiviridae* (17.4%) (Table S3). Based on a Blast-based annotation, a total of 17 viral families were discovered from these four sites with respectively 15 and 14 families for VO and CT, but only 11 and 7 for SP and INRA. The family-level richness of the virome thus appears to be higher for the cultivated than for the unmanaged sites, an observation associated with the absence of several single-stranded RNA (ssRNA) virus families (*Bromoviridae*, *Secoviridae*, *Virgaviridae* and *Benyviridae*) and of the *Caulimoviridae* pararetroviruses from the unmanaged sites (Table S3). On the contrary, double-stranded RNA (dsRNA) virus families *Partitiviridae* and *Totiviridae* and ssRNA families *Endornaviridae*, *Alphaflexiviridae* and *Tombusviridae* were present in all four sites (Table S3).

A viral RdRp clustering approach was used to define operational taxonomic units (OTUs) at a level close to ICTV species using the virAnnot pipeline (Lefebvre *et al*., 2019). In total, 190 viral RdRp OTUs were identified (Table S4) representing 16 viral families. Respectively 73, 50, 55 and 26 OTUs representing 13, 12, 9 and 5 viral families were respectively discovered from the VO, CT, SP and INRA sites, confirming the higher viral family richness in the cultivated sites (Table 1 and Fig. 2a). The lower number of families identified for each site results from the constraint that any virus for which the RdRp core-encoding region is missing (due to incomplete genome coverage) will not be considered by this approach. The OTU-based analysis therefore provides a

lower bound of viral diversity while allowing to analyze the virome at a rank closer to taxonomic species. Double-stranded viruses OTUs account for a larger proportion (54.0% to 70.9%) than ssRNA viruses OTUs (21.8% to 28.0%), except for the INRA site in which the situation was reversed (42.3% ssRNA OTUs vs 30.8% dsRNA ones). Between 7.3% and 26.9% of OTUs could not be annotated by Blast at family level, depending on the site (Fig. 2b). Overall, and as already seen in other phytovirome studies, a large fraction of the detected OTUs (83.7% to 96.3%) putatively correspond to novel viruses since no RdRp-encoding sequence in Genbank fullfiled the identity criterion (>=90% nt or aa identity) to be included in the corresponding OTU (Table 1 and Fig. 2b). The majority of the OTUs for which a Genbank counterpart could be identified correspond to ssRNA viruses (Table S4 and Fig. 2b).

Alpha diversity of fungal communities from mycobiome libraries in different sampling sites

Taking only into account the 16 ITS1 libraries extracted with Powersoil kit for the 4 sampling sites, a total of 1188 unique ASVs were discovered, comprising 4 phyla, 21 classes, 161 families and 247 genera. Of those unique ASVs, 361 ASVs appear to correspond to unknown/novel ASVs (at genus level) (Table S5). Fungal communities in each library were dominated by *Ascomycota* and *Basidiomycota*, with a relative abundance of 53.2%+/-21.0% and 46.4%+/-20.8%, respectively. At class level, *Dothideomycetes* and *Tremellomycetes we*re dominant ones with a relative abundance of 40.2%+/-19.1% and 39.6%+/-20.6% (Table S6).

Richer and more diverse fungal communities were observed for the unmanaged SP and INRA sampling sites (Fig. 2c) with an average of 191.3+/-54.8 and 252.3+/-34.3 ASVs per library, respectively, and with a total of 483 and 639 unique ASVs, respectively (Table S7). This translates in higher fungal diversity in these sites (Fig. 2d) with average Shannon indexes of 3.72+/-0.33 and 3.86+/-0.15, respectively, as compared to values of 3.0+/-0.41 and 2.72+/-0.41 for the cultivated VO and CT sampling sites (Table S7). This difference in fungal diversity cannot be ascribed to a higher diversity of the sampled plant species since fewer plant species were sampled in the

unmanaged sites than in the cultivated ones [respectively 29 (SP) and 33 (INRA) sampled plant species as compared to 34 (CT) and 40 (VO)].

Mycobiome composition: largely consistent within ecosystem whatever sampled host plants but specific between sampling sites

For each sampling site, 4 different plant pools were analyzed, which assemble different plant species. It is thus possible to use our data to compare the mycobiome of different plant species growing together at the same sampling site. The analysis of Venn diagrams of fungal ASVs within each sampling site show that a significant proportion of the ASVs detected are shared between plant pools with 31.1% to 42.9% (on average 35.5%+/-5.2%) of ASVs shared between at least two pools and a core of on average 12.4% +/- 4.8% of ASVs shared between all pools of a sampling site (extremes 9.2-19.3%) (Fig. S1). For these core ASVs of each site (31, 45, 53 and 59 ASVs for VO, SP, CT and INRA, respectively, Fig. S1), a significant proportion (55.3%) are shared between at least two sites, suggesting they correspond to broadly distributed fungal taxa. The 14 most common ASVs shared between all the 16 tested libraries are annotated as *Alternaria infectoria*, *Bensingtonia* sp., *Botrytis caroliniana*, *Cystofilobasidium macerans*, *Epicoccum nigrum*, *Filobasidium stepposum*, *Filobasidium wieringae*, *Holtermanniella wattica*, *Mycosphaerella tassiana*, *Sporobolomyces roseus*, *Stemphylium* sp., *Symmetrospora coprosmae*, *Vishniacozyma carnescens*, *Vishniacozyma victoriae* (Table S5).

Subsequently, the compositional dissimilarities between pools were quantified using a Jaccard metric calculated on presence/absence (binary) data of ASVs (Table S5). Principal coordinates analysis (PCoA) and hierarchical clustering analysis (HCA) on the distance matrixes revealed that fungal communities from pools of a sampling site (that therefore share no common plant species) are more closely related than pools from different sites (Fig. 3), so that the composition difference is strongly correlated with the sampling site (ANOSIM test: R=0.89, p =1E-04) (Table S8 and Fig. 3). Secondly, factors such as ecosystem type (cultivated/unmanaged) and geolocation

(Bordeaux/Bergerac) also contribute to the composition dissimilarity with respectively R =0.52, *p* =4E-04 and R=0.48, *p* = 3E-04 (Table S8 and Fig. 3). The contributions of factors causing the compositional difference were also tested by ADONIS statistics, providing essentially similar results in particular for sampling site contribution (R^2 =0.37, *p* =1E-04), followed by ecosystem type and geolocation (R^2 >0.13, *p* < 5E-04) (Table S8).

Culturome and mycovirome diversity analysis following a culturomics approach

Several families of dsRNA viruses have members with either plant or fungal hosts, so that it is not easy to decide whether the agents detected are *bona-fide* plant-infecting viruses or infect fungi associated with the plant samples analyzed. The situation is further complicated by recent reports of cross-kingdom transmission (Andika *et al.*, 2017; Nerva *et al.*, 2017). In an effort to begin to address this complex question, we characterized the mycobiome and mycovirome of fungal populations that had been cultured from the plant populations sampled in our 4 study sites.

From 480 to 1270 fungal colonies were obtained for each sampling site through a culturomics approach (Table S9). Using the cultured fungal pools thus obtained, fungal metabarcoding and dsRNA-based virome analyses were then performed in the same fashion as for the plant pool samples. As expected and despite the relatively large number of cultivated colonies involved in these experiments, the cultivated fungal ASV output data shows that only a small fraction of the ASVs identified from the plant samples were identified among the cultivated fungi ASVs (4.8% to 13.8%, average 9.0% +/- 3.9%) (Table S9 and Fig. S2). Although a significant fraction of the cultivated fungi ASVs were not detected by the metabarcoding performed on the plant samples (15.4% to 42.9%) an even larger fraction had already been detected from the plant samples (57.1 to 84.6%, average 67.5% +/- 12.6%) (Table S9 and Fig. S2).

In order to maximize the ability to detect shared viruses between the plant- and fungus-associated viromes the non-normalized datasets were used. Remarkably, the viromes obtained from the cultivated fungal pools were almost completely different from the viromes obtained from the plant

pools. In total, based on the Blast annotation, the mycoviromes collectively comprised 14 viral families (7 ssRNA families and 7 dsRNA families, Table S10). *Totiviridae*, together with *Chrysoviridae*, *Endornaviridae*, *Alphaflexiviridae* and *Partitiviridae* were detected from all the mycoviromes and phytoviromes. On the other hand, a range of families were only detected from the cultured fungi, including the *Gammaflexiviridae*, *Hypoviridae*, *Tymoviridae*, *Narnaviridae*, *Fusarividae and Birnaviridae*. Also contrasting with the phytovirome data, the mycovirome of the INRA site proved not less diverse than at other sites with 8 families discovered (Table S10). At the more precise viral RdRp OTU level, although a large fraction of the fungal ASVs in cultured fungal pools are shared with the plant samples mycobiomes (average 67.5% +/- 12.6%, see above), the viromes from the two types of samples were found almost totally different, with only 2 OTUs shared for the CT site (out of a total of 29), while no shared OTU could be detected in the other 3 sites out of a total of 54 viral OTUs detected from the corresponding fungal cultures (Fig. S2). The reciprocal mapping of the reads of one virome type against the contigs of the other type confirmed that only a very minor fraction of agents is shared between the plant and fungal cultures derived viromes (data not shown).

DISCUSSION

While the sampled host plant species certainly affects the fungal community, a core mycobiome was shared, for a given site, between plant pools which gather different plant species. For each site, 31.1% to 42.9% of the mycobiome ASVs were present in at least two pools (Fig. S1), the corresponding core mycobiome representing a “signature” of the sampling site but also containing some widespread ASVs also represented in the mycobiome of other sampling sites. This core mycobiome, in particular site-specific ASVs, explains the clustering and PCoA analysis that group together different plant pools from a given site and unambiguously separate them from other sampling sites (Fig. 3). Theses variations are potentially associated with their mutual environments.

As reviewed in Vacher *et al.* (2016), environmental conditions have been recognized to significantly affect the assemblage of phyllosphere fungi such as elevation, landscape, climatic condition of a continent and across latitudes, and season.

The comparison of the aggregated mycobiomes from different sites showed that leaf-associated mycobiomes are consistently richer in unmanaged than in cultivated sampling sites (Fig. 2c, 2d). Even if cultivated plants were not sampled in the results reported here, they represent a high proportion of the plant biomass at the cultivated sampling sites and the lower diversity of their mycobiome (Compant *et al.*, 2019) may have impacted that of the weeds and wild plants growing nearby. Another hypothesis could be that fungicide treatments applied in cultivated sites may have reduced fungal diversity on the sampled plants. These two hypotheses are, by the way, not mutually exclusive.

Remarkably, a different picture emerged from phytovirome analysis in that more viral families were found from cultivated ecosystems than from unmanaged sampling sites (Table 1). This result parallels that of Bernardo *et al*. (2018) who also observed a higher family level virus diversity in cultivated areas. The results are less clear-cut when considering viral richness as estimated by the number of OTUs, which represent a proxy to viral species (Table 1). While diversity at the INRA unmanaged site was low and that at the VO cultivated site high, comparable and intermediate numbers of OTUs were observed in the other two CT and SP sites (CT cultivated, SP unmanaged). The finding of a lower viral richness for sites with a higher mycobiome diversity suggests that virome and mycobiome richness may not be influenced by the same drivers. Differences in dispersion mechanisms between fungi and viruses or the contrasted impact of fungicide treatments in mycobiomes and viromes are certainly among potential driver candidates. Domestication and cultivation, by reducing biodiversity have been suggested to be responsible for increased viral infections in cultivated ecosystems (Roossinck & García-Arenal, 2015). Such an effect may also

have contributed to the results reported here if spill-over of frequent infections in crops contributes a significant share of the virome of weeds/wild plants growing side by side with the crops.

A study of Thapa *et al.* (2015) has demonstrated for a few selected plant species in an unmanaged ecosystem that host species played a significant effect on virome composition as compared to location and sampling time. The results reported here show extremely high site specificity of the phytoviromes, with a high fraction of 93.2% of viral OTUs solely detected in one of the study sites, to be compared with the corresponding values of 74.7% and 55% respectively for the mycobiome ASVs and the sampled plant species (Fig. S3). Under our experimental conditions, the virome therefore appears to be more site-specific than either the mycobiome or the sampled plant populations. It should however be considered that this observation is likely only valid for viruses present at a high frequency in the sampled plant populations. Indeed, with only 5-7 individual plants sampled per plant species, the ability to detect viruses with a low, less than 10% prevalence in the sampled species, would have been limited. It is therefore possible that deeper sampling of each plant species, involving more numerous individual plants may provide in the future a different picture by allowing to take into account low prevalence viruses.

Fungal culturomics of plant leaves have made clear that *in vitro* culture-based approaches grossly underestimate fungal diversity (Roossinck, 2015a), and the results reported here are in line with this general observation. Indeed, only 4.8% to 13.8% of fungal ASVs were recovered here as fungal cultures (Table S10). However, it is noteworthy that the culturomics provided a significant fraction of cultivated fungi ASVs (15.4% to 42.9%, Fig. S2 and Table S9) or of viral OTUs that were not detected during the direct analysis of plant samples, highlighting the incompleteness of these efforts. The analyzed phytoviromes and mycoviromes, although derived from the same initial samples proved remarkably different. In particular a range of viral families were specifically detected from the mycoviromes: *Gammaflexiviridae* (all four sites), *Hypoviridae* and *Narnaviridae* (3 sites), *Fusariviridae* (2 sites), *Birnaviridae* and *Tymoviridae* (1 site) (Table S10).

Similarly there was almost no congruence between phytoviromes and mycoviromes at the OTU level (Fig. S2). These results are in contrast with some observations, in particular those reported by Al Rwahnih *et al.* (2011) in which a limited culturomics effort, involving only 11 fungal colonies, allowed to demonstrate a mycovirus status for 5 of the 25 (20%) viruses identified in a grapevine virome. While the culturomics effort reported here is 2 to 3 orders of magnitude higher, the proportion of matched OTUs is at least one order of magnitude lower. One possible hypothesis to explain these differences may be linked to the pooling strategy used here which, while allowing the analysis of many more individual samples, may favor the detection of highly prevalent or high concentration viruses. In this respect, further efforts are clearly needed to better understand the links between the mycovirome and the plant-associated virome.

The results presented here provide a large scale parallel analysis of the virome, mycovirome and mycobiome associated with complex plant populations in cultivated and unmanaged ecosystems. While the results obtained confirm a higher viral family richness in cultivated environments (Bernardo *et al.*, 2018), they suggest that mycobiome and virome might be under the influence of different drivers, an observation that clearly deserves further confirmatory efforts.

ACKNOWLEDGEMENTS

The help of Laurence Svanella-Dumas, Chantal Faure and Shuo Liu from the Virology team (UMR 1332 BFP, Villenave d'Ornon, France) for on-site plant sampling and fungal culturomics. and of Charlie Pauvert (UMR Biogeco, Pierroton, France) for fungal metabarcoding data analysis is gratefully acknowledged. The authors thank the Bordeaux Genome Transcriptome Facility (INRA, Pierroton, France) and the Genotoul platform (INRA, Toulouse, France) for MiSeq metabarcoding sequencing and Illumina sequencing respectively. YM was supported by a PhD grant from the China Scholarship Council. The authors also thank A. Raoult and F. Villeneuve (Centre Technique Interprofessionnel des Fruits et Légumes, Lanxade) and F. Dorlhac de Borne (Imperial Tobacco, Bergerac) for access to some sampling sites.

AUTHOR CONTRIBUTION

Thierry Candresse and Armelle Marais designed the experiments and supervised the progress, Corinne Vacher provided many useful suggestion for ITS sequencing. Yuxin Ma performed the molecular experiments, data analysis and interpretation. Tania Fort provided R scripts and contributed to the data processing of fungal ITS datasets. YM and TC wrote the manuscript, TF and CV with other authors provided critical reading of this manuscript and its further improvement. Marie Lefebvre and Sébastien Theil developed the virAnnot pipeline and performed the viral sequence processing and OTU annotation.

REFERENCES

Al Rwahnih M, Daubert S, Úrbez-Torres JR, Cordero F, Rowhani A. 2011. Deep sequencing evidence from single grapevine plants reveals a virome dominated by mycoviruses. *Archives of Virology* **156**(3): 397-403.

Anderson MJ. 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26**(1): 32-46.

Anderson PK, Cunningham AA, Patel NG, Morales FJ, Epstein PR, Daszak P. 2004. Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. *Trends in Ecology & Evolution* **19**(10): 535-544.

Andika IB, Wei S, Cao C, Salaipeth L, Kondo H, Sun L. 2017. Phytopathogenic fungus hosts a plant virus: A naturally occurring cross-kingdom viral infection. *Proceedings of the National Academy of Sciences of the United States of America* **114**(46): 12267-12272.

Arnold AE, Mejía LC, Kyllo D, Rojas EI, Maynard Z, Robbins N, Herre EA. 2003. Fungal endophytes limit pathogen damage in a tropical tree. *Proceedings of the National Academy of Sciences* **100**(26): 15649-15654.

Bernardo P, Charles-Dominique T, Barakat M, Ortet P, Fernandez E, Filloux D, Hartnady P, Rebelo TA, Cousins SR, Mesleard F, et al. 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.

Bringel F, Couée I. 2015. Pivotal roles of phyllosphere microorganisms at the interface between plant functioning and atmospheric trace gas dynamics. *Frontiers in Microbiology* **6**: 486-486.

Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S, Cruaud C, de Vargas C, Gasol JM, et al. 2015. Patterns and ecological drivers of ocean viral communities. *Science* **348**(6237): 1261498.

Cadle-Davidson L, Bergstrom G. 2004. The effects of postplanting environment on the incidence of soilborne viral diseases in winter cereals. *Phytopathology* **94**(5): 527-534.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016a. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* **13**: 581.

Callahan BJ, Sankaran K, Fukuyama J, McMurdie P, Holmes S. 2016b. Bioconductor workflow for microbiome data analysis: from raw reads to community analyses [version 1; peer review: 3 approved]. *F1000Research* **5**(1492).

Clarke KR. 1993. Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology* **18**(1): 117-143.

Compant S, Samad A, Faist H, Sessitsch A. 2019. A review on the plant microbiome: Ecology, functions, and emerging trends in microbial application. *Journal of Advanced Research*.

Cooper I, Jones RAC 2006. Wild Plants and Viruses: Under‐Investigated Ecosystems. *Advances in Virus Research*: Academic Press, 1-47.

Cordier T, Robin C, Capdevielle X, Fabreguettes O, Desprez‐Loustau ML, Vacher C. 2012. The composition of phyllosphere fungal assemblages of E uropean beech (F agus sylvatica) varies significantly along an elevation gradient. *New Phytologist* **196**(2): 510-519.

Coutts B, Jones R. 2002. Temporal dynamics of spread of four viruses within mixed species perennial pastures. *Annals of Applied Biology* **140**(1): 37-52.

Doty SL 2017. Endophytic N-Fixation: Controversy and a Path Forward. *Functional Importance of the Plant Microbiome*: Springer, 7-20.

Elena SF, Fraile A, García-Arenal F 2014. Evolution and emergence of plant viruses. *Advances in Virus Research*: Elsevier, 161-191.

Fakhro A, Andrade-Linares DR, von Bargen S, Bandte M, Büttner C, Grosch R, Schwarz D, Franken P. 2010. Impact of Piriformospora indica on tomato growth and on interaction with fungal and viral pathogens. *Mycorrhiza* **20**(3): 191-200.

Fürnkranz M, Wanek W, Richter A, Abell G, Rasche F, Sessitsch A. 2008. Nitrogen fixation by phyllosphere bacteria associated with higher plants and their colonizing epiphytes of a tropical lowland rainforest of Costa Rica. *The ISME journal* **2**(5): 561.

Galan M, Razzauti M, Bard E, Bernard M, Brouat C, Charbonnel N, Dehne-Garcia A, Loiseau A, Tatard C, Tamisier L. 2016. 16S rRNA amplicon sequencing for epidemiological surveys of bacteria in wildlife. *MSystems* **1**(4): e00032-00016.

Ghabrial SA, Caston JR, Jiang D, Nibert ML, Suzuki N. 2015. 50-plus years of fungal viruses. *Virology* **479-480**: 356-368.

Hacquard S, Spaepen S, Garrido-Oter R, Schulze-Lefert P. 2017. Interplay Between Innate Immunity and the Plant Microbiota. *Annual Review of Phytopathology* **55**: 565-589.

Hamers L, Hemeryck Y, Herweyers G, Janssen M, Keters H, Rousseau R, Vanhoutte A. 1989. Similarity measures in scientometric research: the Jaccard index versus Salton's cosine formula. *Inf. Process. Manage.* **25**(3): 315-318.

Huerta-Cepas J, Bork P, Serra F. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution* **33**(6): 1635-1638.

Illumina. 2017. Effects of index misassignment on multiplexing and downstream analysis.

Jumpponen A, Jones K. 2010. Seasonally dynamic fungal communities in the Quercus macrocarpa phyllosphere differ between urban and nonurban environments. *New Phytologist* **186**(2): 496-513.

Jumpponen A, Jones KL. 2009. Massively parallel 454 sequencing indicates hyperdiverse fungal communities in temperate Quercus macrocarpa phyllosphere. *New Phytologist* **184**(2): 438-448.

Koskella B. 2013. Phage-mediated selection on microbiota of a long-lived host. *Current Biology* **23**(13): 1256-1260.

Laforest-Lapointe I, Paquette A, Messier C, Kembel SW. 2017. Leaf bacterial diversity mediates plant diversity and ecosystem function relationships. *Nature* **546**(7656): 145-147.

Laforest-Lapointe I, Whitaker BK. 2019. Decrypting the phyllosphere microbiota: progress and challenges. *American Journal of Botany* **106**(2): 171-173.

Lindow SE, Brandl MT. 2003. Microbiology of the phyllosphere. *Applied and Environmental Microbiology* **69**(4): 1875-1883.

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

Marais A, Faure C, Bergey B, Candresse T. 2018. Viral Double-Stranded RNAs (dsRNAs) from Plants: Alternative Nucleic Acid Substrates for High-Throughput Sequencing. *Methods in Molecular Biology* **1746**: 45-53.

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

McArdle BH, Anderson MJ. 2001. Fitting Multivariate Models to Community Data: A Comment on Distance-Based Redundancy Analysis. *Ecology* **82**(1): 290-297.

McLeish MJ, Fraile A, García-Arenal F. 2019. Evolution of plant–virus interactions: host range and virus emergence. *Current Opinion in Virology* **34**: 50-55.

McMurdie PJ, Holmes S. 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PloS One* **8**(4): e61217.

Morris CE. 2001. Phyllosphere. *e LS.*

Morris CE, Conen F, Alex Huffman J, Phillips V, Pöschl U, Sands DC. 2014. Bioprecipitation: a feedback cycle linking Earth history, ecosystem dynamics and land use through biological ice nucleators in the atmosphere. *Global Change Biology* **20**(2): 341-351.

Moyes AB, Kueppers LM, Pett‐Ridge J, Carper DL, Vandehey N, O'Neil J, Frank AC. 2016. Evidence for foliar endophytic nitrogen fixation in a widely distributed subalpine conifer. *New Phytologist* **210**(2): 657-668.

Nerva L, Varese GC, Falk BW, Turina M. 2017. Mycoviruses of an endophytic fungus can replicate in plant cells: evolutionary implications. *Scientific Reports* **7**(1): 1908.

Nuss D. 2008. Hypoviruses.

Op De Beeck M, Lievens B, Busschaert P, Declerck S, Vangronsveld J, Colpaert JV. 2014. Comparison and validation of some ITS primer pairs useful for fungal metabarcoding studies. *PloS One* **9**(6): e97629.

Osono T. 2006. Role of phyllosphere fungi of forest trees in the development of decomposer fungal communities and decomposition processes of leaf litter. *Canadian Journal of Microbiology* **52**(8): 701-716.

Penuelas J, Rico L, Ogaya R, Jump A, Terradas J. 2012. Summer season and long‐term drought increase the richness of bacteria and fungi in the foliar phyllosphere of Quercus ilex in a mixed Mediterranean forest. *Plant Biology* **14**(4): 565-575.

Ramette A. 2007. Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology* **62**(2): 142-160.

Roossinck MJ. 2005. Symbiosis versus competition in plant virus evolution. *Nature Reviews: Microbiology* **3**(12): 917-924.

Roossinck MJ. 2012. Plant Virus Metagenomics: Biodiversity and Ecology. *Annual Review of Genetics* **46**: 359-369.

Roossinck MJ. 2015a. Metagenomics of plant and fungal viruses reveals an abundance of persistent lifestyles. *Frontiers in Microbiology* **5**.

Roossinck MJ. 2015b. Plants, viruses and the environment: ecology and mutualism. *Virology* **479**: 271-277.

Roossinck MJ, García-Arenal F. 2015. Ecosystem simplification, biodiversity loss and plant virus emergence. *Current Opinion in Virology* **10**: 56-62.

Roossinck MJ, Saha P, Wiley GB, Quan J, White JD, Lai H, Chavarria F, Shen GA, Roe BA. 2010. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. *Molecular Ecology* **19**: 81-88.

Rosado BH, Almeida LC, Alves LF, Lambais MR, Oliveira RS. 2018. The importance of phyllosphere on plant functional ecology: a phyllo trait manifesto. *New Phytologist* **219**(4): 1145-1149.

Rosario K, Breitbart M. 2011. Exploring the viral world through metagenomics. *Current Opinion in Virology* **1**(4): 289-297.

Saleem M, Meckes N, Pervaiz ZH, Traw MB. 2017. Microbial Interactions in the Phyllosphere Increase Plant Performance under Herbivore Biotic Stress. *Frontiers in Microbiology* **8**(41).

Sapkota R, Jørgensen LN, Nicolaisen M. 2017. Spatiotemporal Variation and Networks in the Mycobiome of the Wheat Canopy. *Frontiers in plant science* **8**: 1357-1357.

Sapkota R, Knorr K, Jørgensen LN, O'Hanlon KA, Nicolaisen M. 2015. Host genotype is an important determinant of the cereal phyllosphere mycobiome. *New Phytologist* **207**(4): 1134-1144.

Schwarz D, Beuch U, Bandte M, Fakhro A, Büttner C, Obermeier C. 2010. Spread and interaction of Pepino mosaic virus (PepMV) and Pythium aphanidermatum in a closed nutrient solution recirculation system: effects on tomato growth and yield. *Plant Pathology* **59**(3): 443-452.

Seabloom EW, Borer ET, Mitchell CE, Power AG. 2010. Viral diversity and prevalence gradients in North American Pacific Coast grasslands. *Ecology* **91**(3): 721-732.

Shaul O, Galili S, Volpin H, Ginzberg I, Elad Y, Chet I, Kapulnik Y. 1999. Mycorrhiza-induced changes in disease severity and PR protein expression in tobacco leaves. *Molecular Plant-Microbe Interactions* **12**(11): 1000-1007.

Shi M, Lin X-D, Tian J-H, Chen L-J, Chen X, Li C-X, Qin X-C, Li J, Cao J-P, Eden J-S, et al. 2016. Redefining the invertebrate RNA virosphere. *Nature* **540**: 539.

Siddell SG, Walker PJ, Lefkowitz EJ, Mushegian AR, Adams MJ, Dutilh BE, Gorbalenya AE, Harrach B, Harrison RL, Junglen S, et al. 2019. Additional changes to taxonomy ratified in a special vote by the International Committee on Taxonomy of Viruses (October 2018). *Archives of Virology* **164**(3): 943-946.

Simmonds P. 2015. Methods for virus classification and the challenge of incorporating metagenomic sequence data. *Journal of General Virology* **96**(Pt 6): 1193-1206.

Spence N, Basham J, Mumford R, Hayman G, Edmondson R, Jones D. 2006. Effect of Pepino mosaic virus on the yield and quality of glasshouse‐grown tomatoes in the UK. *Plant Pathology* **55**(5): 595-606.

Stobbe AH, Roossinck MJ. 2014. Plant virus metagenomics: what we know and why we need to know more. *Frontiers in plant science* **5**(150).

Susi H, Filloux D, Frilander MJ, Roumagnac P, Laine AL. 2019. Diverse and variable virus communities in wild plant populations revealed by metagenomic tools. *PeerJ* **7**: e6140.

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

UNITE C. 2019. UNITE general FASTA release for Fungi. Version 18.11.2018. . *UNITE Community.* .

Unterseher M, Schnittler M. 2009. Dilution-to-extinction cultivation of leaf-inhabiting endophytic fungi in beech (Fagus sylvatica L.)--different cultivation techniques influence fungal biodiversity assessment. *Mycological Research* **113**(5): 645-654.

Vacher C, Hampe A, Porté AJ, Sauer U, Compant S, Morris CE. 2016. The Phyllosphere: Microbial Jungle at the Plant–Climate Interface. *Annual Review of Ecology, Evolution, and Systematics* **47**(1): 1-24.

van der Valk T, Vezzi F, Ormestad M, Dalén L, Guschanski K. 2019. Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. *Molecular Ecology Resources*.

Vorholt JA. 2012. Microbial life in the phyllosphere. *Nature Reviews Microbiology* **10**: 828.

Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* **73**(16): 5261-5267.

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*: Springer-Verlag New York.

Wren JD, Roossinck MJ, Nelson RS, Scheets K, Palmer MW, Melcher U. 2006. Plant Virus Biodiversity and Ecology. *PLoS Biology* **4**(3): e80.

Table 1. Main characteristics of the plant-associated viromes from different sampling sites

Site	**VO**	**CT**	**SP**	**INRA**
Coordinates	***44°46'50.9"N 0°33'24.6"W***	***44°50'30.4"N 0°25'34.8"E***	***44°51'56.7"N 0°17'36.1"E***	***44°47'19.7"N 0°34'32.7"W***
Ecosystem_type	***Cultivated***	***Cultivated***	***Unmanaged***	***Unmanaged***
Sampling_dates	***16-may-2017***	***11-july-2017***	***11-july-2017***	***5-apr-2017***
Plant_species	*40*	*33*	*29*	*34*
Pooled_libraries names	*lib1-lib9*	*lib10-lib18*	*lib21-lib22*	*lib19-lib20*
Normalization				
Total reads (pairs)	*2739314*	*9205641*	*976948*	***634432***
Subsampling_depth (pairs)	634432	634432	634432	634432
Blast annotation				
Reads in viral contigs	289019	651841	234088	106501
% reads in viral contigs	22.8%	51.4%	18.4%	8.4%
dsRNA / ssRNA viruses reads	1.6	1.1	19.6	0.7
Viral families	15	14	11	7
RdRp-OTU classification				
Total no. of OTUs	73	50	55	26
Viral families	13	12	9	5
Percent OTUs with Genbank RdRp identity >=90%	10.0%	16.3%	3.7%	12.0%
Percent OTUs with Genbank RdRp identity <90%	90.0%	83.7%	96.3%	88.0%

LEGENDS TO THE FIGURES

Figure 1. Schematic representation of the sample processing and sequencing strategies (metagenomics and culturomics) for viral communities from plants samples and fungal cultures (phytovirome and mycovirome) and for fungal communities from plant samples and fungal cultures (mycobiome and culturome) analyses.

Figure 2. RdRp Operational Taxonomy Units (OTUs) virome composition and known/novel status of RdRp OTUs at each sampling site. (**A**) Virome composition based on family level OTUs annotation. (**B**) Pie charts showing the proportion of ssRNA, dsRNA and unclassified OTUs. Colors separate in each group the known viral OTUs (in green) for which a RdRp with >=90% identity was identified in Genbank and the potentially novel viral OTUs (in grey). Box plots illustrating fungal community richness and diversity in plant pools from cultivated or unmanaged sampling sites reflected by (**C**) the number of detected amplicon sequence variants (ASVs) and (**D**) Shannon diversity index calculated using read numbers as a proxy to individual ASV prevalence.

Figure 3. Principal coordinates analysis (PCoA) and hierarchical clustering analysis of mycobiome compositions for independent plant pools coming from the same or from different sampling sites. Plant pools from the same site do not contain shared plant species. (A) PCoA (A) and dendrogram (B) calculated using the Jaccard-Binary distance based on presence/absence of amplicon sequence variants (ASVs) for each library. Different shapes indicates the plant pools from a specific sampling sites (CT, INRA, SP, VO). The shapes are colored according to the sampling site status (cultivated or unmanaged).

Figure 1

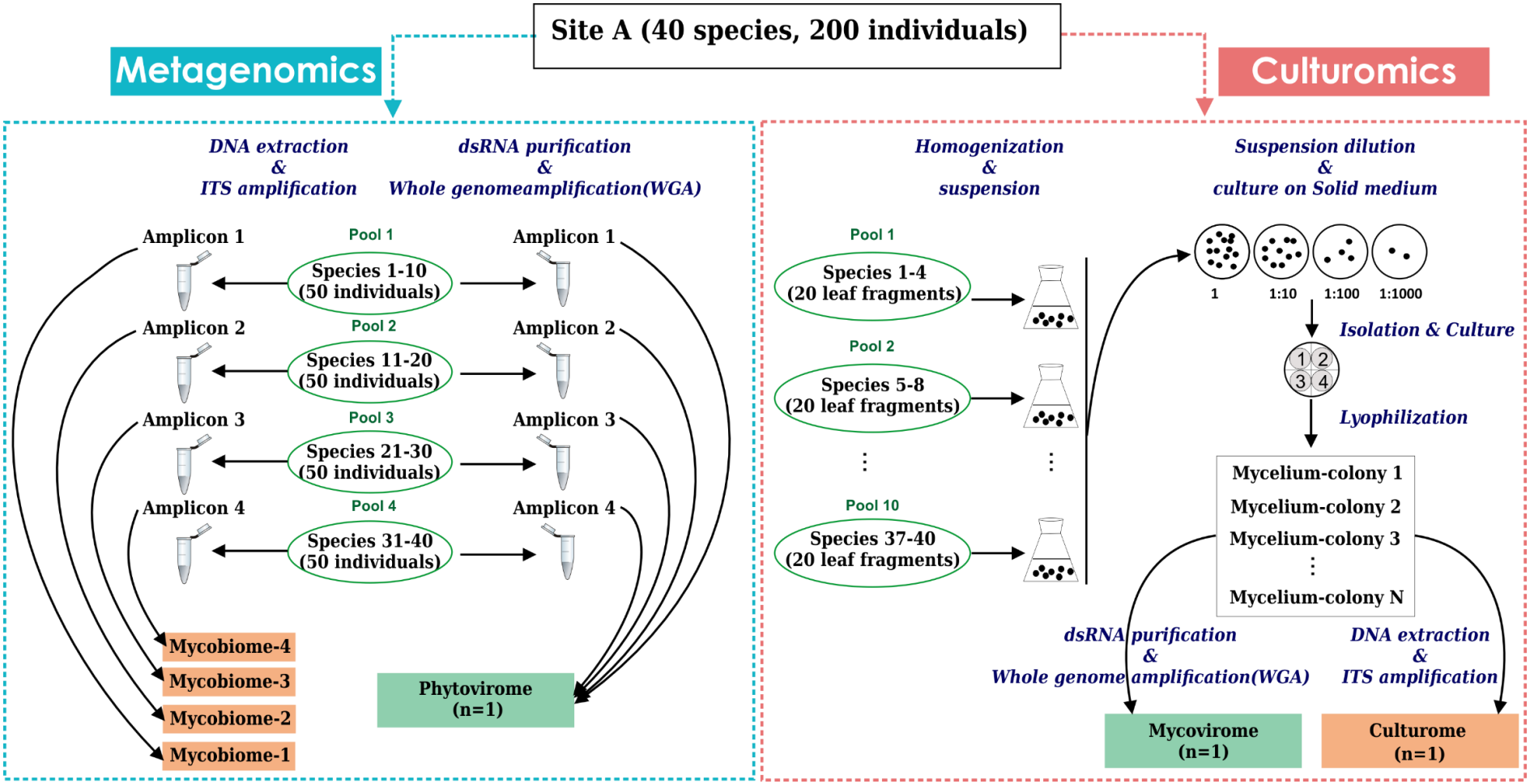

Site A (40 species, 200 individuals)
Metagenomics
Culturomics
DNA extraction & ITS amplification
dsRNA purification & Whole genomeamplification(WGA)
Amplicon 1
Amplicon 2
Amplicon 3
Amplicon 4
Pool 1
Species 1-10 (50 individuals)
Pool 2
Species 11-20 (50 individuals)
Pool 3
Species 21-30 (50 individuals)
Pool 4
Species 31-40 (50 individuals)
Mycobiome-4
Mycobiome-3
Mycobiome-2
Mycobiome-1
Phytovirome (n=1)
Homogenization & suspension
Suspension dilution & culture on Solid medium
Pool 1
Species 1-4 (20 leaf fragments)
Pool 2
Species 5-8 (20 leaf fragments)
Pool 10
Species 37-40 (20 leaf fragments)
1
1:10
1:100
1:1000
Isolation & Culture
1 2 3 4
Lyophilization
Mycelium-colony 1
Mycelium-colony 2
Mycelium-colony 3
Mycelium-colony N
dsRNA purification & Whole genome amplification(WGA)
DNA extraction & ITS amplification
Mycovirome (n=1)
Culturome (n=1)

Figure 2

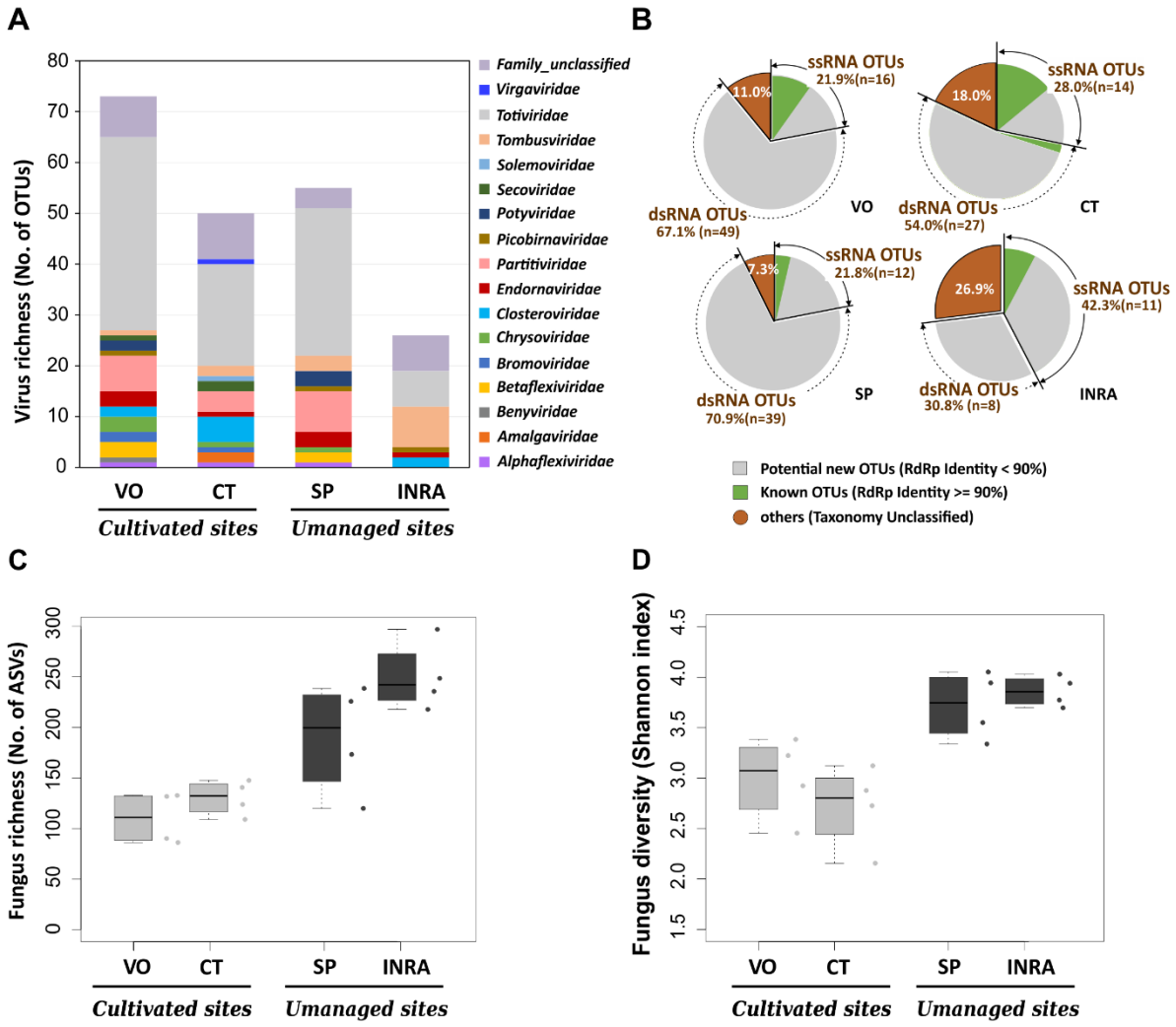
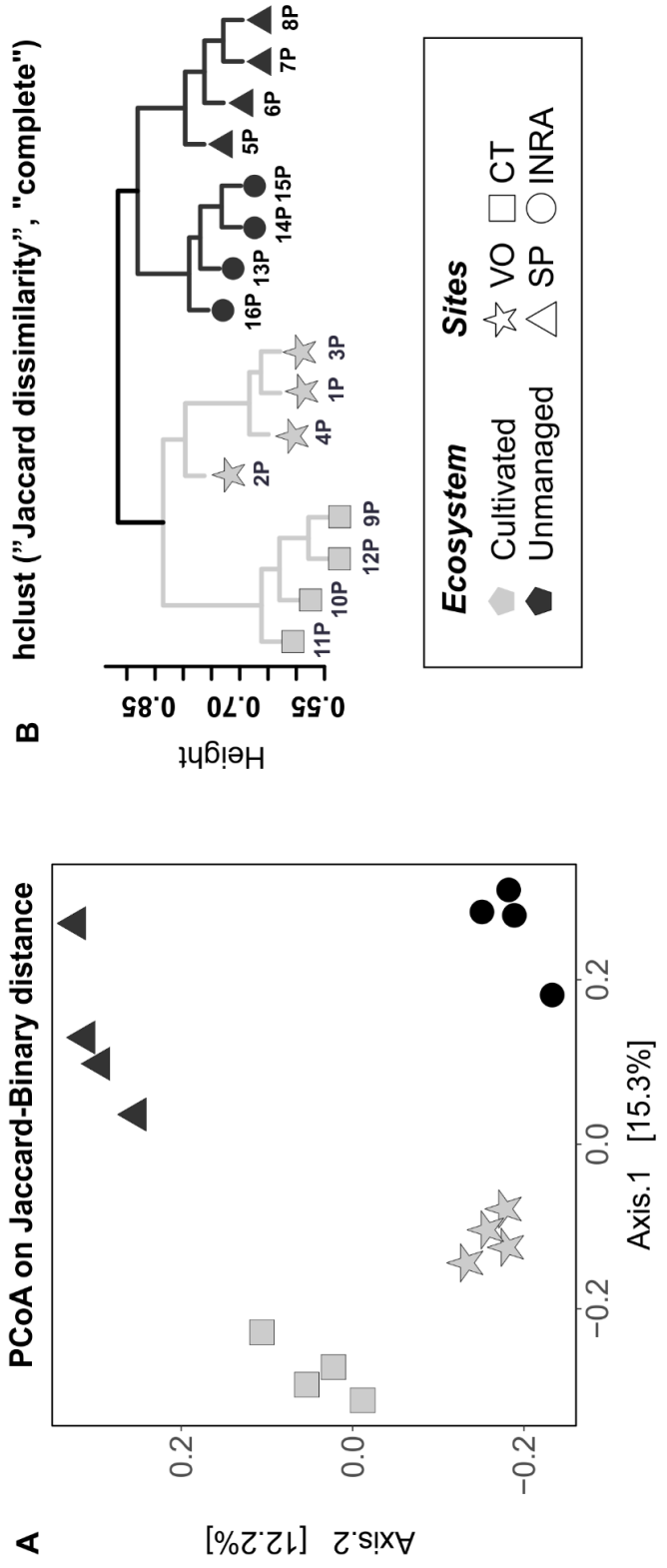

A
Virus richness (No. of OTUs)
Family_unclassified
Virgaviridae
Totiviridae
Tombusviridae
Solemoviridae
Secoviridae
Potyviridae
Picobirnaviridae
Partitiviridae
Endornaviridae
Closteroviridae
Chrysoviridae
Bromoviridae
Betaflexiviridae
Benyviridae
Amalgaviridae
Alphaflexiviridae
VO
CT
SP
INRA
Cultivated sites
Umanaged sites
B
ssRNA OTUs 21.9%(n=16)
11.0%
dsRNA OTUs 67.1% (n=49)
VO
ssRNA OTUs 28.0%(n=14)
18.0%
dsRNA OTUs 54.0%(n=27)
CT
ssRNA OTUs 21.8%(n=12)
7.3%
dsRNA OTUs 70.9%(n=39)
SP
ssRNA OTUs 42.3%(n=11)
26.9%
dsRNA OTUs 30.8% (n=8)
INRA
Potential new OTUs (RdRp Identity < 90%)
Known OTUs (RdRp Identity >= 90%)
others (Taxonomy Unclassified)
C
Fungus richness (No. of ASVs)
VO
CT
SP
INRA
Cultivated sites
Umanaged sites
D
Fungus diversity (Shannon index)
VO
CT
SP
INRA
Cultivated sites
Umanaged sites

Figure 3


A
PCoA on Jaccard-Binary distance
Axis.2 [12.2%]
Axis.1 [15.3%]
0.2
0.0
−0.2
B
hclust ("Jaccard dissimilarity", "complete")
Height
0.55
0.70
0.85
11P
10P
12P
9P
2P
4P
1P
3P
16P
13P
14P
15P
5P
6P
7P
8P
Ecosystem
Cultivated
Unmanaged
Sites
VO
SP
CT
INRA

Supplementary Materials [All supplementary Tables are available at https://doi.org/10.15454/UL0OLW, those that fit easily in A4 format are also provided here]

Table S1 Individual plant samples and species for phytovirome and mycobiome analysis.

Table S2 Phytoviromes and mycoviromes from different sites with information on HTS output.

Table S3 Number of reads integrated in contigs belonging to different viral families as determined by Blast-based annotation.

Table S4 Viral RdRp OTUs identified from the VO, CT, SP and INRA sampling sites. The number of reads integrated in each OTUs is shown for each site, together with the BastX first Genbank hit information for each OTU. OTUs corresponding to known viruses are highlighted in green.

Table S5 Identified amplicon sequence variants (ASVs) with their taxonomic assignation and the corresponding number of reads identified in each library. The well represented ASVs in all the libraries are highlighted in grey.

Table S6 Relative abundance of fungal phyla and classes in each library.

Table S7 Number of ITS1 ASVs detected and diversity indices.

Table S8 ANOSIM and ADONIS tests for potential factors affecting mycobiome composition.

Table S9 Comparison of fungal richness reflected by the statistics of fungal ASVs calculated using non-normalized data between plant mycobiomes and fungal culturomes.

Table S10 Viral families detected by Blast annotation of contigs from plant or cultured fungal samples for the four different study sites.

Table S2. Fungal pools used for dsRNA or DNA extraction for mycovirome and culturomics studies.

Library*	Normalization	Paired-end reads	Percentage of reads in eucaryota contigs	Percentage of reads in bacteria contigs	Percentage of reads in virus contigs	Percentage of reads in unknown contigs
Phytoviromes$						
VO	yes	*634432*	55.0%	0.3%	25.7%	18.9%
CT	yes	*634432*	21.8%	0.6%	56.1%	21.6%
SP	yes	*634432*	38.7%	0.0%	44.5%	16.8%
INRA	yes	*634432*	15.0%	0.6%	59.7%	24.7%
Negative control Phytovirome	no	*154440*	27.6%	59.7%	4.7%	8.0%
Mycoviromes						
VO	no	*307324*	0.8%	0.4%	33.3%	65.5%
CT	no	*533238*	2.0%	0.5%	52.4%	45.1%
SP	no	*702596*	2.5%	0.0%	61.0%	36.4%
INRA	no	*501986*	14.9%	2.5%	13.4%	69.1%
Negative control Mycovirome	no	*1641114*	29.8%	58.4%	2.9%	8.8%

*** For all whole genome amplifications (WGA), a minimum of two independent amplifications with different MID tags were performed for each site.**

$ For Phytovirome analysis, in order to allow a more precise comparison of the virome between sites, all the sequences with different MID tags but from the same sample were pooled and then normalized to a given depth.

Table S3. Number of reads integrated in contigs belonging to different viral families as determined by Blast-based annotation.

Site	**VO**	**CT**	**SP**	**INRA**
Ecosystem_type	***Cultivated***	***Cultivated***	***Unmanaged***	***Unmanaged***
Plant_species	*40*	*33*	*29*	*34*
Normalization depth	634432 pairs	634432 pairs	634432 pairs	634432 pairs
unclassified	2888	67948	2912	5349
Alphaflexiviridae	2511	113495	1902	45508
Amalgaviridae	0	0	875	0
Benyviridae	156	10	0	0
Betaflexiviridae	1238	0	2285	0
Bromoviridae	1026	3984	0	0
Caulimoviridae	52	16	0	0
Chrysoviridae	4740	68821	2282	0
Closteroviridae	65558	26484	0	2379
Endornaviridae	65906	201454	191769	36746
Luteoviridae	0	20	217	91
Partitiviridae	92761	58563	21813	7601
Potyviridae	35499	0	1048	0
Reoviridae	41	1244	135	0
Secoviridae	4892	54924	0	0
Tombusviridae	135	7840	5754	8420
Totiviridae	11577	4450	3096	407
Virgaviridae	39	42588	0	0
Total viral reads	**289019**	**651841**	**234088**	**106501**

Table S7. Number of ITS1 ASVs detected and diversity indices.

Site	Sample	ASV	Shannon	Simpson	Evenness
VO	VO-myco-P1	91	2.92	0.90	0.65
	VO-myco-P2	134	3.22	0.92	0.66
	VO-myco-P3	133	3.38	0.94	0.69
	VO-myco-P4	87	2.45	0.83	0.55
	Total unique ASVs	***268***	*na*	*na*	*na*
	Mean	***111.3***	***3.00***	***0.90***	***0.64***
	SD	***25.7***	***0.41***	***0.05***	***0.06***
SP	SP-myco-P1	241	3.94	0.96	0.72
	SP-myco-P2	121	3.34	0.93	0.70
	SP-myco-P3	228	4.05	0.97	0.75
	SP-myco-P4	175	3.55	0.93	0.69
	Total unique ASVs	***483***	*na*	*na*	*na*
	Mean	***191.3***	***3.72***	***0.95***	***0.71***
	SD	***54.8***	***0.33***	***0.02***	***0.03***
CT	CT-myco-P1	125	2.73	0.80	0.56
	CT-myco-P2	110	2.16	0.69	0.46
	CT-myco-P3	149	2.88	0.83	0.58
	CT-myco-P4	142	3.12	0.90	0.63
	Total unique ASVs	***275***	*na*	*na*	*na*
	Mean	***131.5***	***2.72***	***0.81***	***0.56***
	SD	***17.5***	***0.41***	***0.09***	***0.07***
INRA	INRA-myco-P1	238	3.94	0.96	0.72
	INRA-myco-P2	220	3.70	0.93	0.69
	INRA-myco-P3	251	3.77	0.93	0.68
	INRA-myco-P4	300	4.03	0.95	0.71
	Total unique ASVs	***639***	*na*	*na*	*na*
	Mean	***252.3***	***3.86***	***0.94***	***0.70***
	SD	***34.3***	***0.15***	***0.01***	***0.02***

Table S8. ANOSIM and ADONIS test for potential factors affecting mycobiome composition.

Factors	Jaccard-Binary - ANOSIM		Jaccard-Binary - ADONIS	
	R	*P* - value	R^2	*P* - value
Sampling_site	0.8859	1.00E-04	0.37039	1.00E-04
Ecosystem_type	0.5232	4.00E-04	0.14294	4.00E-04
Location	0.4821	3.00E-04	0.13202	5.00E-04

Table S9. Comparison of fungal richness reflected by the statistics of fungal ASVs calculated using non-normalized data between plant mycobiomes and fungal culturomes.

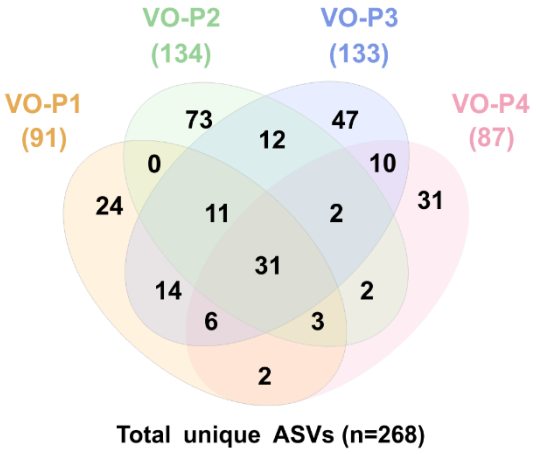
	VO	**CT**	**SP**	**INRA**		
Geo_Location	Bordeaux	Bergerac	Bergerac	Bordeaux		
Ecosystem_type	Cultivated	Cultivated	Unmanaged	Unmanaged	**Mean**	**SD**
Plant species	40	33	29	34		
Fungus colonies	1279	480	590	1060		
Fungal ASVs in culturome **(F)**	75	63	52	68	***64.5***	***9.7***
Fungal ASVs in plant mycobiome **(P)**	378	348	633	838	***549.3***	***231.1***
Fungal ASVs detected from both culturome and mycobiome **(S)**	52	36	44	40	***43.0***	***6.8***
Percentage of plant mycobiome ASVs shared with culturome **(S/P)**	13.8%	10.3%	7.0%	4.8%	***9.0%***	***3.9%***
Percentage of culturome ASVs not shared with plant mycobiome ASVs (**(F-S)/F)**	30.7%	42.9%	15.4%	41.2%	***32.5%***	***12.6%***
Percentage of culturome ASVs shared with plant mycobiome ASVs **(S/F)**	69.3%	57.1%	84.6%	58.8%	***67.5%***	***12.6%***

Supplementary Figure S1. Venn diagrams showing the shared fungal ASVs between different plant pools from the same sampling site. (A) VO site, **(B)** SP site**, (C)** CT site and **(D)** INRA site. The total number of unique ASVs in each site is indicated.

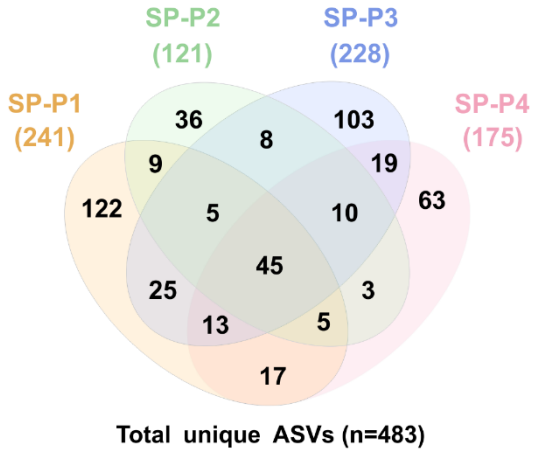
Supplementary Figure S2. Scaled Venn diagrams demonstrating the shared fungal ASVs/shared viral OTUs between the mycobiomes/viromes of plants pools and of cultured fungi pools obtained from the plant pools from different sampling sites.

Supplementary Figure S3. Bar plot showing the frequency of sampled plant species, detected fungal ASVs and detected viral OTUs between the four study sites.

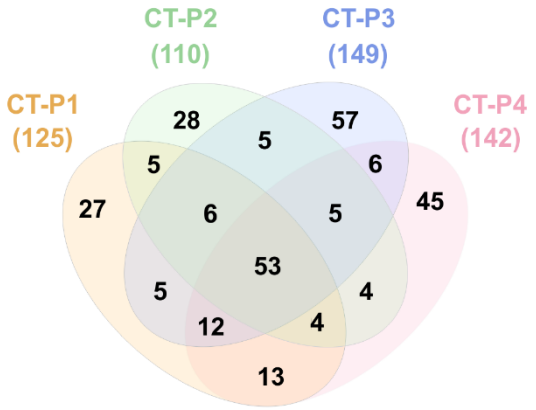
Supplementary Figure S1



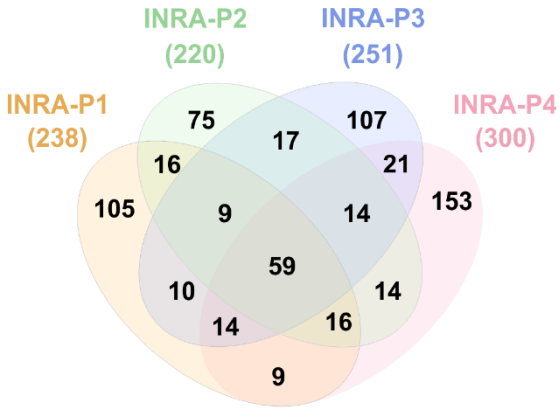
Total unique ASVs (n=268)



Total unique ASVs (n=483)

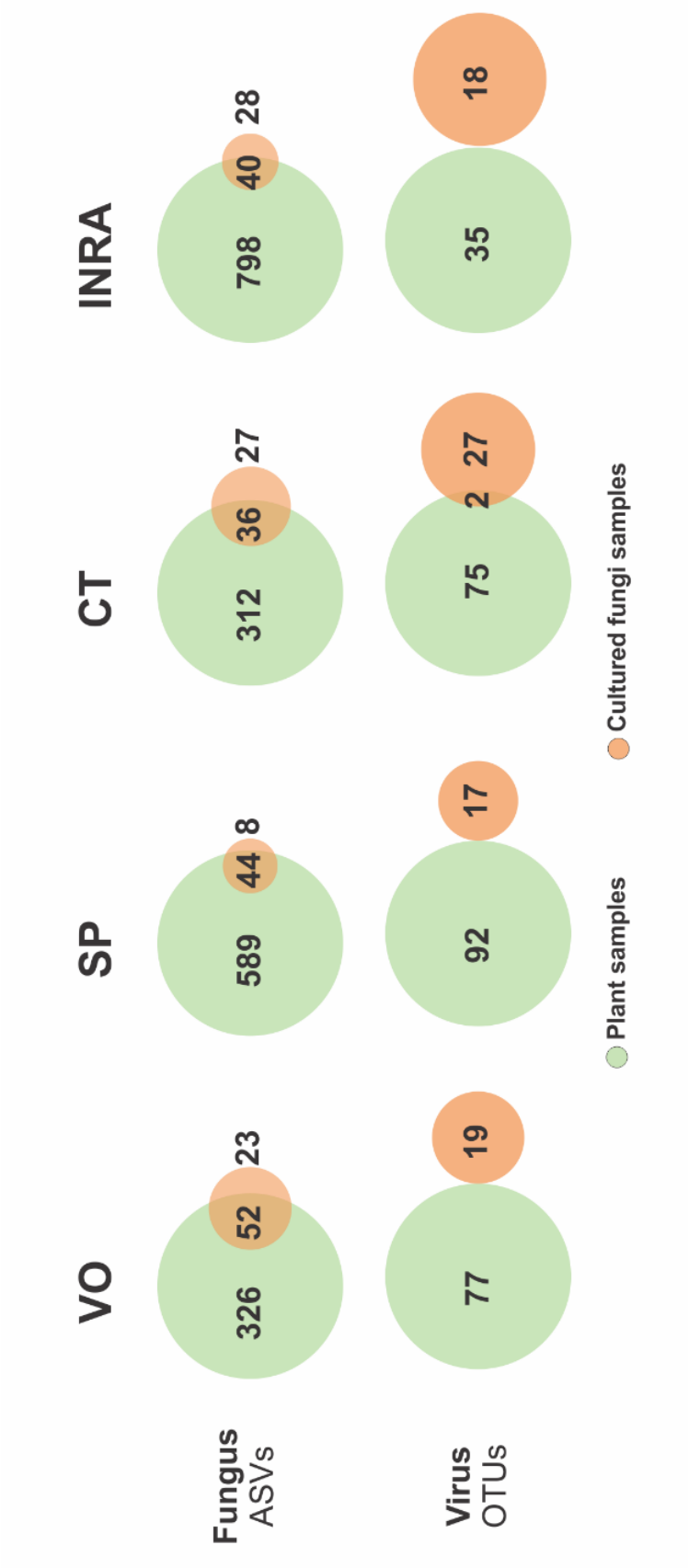


Total unique ASVs (n=275)

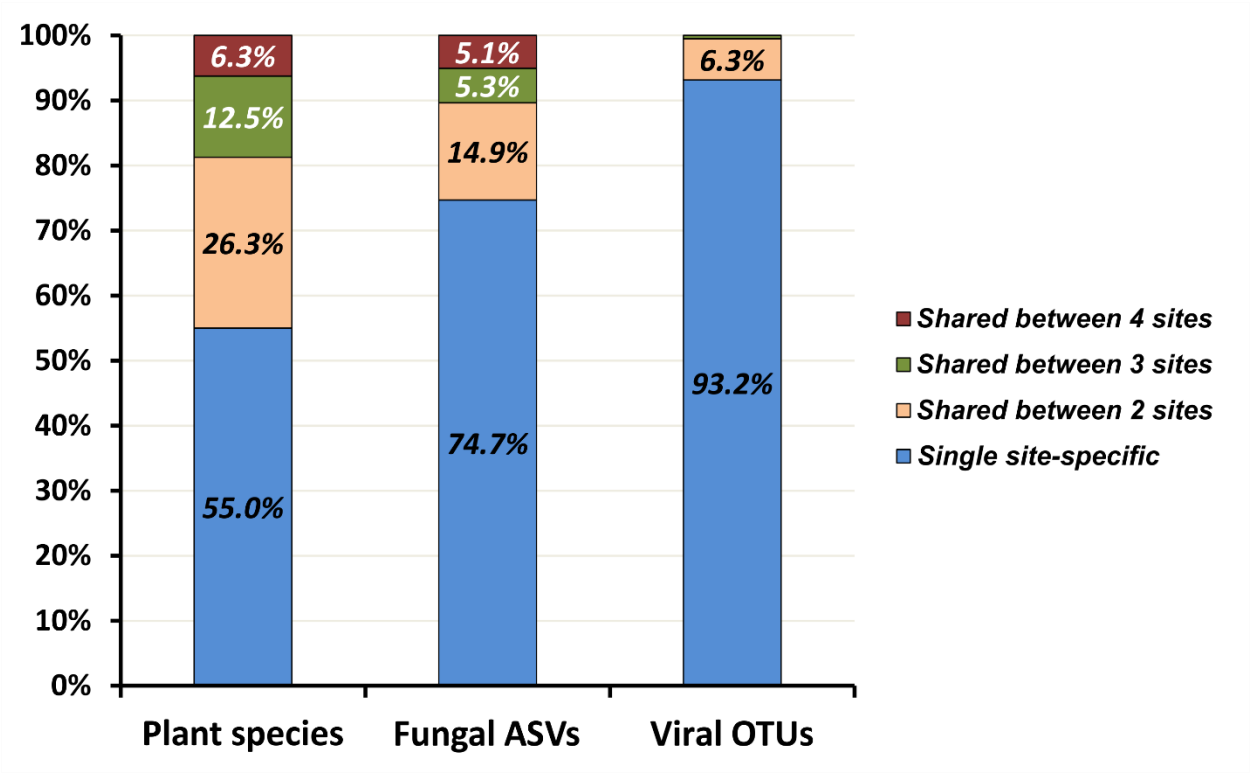


Total unique ASVs (n=639)

Supplementary Figure S2



Supplementary Figure S3



ANNEXES to CHAPTER III

Annex A to Chapter III – "Reproducibility of phytovirome composition analysis using random whole genome amplification"

HTS-based metagenomics targeting dsRNA nucleic acids or VANA has now been used in several studies (Bernardo *et al.*, 2018, Marais *et al.*, 2018, Roossinck *et al.*, 2010, Thapa *et al.*, 2015). However, there have been few efforts to date to evaluate the reproducibility of phytovirome description using whole genome amplification (WGA) of purified templates obtained from complex plant pools such as those coming from the so-called "lawn-mower" sampling approach (Roossinck *et al.*, 2015). Using the purified dsRNA samples obtained from two of the sampling sites of the study described in the manuscript of Chapter III (VO and CT), we evaluated the reproducibility of the whole genome amplification as well as the potential existence of an amplification bias linked to the use of particular MID tag identifiers used.

For each sample, triplicate amplifications involving 3 different MID tag were performed, generating a total of 9 libraries per sample (Table A1). Following HTS, the sequence reads were cleaned, normalized by subsampling to the same sequencing depth, assembled into contigs and submitted to an RdRp OTU clustering approach using the virAnnot pipeline and a 10% nucleotide/amino acid divergence cut-off (Lefebvre et al., submitted for publication).

On average, 12.1+/-0.9 viral families and 40.3+/-6.3 viral RdRp OTUs were detected for each of the libraries for the VO sampling site and 10.6+/-1.1 viral families with 33.9+/-3.2 OTUs for the libraries of the CT site (Table A1). ANOVA tests were used to compare for each site the libraries with different MID tags on several parameters, including viral reads counts/proportion, number of viral families and number of viral OTUs. No statistically significant differences were found between MID groups, suggesting that the choice of MID tag does not introduce significant biases on these quantifiable variables.

The aggregates and intersections of RdRp OTU virome compositions between the different libraries were then analyzed using the 'UpSet' package in R (Figures A1A and B). Approximately

40% (VO site) to 50% (CT site) of the OTUs identified in a given library were found to be reproducibly detected and shared between all 9 libraries amplified from the same purified dsRNA pool. The other OTUs were only detected from some or even from a single one of the libraries (Figures A1A and B). For the VO site, taking into account the aggregate of unique detected OTUs increases the number of OTUs from an average of 40 per library to a grand total of 79 OTUs, with 16 of them (20.2%) representing a fully reproducible core virome. The comparable values for the CT site are respectively 34 and 58 OTUs (17 OTUs or 29.3% for the core virome) clearly highlighting some inter-amplification variability in the detected virome (Figures A1A and B). Overall, only close to 50-60% (49-57%) of OTUs were detected from 5 or more of the 9 libraries generated for a given site. However, ANOSIM analyses of virome similarities between replicates show that these representational differences universally occurred among PCR replicates, irrespective of the specific MID tag for both the VO (R=-0.3374, *p*-value=0.974) and CT site with (R=-0.06584, *p*-value=0.622) (Figures A1C and D), .which is consistent with the quantifiable evaluation results described above and are also applicable to non-normalized datasets (data not shown).

Given that viruses do not share conserved regions that could be used for targeted (barcoding) amplification, the use of a WGA procedure is critical step in virome analysis. The results presented in this Annex show that the choice of MID tag does not seem to introduce a bias during this step. They also surprisingly show a significant variability in the WGA amplification output between independent PCR replicates. The pools analyzed here are very complex and it is conceivable that random stochastic effects during the first PCR cycles might influence the end results and the representation in the final sequenced library of particular agents. In particular, it should be noted that the failure to detect an OTU does not necessarily mean that the corresponding virus is absolutely not represented among the sequencing reads but merely that not contig encoding the polymerase core domain and passing the quality criteria could be assembled from them. Indeed, a

detailed analysis shows that the core OTUs shared between replicates tend to aggregate high number of reads while more erratic OTUs tend to correspond to low reads number (not shown). However, OTUs represented by high reads numbers were on occasion also observed to vary between replicates so that merely increasing the sequencing depth may not completely solve the representation problems highlighted here. On the other hand, reducing the complexity of the pools should at least partially alleviate this difficulty. In any case, when comparing two independent WGA amplification, shared OTUs appear to represent on the order of 75% of the total OTUs, a fraction significant enough to establish the relatedness of the corresponding viromes. As suggested for other microbiome studies (Zinger *et al.*, 2019), a few guidelines for virome analysis from complex "lawn mower" pools based on the results in this study are suggested including (i) limiting if experimentally feasible the complexity of pools (or conversely increasing sequencing depth) and (ii) performing independent multiple amplifications that can later be pooled prior to sequencing.

References

Bernardo P, Charles-Dominique T, Barakat M, Ortet P, Fernandez E, Filloux D, Hartnady P, Rebelo TA, Cousins SR, Mesleard F, Cohez D, Yavercovski N, Varsani A, Harkins GW, Peterschmitt M, Malmstrom CM, Martin DP, Roumagnac P (2018). Geometagenomics illuminates the impact of agriculture on the distribution and prevalence of plant viruses at the ecosystem scale. ISME J 12**:** 173-184.

Marais A, Faure C, Bergey B, Candresse T (2018). Viral Double-Stranded RNAs (dsRNAs) from Plants: Alternative Nucleic Acid Substrates for High-Throughput Sequencing. Methods Mol Biol 1746**:** 45-53.

Roossinck MJ, Saha P, Wiley GB, Quan J, White JD, Lai H, Chavarria F, Shen GA, Roe BA (2010). Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol Ecol 19**:** 81-88.

Roossinck MJ, Martin DP, Roumagnac P (2015). Plant Virus Metagenomics: Advances in Virus Discovery. Phytopathology 105**:** 716-727.

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

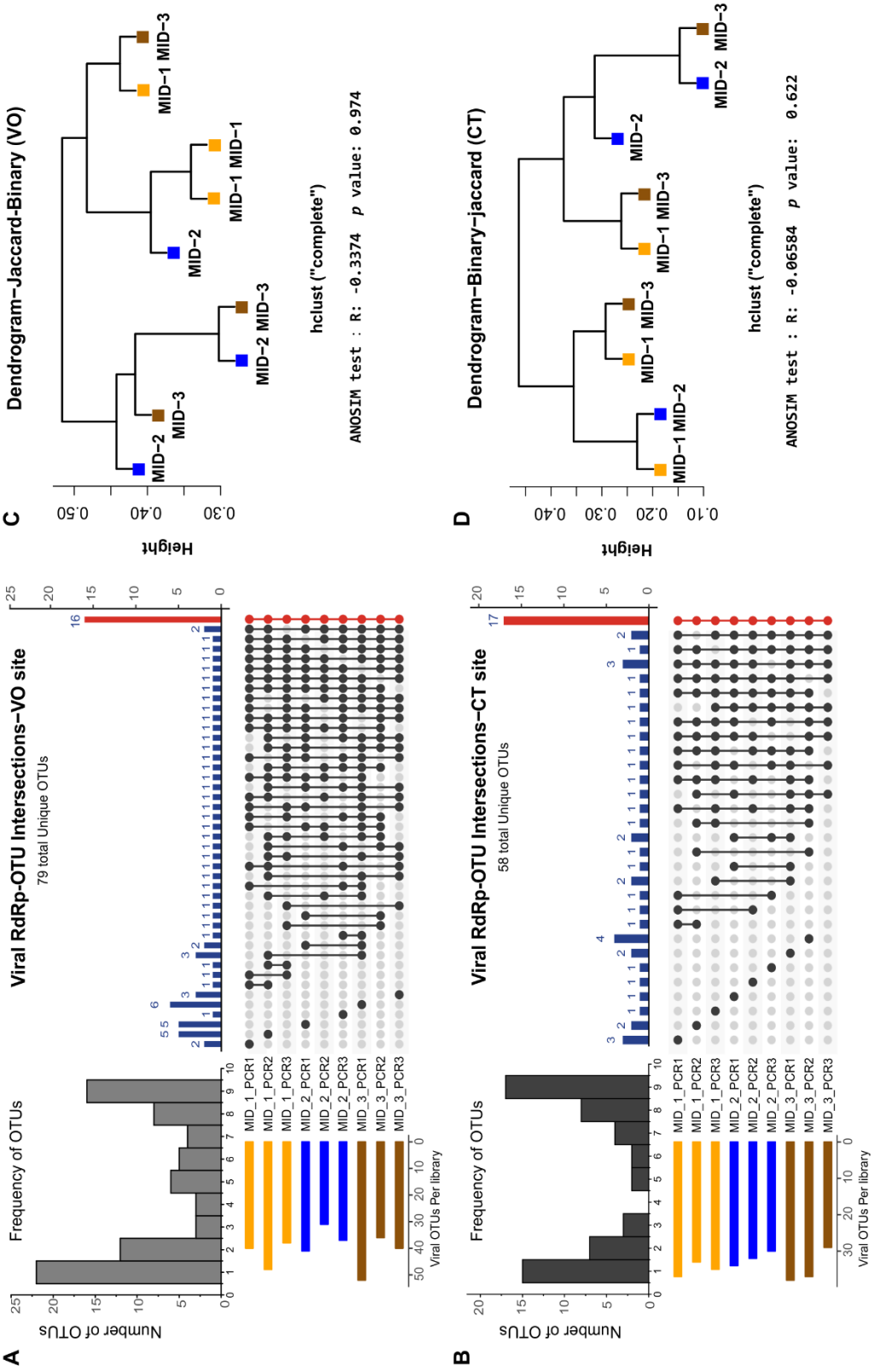
Zinger L, Bonin A, Alsos IG, Bálint M, Bik H, Boyer F, Chariton AA, Creer S, Coissac E, Deagle BE (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. Mol Ecol 28**:** 1857-1862.

Table A1. Libraries for phytovirome studies and characteristics of high throughput sequencing output for normalized datasets.

LIBRARY INFORMATION				NORMALIZED OUTPUT STATISTICS													
						Annotation of contigs and reads									Viral families and OTUs		
Library	Source plant pools	MID	Name	Percentage of assembled reads	Number of contigs	Eukaryota contigs	Bacteria contigs	Virus contigs	Unknown contigs	Percentage of Virus contigs	Reads in eucaryota contigs	Reads in bacteria contigs	Reads in virus contigs	Reads in unknown contigs	Percentage of reads in virus contigs	Contig_families	Viral OTUs	OTU-families
Lib1	**VO-viro-P**	MID-GENCO8	VO-viro-P-MID1.1	63%	**712**	413	5	124	168	17,4%	136366	119	65241	57754	25,1%	13	40	11
Lib2	**VO-viro-P**	MID-GENCO8	VO-viro-P-MID1.2	57%	**638**	393	3	100	142	15,7%	136335	173	65414	34692	27,6%	12	48	11
Lib3	**VO-viro-P**	MID-GENCO8	VO-viro-P-MID1.3	84%	**911**	493	4	156	258	17,1%	181974	227	92361	71202	26,7%	12	38	10
			MEAN	***0,681***	***754***	***433***	***4***	***127***	***189***	***16,7%***	***151558***	***173***	***74339***	***54549***	***26,5%***	***12,3***	***42,0***	***10,7***
			SD	***0,140***	***141***	***53***	***1***	***28***	***61***	***0,9%***	***26341***	***54***	***15608***	***18465***	***1,3%***	***0,6***	***5,3***	***0,6***
Lib4	**VO-viro-P**	MID-GENCO40	VO-viro-P-MID2.1	60%	**662**	363	0	127	172	19,2%	128710	0	56173	62413	22,7%	12	41	10
Lib5	**VO-viro-P**	MID-GENCO40	VO-viro-P-MID2.2	23%	**274**	158	4	47	64	17,2%	52780	86	21034	19202	22,6%	11	31	10
Lib6	**VO-viro-P**	MID-GENCO40	VO-viro-P-MID2.3	59%	**693**	406	2	122	162	17,6%	144029	112	57527	41406	23,6%	12	37	10
			MEAN	***0,472***	***543***	***309***	***2***	***99***	***133***	***18,0%***	***108506***	***66***	***44911***	***41007***	***23,0%***	***11,7***	***36,3***	***10,0***
			SD	***0,213***	***233***	***133***	***2***	***45***	***60***	***1,1%***	***48864***	***59***	***20689***	***21608***	***0,6%***	***0,6***	***5,0***	***0,0***
Lib7	**VO-viro-P**	MID-GENCO37	VO-viro-P-MID3.1	88%	**995**	561	2	182	248	18,3%	178152	37	110179	71912	30,5%	14	52	12
Lib8	**VO-viro-P**	MID-GENCO37	VO-viro-P-MID3.2	33%	**307**	193	1	41	72	13,4%	75468	115	35536	24293	26,2%	11	36	10
Lib9	**VO-viro-P**	MID-GENCO37	VO-viro-P-MID3.3	89%	**983**	582	4	173	223	17,6%	198634	88	89889	77143	24,6%	12	40	11
			MEAN	***0,698***	***762***	***445***	***2***	***132***	***181***	***16,4%***	***150751***	***80***	***78535***	***57783***	***27,1%***	***12,3***	***42,7***	***11,0***
			SD	***0,320***	***394***	***219***	***2***	***79***	***95***	***2,7%***	***65997***	***40***	***38595***	***29121***	***3,1%***	***1,5***	***8,3***	***1,0***
Lib10	**CT-viro-P**	MID-GENCO12	CT-viro-P-MID1.1	47%	**605**	233	3	184	185	30,4%	83448	177	212508	52777	60,9%	12	37	13
Lib11	**CT-viro-P**	MID-GENCO12	CT-viro-P-MID1.2	23%	**358**	144	1	116	97	32,4%	49978	214	100261	24879	57,2%	9	33	12
Lib12	**CT-viro-P**	MID-GENCO12	CT-viro-P-MID1.3	46%	**547**	200	1	196	150	35,8%	63734	27	218157	64382	63,0%	10	35	12
			MEAN	***0,387***	***503***	***192***	***2***	***165***	***144***	***32,9%***	***65720***	***139***	***176975***	***47346***	***60,4%***	***10,3***	***35,0***	***12,3***
			SD	***0,133***	***129***	***45***	***1***	***43***	***44***	***2,7%***	***16823***	***99***	***66497***	***20304***	***2,9%***	***1,5***	***2,0***	***0,6***
Lib13	**CT-viro-P**	MID-GENCO48	CT-viro-P-MID2.1	91%	**1858**	436	12	637	773	34,3%	125537	926	382789	170215	56,3%	11	34	13
Lib14	**CT-viro-P**	MID-GENCO48	CT-viro-P-MID2.2	91%	**2217**	527	16	741	932	33,4%	159021	2991	344099	173962	50,6%	11	32	13
Lib15	**CT-viro-P**	MID-GENCO48	CT-viro-P-MID2.3	30%	**525**	102	5	214	204	40,8%	18787	243	132152	75291	58,4%	11	30	13
			MEAN	***0,706***	***1533***	***355***	***11***	***531***	***636***	***36,2%***	***101115***	***1387***	***286347***	***139823***	***55,1%***	***11,0***	***32,0***	***13,0***
			SD	***0,349***	***892***	***224***	***6***	***279***	***383***	***4,0%***	***73237***	***1431***	***134930***	***55917***	***4,0%***	***0,0***	***2,0***	***0,0***
Lib16	**CT-viro-P**	MID-GENCO15	CT-viro-P-MID3.1	22%	**330**	110	1	107	112	32,4%	30570	86	112984	20787	68,7%	12	38	13
Lib17	**CT-viro-P**	MID-GENCO15	CT-viro-P-MID3.2	20%	**298**	129	2	81	86	27,2%	48350	2580	59576	36206	40,6%	9	37	12
Lib18	**CT-viro-P**	MID-GENCO15	CT-viro-P-MID3.3	44%	**547**	199	5	175	168	32,0%	64528	1536	214711	48694	65,2%	11	29	13
			MEAN	***0,285***	***392***	***146***	***3***	***121***	***122***	***30,5%***	***47816***	***1401***	***129090***	***35229***	***58,2%***	***10,7***	***34,7***	***12,7***
			SD	***0,135***	***135***	***47***	***2***	***49***	***42***	***2,9%***	***16985***	***1252***	***78812***	***13979***	***15,3%***	***1,5***	***4,9***	***0,6***

Figure A1. Virome composition comparisons between libraries obtained by independent whole genome amplifications (WGA) using either the same or different MID tags. (A) and (B) Detection of viral OTUs in the various libraries from the VO and CT sampling sites, respectively as illustrated by UpSet plots and histograms of OTU frequency (upper-left). The number of discovered OTUs are plotted at the bottom-left and colored according to the MID tags (MID-1, 2, 3 are colored orange, blue and brown respectively). The detection of OTUs in different libraries are shown in a matrix layout at the bottom-right, the aggregates based on the groupings and their corresponding numbers are plotted at upper-right. **(C) and (D) Hierarchical Clustering of viromes obtained by independent whole genome amplifications (WGA) using either the same or different MID tags for the VO and CT sampling sites, respectively.** Jaccard dissimilarity metrics were calculated based on the OTU presence/absence data and a clustering was performed using the "complete linkage" method in R. Analysis of similarities **(**ANOSIM) on the jaccard-binary distance matrix, testing whether similarity between groups is greater than the similarity within the groups are shown at the bottom of dendrogram ($0<R<1$ suggesting more similarity within groups; R values close to zero representing no difference between within groups and with groups; $-1<R<0$ suggests more similarity between groups than within groups).

Figure A1


A
Viral RdRp-OTU Intersections–VO site
79 total Unique OTUs
Number of OTUs
Frequency of OTUs
MID_1_PCR1
MID_1_PCR2
MID_1_PCR3
MID_2_PCR1
MID_2_PCR2
MID_2_PCR3
MID_3_PCR1
MID_3_PCR2
MID_3_PCR3
Viral OTUs Per library
B
Viral RdRp-OTU Intersections–CT site
58 total Unique OTUs
Number of OTUs
Frequency of OTUs
Viral OTUs Per library
C
Dendrogram–Jaccard-Binary (VO)
MID–2
MID–3
MID–2 MID–3
MID–2
MID–1 MID–1
MID–1 MID–3
Height
hclust ("complete")
ANOSIM test : R: -0.3374 p value: 0.974
D
Dendrogram–Binary–jaccard (CT)
MID–1 MID–2
MID–1 MID–3
MID–1 MID–3
MID–2
MID–2 MID–3
Height
hclust ("complete")
ANOSIM test : R: -0.06584 p value: 0.622

Annex B to Chapter III – Comparison of two DNA extraction kits and of ITS1 and ITS2 amplicons for the analysis of fungal communities

As described in the manuscript, to study plant-associated mycobiomes, a total of 16 plant pools (50 individual plants per pool, 7 to 10 plant species per pool) were prepared from the 4 different sampling sited (4 pools per site). Total DNA was extracted from these 16 pools using two different kits (Powersoil and DNeasy) and ITS1 and ITS2 regions sequenced using both extracts so that overall the 16 plant pools generated a total of 64 sequencing libraries (16×2×2) plus an additional 4 blank control libraries (two for each extraction kits, respectively amplified for ITS1 and ITS2, Table A2). The multiplexed MiSeq sequencing of theses 64 libraries generated 31011 to 58906 raw reads per library with an average of 43254 +/- 6710 reads (Table A2). After trimming and processing through the DADA 2 pipeline and contamination correction as described in the manuscript, the clean reads per library ranged from 18828 to 43335 with an average of 30192 +/- 6596 reads, showing that on average 69% of reads were retained for further analyses (Table A2).

The number of ITS1 reads (45269 +/- 6288 reads per library) generated with the DADA pipeline was significantly higher (p = 0.0004, paired t-test) than the number of ITS2 reads (41239 +/- 6601 reads per library) (Table A2 and Table A3). ITS1 similarly yield more cleaned reads (avg. 33995 +/- 5443 reads) than ITS2 (average 26388 +/- 5378; p = 1.65E-08) (Table A2, Table A3 and Figure A2A). Moreover, after removal of non-fungal reads according to the taxonomic assignation, a higher proportion of fungal reads was observed for ITS1 libraries (average 30548 +/- 5428 reads, 89.9%) than for ITS2 libraries (average 15129 +/-8072, 55.2%; p = 3.19E-10) (Table A2, Table A3 and Figure A2B).

As expected, a greater number of fungal ASVs were identified from ITS1 amplicons than from ITS2 ones (152 +/- 54 versus 119 +/- 51 ASVs per library, p = 0.0002) no matter what extraction kit had been used (Table A2, Table A3 and Figure A2C). However, the evenness of ITS2 libraries was significant higher that of ITS1 ones (p<0.001) (Table A3 and Figure A2D).

Overall, our results with complex plant pools involving a wide range of plant species demonstrate that ITS1 amplicons generated more sequences taxonomically assigned to fungi and uncovered a greater fungal richness in the complex plant pools analyzed. In particular the ITS1 amplicons showed a higher and much less variable percentage of fungal reads, demonstrating a higher robustness when confronted to a wide range of samples containing different plant species (Figure A2B). As used in several large-scale microbiome projects, only targeting ITS1 region may provide insufficient resolution to distinguish species so the use of additional or alternative markers are suggested (Nilsson *et al.*, 2019). Using ITS1F and ITS2 primer pair for ITS1 targeting may also suffer from primer biases and the presence of an intron that is common in several fungal groups, which can lead to biased amplification (Tedersoo and Lindahl, 2016). On the other hand, no statistically significant differences were found between the two DNA extraction kits on the majority of variables, except for the evenness of ITS2 ASVs (p=0.026) (Table A3 and Figure A2D).

Given that a better performance was obtained with the ITS1 amplicons (higher proportion of fungal reds, richer mycomes), only the ITS1 information was analyzed in detail in the manuscript. On the other hand, given that a comparable performance was obtained with the two kits the reads corresponding to the ITS1 amplicons obtained from DNA extracted with the two kits were aggregated in some cases for the fungal diversity analysis. Effort on refined compositional analysis of fungal communities derived from different approach could be made in the future.

References

Nilsson RH, Anslan S, Bahram M, Wurzbacher C, Baldrian P, Tedersoo L (2019). Mycobiome diversity: high-throughput sequencing and identification of fungi. Nature Reviews Microbiology 17**:** 95-109.

Tedersoo L, Lindahl B (2016). Fungal identification biases in microbiome projects. Environ Microbiol Rep 8**:** 774-779.

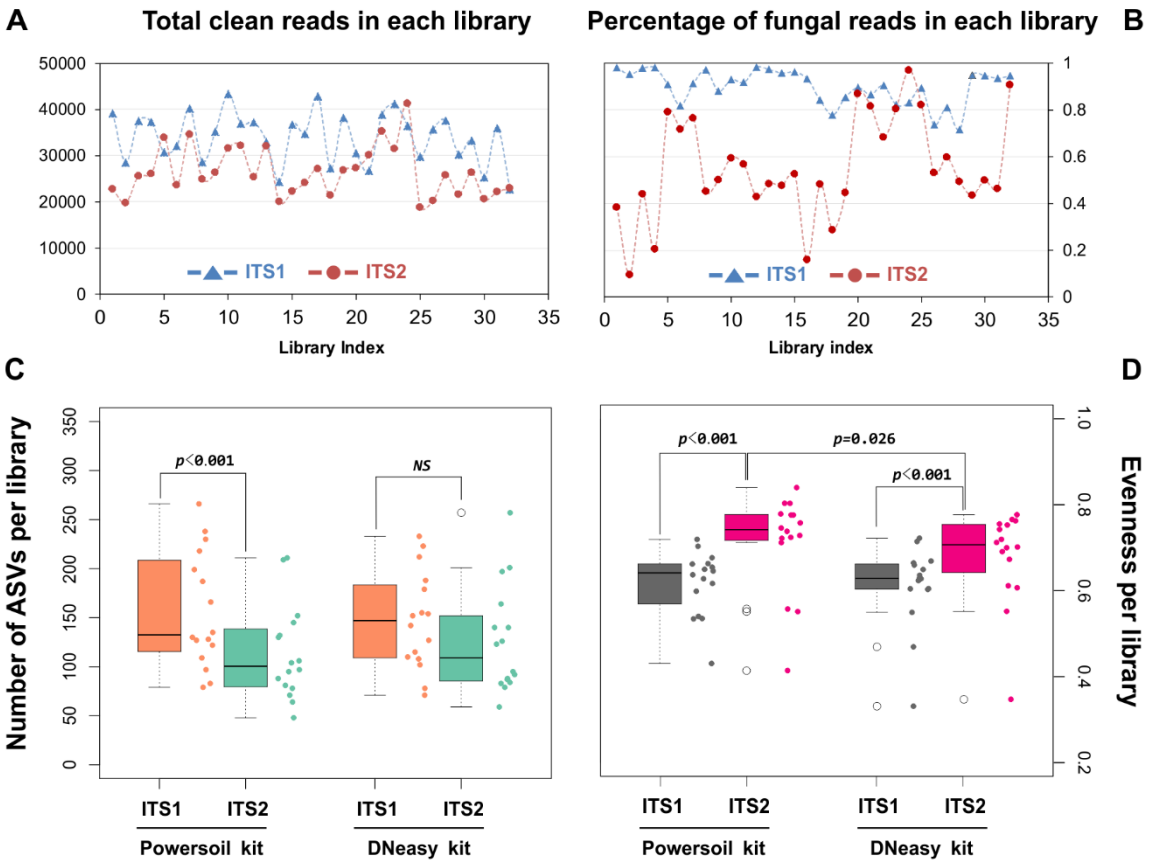
Table A2. Reads, ASV and diversity indices for ITS1 and ITS2 amplicons obtained for the various study sites from DNA extracted using different kits

Kit	Site	Source material pool	Name	ITS1							ITS2						
				Raw reads	clean reads	Fungal reads	ASV	Shannon	Simpson	Evenness	Raw reads	clean reads	Fungal reads	ASV	Shannon	Simpson	Evenness
Powersoil	VO	VO-myco-P1	1P	50455	39149	38379	83	2,78	0,89	0,63	40406	22721	8722	64	3,05	0,91	0,73
		VO-myco-P2	2P	39040	28496	27156	122	3,08	0,91	0,64	33770	19749	1870	48	3,30	0,95	0,85
		VO-myco-P3	3P	51073	37538	36725	127	3,21	0,93	0,66	43677	25607	11255	78	3,16	0,91	0,72
		VO-myco-P4	4P	47096	37261	36536	79	2,40	0,82	0,55	41335	26021	5310	71	3,48	0,95	0,82
	SP	SP-myco-P1	5P	47361	30670	27878	230	3,89	0,95	0,72	47786	33877	26817	209	3,94	0,96	0,74
		SP-myco-P2	6P	40147	32097	26263	130	3,29	0,93	0,68	34439	23595	16930	97	3,39	0,95	0,74
		SP-myco-P3	7P	51052	40189	36685	218	3,94	0,96	0,73	48981	34601	26448	211	4,06	0,97	0,76
		SP-myco-P4	8P	37181	28590	27775	166	3,37	0,92	0,66	36662	24908	11247	130	3,85	0,96	0,79
	CT	CT-myco-P1	9P	47509	35210	30999	109	2,60	0,79	0,55	42840	26310	13164	95	2,57	0,74	0,57
		CT-myco-P2	10P	52351	43335	40317	97	2,04	0,67	0,45	48572	31597	18748	81	1,88	0,55	0,43
		CT-myco-P3	11P	47748	36916	33863	135	2,69	0,80	0,55	50567	32197	18233	106	2,66	0,79	0,57
		CT-myco-P4	12P	46180	37229	36585	128	2,97	0,89	0,61	39336	25346	10861	104	3,49	0,93	0,75
	INRA	INRA-myco-P1	13P	44833	32975	32085	199	3,53	0,94	0,67	50293	32114	15560	145	3,92	0,97	0,79
		INRA-myco-P2	14P	35525	24310	23300	187	3,53	0,92	0,68	31959	20059	9559	132	3,85	0,96	0,79
		INRA-myco-P3	15P	51090	36751	35353	238	3,56	0,91	0,65	40275	22275	11692	152	3,87	0,96	0,77
		INRA-myco-P4	16P	46616	34713	32415	266	3,85	0,94	0,69	37579	24110	3831	88	3,65	0,95	0,82
		Negative Control-Powersoil	NC-Pl	15966	44514	na	na	na	na	na	64402	46171	na	na	na	na	na
DNeasy	VO	VO-myco-P1	1D	53611	42862	36105	108	2,89	0,89	0,62	42128	27120	13054	87	3,24	0,92	0,72
		VO-myco-P2	2D	33465	27211	**21168**	127	3,26	0,92	0,67	34084	21375	6105	84	3,39	0,93	0,77
		VO-myco-P3	3D	48574	38187	32631	142	3,22	0,93	0,65	41046	26852	11950	95	3,25	0,92	0,71
		VO-myco-P4	4D	39872	30507	27314	71	2,40	0,81	0,56	43412	27292	23736	79	3,00	0,92	0,69
	SP	SP-myco-P1	5D	43276	26713	23121	212	3,90	0,95	0,73	45712	30171	24619	201	3,88	0,96	0,73
		SP-myco-P2	6D	47556	38805	35124	154	3,24	0,91	0,64	53932	35307	24080	140	3,52	0,95	0,71
		SP-myco-P3	7D	51948	41216	33892	223	3,97	0,96	0,73	46653	31478	25329	197	4,09	0,97	0,77
		SP-myco-P4	8D	48681	36459	30269	155	3,21	0,91	0,64	54978	41242	39967	257	3,47	0,88	0,62
	CT	CT-myco-P1	9D	42479	29732	26591	102	2,85	0,85	0,62	40696	18828	15465	88	2,78	0,80	0,62
		CT-myco-P2	10D	42740	35671	26246	78	1,51	0,45	0,35	33026	20246	10742	59	1,48	0,45	0,36
		CT-myco-P3	11D	58906	37604	30488	115	2,29	0,69	0,48	43637	25755	15378	92	2,56	0,77	0,57
		CT-myco-P4	12D	37969	30264	21666	110	2,90	0,87	0,62	33505	21623	10664	83	3,11	0,91	0,70
	INRA	INRA-myco-P1	13D	47426	33270	31539	188	3,55	0,94	0,68	37550	26312	11424	123	3,80	0,96	0,79
		INRA-myco-P2	14D	37528	25275	23915	179	3,33	0,89	0,64	31011	20603	10296	126	3,76	0,96	0,78
		INRA-myco-P3	15D	46935	35987	33683	233	3,72	0,93	0,68	35663	22186	10278	140	3,79	0,96	0,77
		INRA-myco-P4	16D	32409	22662	21470	152	3,33	0,90	0,66	34134	22942	20816	164	3,85	0,96	0,76
		Negative Control-Dneasy	NC-D	51072	13818	na	na	na	na	na	44650	32504	na	na	na	na	na

Table A3. Paired *t*-test evaluating the statistical significance of differences for different variables between ITS1 and ITS2 amplicon libraries and between Powersoil and DNeasy kits amplicon libraries.

***p*-value (paired test)**	**Raw reads**	**Clean reads**	**Fungal reads**	**Observed_ASV**	**Shannon (H)**	**Simpson (1 - D)**	**Evenness (J')**
ITS1 vs ITS2	**0,0004**	**1,65E-08**	**3,19E-10**	**0,0002**	**6,24E-05**	**0,01752**	**2,51E-08**
Powersoil vs DNeasy	**Raw reads_its1**	**Clean reads_its1**	**Fungal reads_its1**	**Observed_ASV_its1**	**Shannon (H)_its1**	**Simpson (1 - D)_its1**	**Evenness (J')_its1**
	0,4691	0,3033	0,0144	0,2063	0,2324	0,1773	0,3988
	Raw reads_its2	**Clean reads_its2**	**Fungal reads_its2**	**Observed_ASV_its2**	**Shannon (H)_its2**	**Simpson (1 - D)_its2**	**Evenness (J')_its2**
	0,6417	0,8389	0,1192	0,2345	0,2316	0,2423	**0,02625**

Figure A2. Comparison of ITS1 or ITS2 amplicons-based mycobiome description. (A) Total number of clean reads in each library generated by sequencing ITS1 or ITS2 amplicons. **(B)** Percentage of fungal reads in each library generated by sequencing ITS1 or ITS2 amplicons. Boxplot of the number of amplicon sequence variants (ASVs) reflecting the fungal community richness **(C)** and values of Pielou's evenness **(D)** in libraries generated by sequencing ITS1 or ITS2 amplicons generated from total nucleic acids purified using the Powersoil or DNeasy kits. Only significant *p* values of paired *t*-tests are indicated above the connecting lines.



CHAPTER Ⅳ

Metagenomic analysis of virome cross-talk between cultivated Solanum lycopersicum and wild Solanum nigrum

Metagenomic analysis of virome cross-talk between cultivated *Solanum lycopersicum* and wild *Solanum nigrum*

Yuxin Ma, Armelle Marais, Marie Lefebvre, Chantal Faure, and Thierry Candresse*

UMR 1332 BFP, INRA, Univ. Bordeaux, CS20032, 33882 Villenave d'Ornon cedex, France

Abstract: 148 word

Text: 5029 words (Introduction- Materials and Methods-Results-Discussion-Acknowledgements)

* Corresponding author thierry.candresse@inra.fr

The assembled viral genomic reported here have been deposited in Genbank under accession numbers **MN216346** to **MN216389**. Cleaned virome sequence reads have been deposited on the INRA National Data Portal under the identifier https://doi.org/10.15454/S486RR.

ABSTRACT

Wild plants and weeds growing close to crops constitute a potential reservoir for future epidemies or for the emergence of novel viruses but the frequency and directionality of viral flow between cultivated and wild plants remains poorly documented in many cases. Here, we studied the diversity and potential flow of viral populations between tomato (*Solanum lycopersicum*) and neighboring european black nightshade (*Solanum nigrum*) using high throughput sequencing (HTS) based metagenomics. A large variability in virome richness with only 17.9% shared Operational Taxonomy Units between tomato and nightshade could not be linked to a particular host or to local conditions. A detailed population analysis based on assembled contigs for potato virus Y (PVY), broad wilt bean virus 1 and a new ilarvirus tentatively named Solanum nigrum ilarvirus 1 provides information on the circulation of these viruses between these two *Solanum* species and enriches our knowledge of the tomato virome.

KEYWORDS: metagenomics, virome, double stranded RNA (dsRNA), tomato, spillover

INTRODUCTION

Through the past decade, metagenomics based on high throughput sequencing (HTS) has been widely used in the plant virology field, advancing our knowledge on the diversity of plant viruses. Specifically, metagenomics allowed to discover unknown viruses, explore the intraspecific genetic diversity of known viruses, and study virus ecology and epidemiology (Massart et al., 2014; Roossinck et al., 2015; Villamor et al., 2019). Plant viruses cause epidemics on all major cultures of agronomic importance, representing a serious threat to global food security. As a consequence, virologists have for a long time focused their efforts on economically important crops, often neglecting bordering weeds and wild plants (Wren et al., 2006). However, agro-ecosystems are complex environments in which crop plants sometimes interact with the in-plot and bordering weeds and wild plants while viruses may be transferred between wild plants/weeds and crops and vice versa by a variety of mechanisms and vectors. Thus wild plants or weeds may constitute "reservoirs" of viruses that may subsequently spread to cultivated plants while crops may constitute a source from which viral infections may spillover to the wild plants/weeds compartment (Power and Mitchell, 2004).

Overall, our understanding of the details of fluxes of viruses from crops to weeds and from weeds to crops remains frequently limited. A role of a weed population as a reservoir or spillover from a crop are often assumed but the techniques used for the characterization of viral populations by classical plant virus epidemiology, in particular serological ones, frequently do not provide sufficient intra-specific resolution. It is then difficult to ensure that the co-occurrence of a virus in crops and weeds reflects the transfer of isolates rather than the existence of separate viral populations adapted to the two host populations.

Tomato (*Solanum lycopersicum*) is one of the most popular and extensively consumed vegetable crops. There are at least 136 characterized viral species that are capable of infecting tomato and due to global climate changes and increased international trade, the spread of known viruses to new geographic areas and the emergence of new viruses have been frequently detected in particular in recent years (Brunt, 1996; Hanssen et al., 2010). Torradoviruses (family *Secoviridae*) are an example of a group of recently emerged plant viruses, many of which affect tomato. These include for example tomato torrado virus (ToTV), which was first described from tomato in Mexico (Verbeek et al., 2008) and reported more recently in France (Verdin et al., 2009) and in other host plant species (van der Vlugt et al., 2015), as well as tomato marchitez virus (ToMarV; (Verbeek et al., 2008)) and tomato chocolàte virus (ToChV; (Verbeek et al., 2010)). Another example of recent emergence of a virus in tomato concerns tomato brown rugose fruit virus (ToBRFV), a tobamo like virus which was discovered from tomato in Israel in 2014 (Luria et al., 2017) and that has spread since then to many countries including Jordan, Mexico, the United States (Southern California), Germany, Italy, Turkey, the Netherlands and Saudi Arabia. The source(s) and cause(s) of the emergence of such novel agents is(are) frequently unknown but weed and wild plants are often considered as a major sources of future emerging viruses than may occasionally be transferred to crops (Anderson et al., 2004; Elena et al., 2014; McLeish et al., 2019).

Recently, during a study characterizing the virome of 170 field-grown tomatoes collected in China by small RNAs sequencing, Xu et al. (2017) showed that the tomato viral community is dominated by a few species, most of them being positive-sense ssRNA viruses. Multiple infections were found to be frequent as well as recombination events in viral genomes (Xu et al., 2017).

European black nightshade (*Solanum nigrum*), a wild relative of tomato is a widespread weed in many countries. However, in southern India it is widely consumed and cultivated on a commercial scale (Jamuna et al., 2017) and sometimes also used as for its medicinal properties (Javed et al., 2011). *S. nigrum* is known to harbor a wide range of viruses such as begomoviruses, orthotospoviruses, potyviruses, tobraviruses under field conditions, and has often been suspected to act as a reservoir host for viruses infecting solanaceous crops (Holm et al., 1979; Jamuna et al., 2017).

In the present study, using a metagenomics approach, we investigated and compared the virome in tomato samples and in the related *S. nigrum* populations collected either in tomato fields or in various other environments. The comparison of these viromes provides novel insight into the viral fluxes between these two species.

MATERIALS AND METHODS

Study sites and plant samples

Virome richness and composition were analyzed in tomato (*Solanum lycopersicum*) and in European black nightshade (*Solanum nigrum*) that were growing either close to the sampled tomato crops or in other sites, unmanaged or involving unrelated crops (sunflower, maize, sorghum and alfalfa, Table S1). In total, tomato crops were sampled in seven sites and nightshade populations in six of the seven tomato sites, plus in five non-tomato sites (Table S1). For each sampled plant populations, leaves from a total of 100 individual plants were collected in summer 2017 or 2018 and assembled in two pools corresponding to fifty individual plants (0.1g of

leaf/plant) for nucleic acids extraction. No specific efforts were made to select symptomatic plants, but plants with obvious fungal attack, insect colonization or necrotized parts were excluded.

Double-stranded RNAs purification, library preparation and Illumina HiSeq sequencing

Double-stranded RNAs were purified from each plant pool by two rounds of CF11 cellulose chromatography and converted to cDNA according to the protocol described by Marais et al. (2018). In parallel, a negative control blank was similarly prepared using only buffer. Whole genome amplifications (WGAs) were performed on each cDNA sample (using the same MID tag for the two pools of each sampling site), the PCR products were purified using the MinElute PCR Purification Kit (Qiagen) and their concentration determined spectrophotometrically (Marais et al., 2018). Equal quantities of the amplification products from the two pools of each sampling site were then regrouped and independent sequencing libraries prepared for each site and sequenced in multiplexed format (2×150 bp) on an Illumina HiSeq 3000 system at the GenoToul platform (INRA Toulouse, France). Cleaned virome sequence reads have been deposited on the INRA National Data Portal under the identifier https://doi.org/10.15454/S486RR.

Bioinformatics analyses: Reads cleaning, contigs assembly and annotation, Operational Taxonomic Units (OTU) clustering

Following demultiplexing, adapters and MID tags were removed with *cutadapt* (Martin, 2011), and reads were quality trimmed (minimum quality score 20, minimum length 70 nucleotides). In order to limit inter-sample cross talk associated with index-hopping (Illumina, 2017; van der Valk et al., 2019), only reads having identical MID tags on both pair members were retained for further

analyses (Table S1). Contigs were *de novo* assembled for each library using IDBA-UD (https://academic.oup.com/bioinformatics/article/28/11/1420/266973).

All contigs were annotated using BlastN and BlastX against the NCBI Genbank non redundant nucleotide (nt) or protein databases with a conservative e-value cut-off of 10^{-4}. In this way, contigs were assigned to one of the following categories: virus, eukaryote, bacteria, algae, and unknown. For viral contigs, a family-level annotation was derived from the NCBI taxonomic information for the first Blast hit.

A clustering approach (Lefebvre et al., 2019) was used to define operational taxonomy units, following the strategy highlighted by Simmonds (2015). Briefly, a search of RNA-dependent RNA polymerase (RdRp) conserved protein motifs was performed in all contigs using Reversed Position Specific Blast (RPS-Blast) (Altschul et al., 1997) against the pfam database (Bateman et al., 2018). The contigs encoding a viral RdRp motif (Table S1) were retrieved and aligned with reference sequences and distance matrices computed with the ETE3 toolkit (Huerta-Cepas et al., 2016). These matrices were used to perform a clustering allowing to regroup in a single OTU all contigs differing by less than a set cut-off divergence value (Murtagh and Legendre, 2014). We used a 10% divergence cut-off value, because it has been shown to generate in many viral families OTUs that are a relatively good approximation of taxonomic species (Lefebvre et al., 2019). OTUs were thus defined for each RdRp family, allowing to generate an OTU table indicating for each sampling site the presence/absence and the number of reads integrated in each identified OTU (Table S1).

Further viral genome assemblies, sequence comparisons and phylogenetic analyses

When needed, contigs were extended by repeated rounds of mapping of quality-trimmed reads using CLC Genomics Workbench 11 (CLC-GWB). For some isolates/viruses, genomic scaffolds were assembled by mapping contigs and/or reads on a reference genome using CLC-GWB. Long contigs or scaffolds showing more than 75% completeness for cucumber mosaic virus (CMV), southern tomato virus (STV), broad wilt bean virus 1 (BBWV1, both genomic RNAs), the new ilarvirus (all three genomic RNAs) and potato virus Y (PVY) were used for phylogenetic analyses and have been deposited in Genbank (Accession numbers MN216356 to MN216369 (Table S2).

Multiple sequence alignments of contigs/scaffolds obtained from HTS data and of reference isolates retrieved from Genbank (or alignments of deduced encoded proteins) were performed using the ClustalW algorithm (Thompson et al., 1994) as implemented in MEGA 6.0 (Kumar et al., 2008). Phylogenetic trees were reconstructed in MEGA 6.0, using strict nucleotide or amino acid distances and the Neighbor Joining (NJ) algorithm. Branch support was evaluated by bootstrap analysis (1000 replicates).

RESULTS

Comparison of the tomato and nightshade viromes at different sampling sites

A total of 20 viral families were discovered by Blast annotation taking into account the different libraries (18 sampled plant populations) with an average of 4.3±3.3 families per library, but with a very large variability between the sampled plant populations. The tomato sample from the TOM3 site showed the highest number of viral families (13, Figure 1) followed by another tomato sample (TOM7, 9 viral families) and nightshade samples from the TOM2 and NIG3 sites (8 viral families). The *Potyviridae* family was represent in a total of 13 samples including both tomato (six samples)

and nightshade (seven samples, of which five were from tomato sites; Figure 1). The family *Totiviridae* was represented in eight samples while at the other extreme the *Tombusviridae* family was represented in a single tomato sample from the TOM3 site. Given the high between-populations variability it was not possible to establish statistically meaningful differences in family-level richness between the tomato and nightshade populations (Figure 1).

Taking into account all sampling sites, a total of 87 unique RNA-dependent RNA polymerase (RdRp) OTUs were detected (Table S1). Similar to the family-level analysis, a very large variability was observed in the number of OTUS detected per site. The richer viromes were found in the TOM7 site tomato population and in the NIG3 nightshade population, with respectively 38 and 27 OTUs, followed by 26 OTUs for the TOM3 site tomato population. In all other samples less than 8 RdRp OTUs were detected (Table S1).

In total, 62 OTUs were identified from tomato samples and 44 from nightshade ones but this difference is largely the consequence of a single tomato sample (TOM7) which is particularly rich in unique mycovirus-like OTUs (Table S1). Nineteen OTUs (21.8% of total) were shared between the two plant species, most of them from the families *Totiviridae, Partitiviridae and Chrysoviridae* as well as unclassified mycovirus-like OTUs. RdRP_1-OTU_8 which corresponds to potato virus Y was the most widely shared OTU (Table S1, see below). It explains the wide prevalence of the *Potyviridae* family described above. Twenty-five OTUs were found to be nightshade-specific, among which RdRP_2-OTU_13 corresponds to cucumber mosaic virus (CMV) and RdRP_1-OTU_14 to broad bean wilt virus 1 (BBWV1) (Table S1, see below). Fourty-three OTUs were found to be tomato-specific, some of which have extremely high identity levels with known viruses

such as Sclerotinia sclerotiorum hypovirus 1, Sclerotinia sclerotiorum umbra-like virus 2 or Botrytis virus F and very likely correspond to these agents (Table S1).

There were overall only very few OTUs shared between tomato and nightshade samples for a given sampling site, with PVY being the most frequent. In five sites out of six, no OTU (sites TOM2 and TOM6) or only one OTU (sites TOM1, TOM4, and TOM5) were shared, whereas in site TOM3, four OTUs were shared (Table S1, Figure S1).

Near complete genome reconstruction for selected viral agents

For several viruses, long, high quality contigs were obtained during the initial trimmed reads assembly. This concerned in particular several single-stranded RNA viruses: cucumber mosaic virus (CMV), broad wilt bean virus 1 (BBWV1, both genomic RNAs), potato virus Y (PVY), and a new ilarvirus (all three genomic RNAs) as well as a double-stranded RNA virus of the *Amalgaviridae* family, southern tomato virus (STV). In a few cases, the viral genome was unambiguously covered by a few contigs that were either non-overlapping or had only a short overlap and which were therefore manually assembled into a scaffold by mapping contigs on a reference genome. All contigs and scaffolds were validated by visual inspection of read mappings at high stringency to ensure the absence of assembly artifacts. The corresponding sequences have been deposited in Genbank (Accession numbers MN216346 to MN216389, Table S2).

Multiple alignments and phylogenetic analyses (see below) were used to identify representative contigs for the various phylogenetic clusters of each virus. These representative contigs were in turn used a targets for the mapping of the trimmed reads of all libraries at high stringency. This allowed to evaluate the representation of each virus/variant in the virome of each sampled plant

population. The low background of viral reads observed in the negative control, probably resulting from low level experimental contamination or from inter sample cross talk due to index-hopping (Illumina, 2017), was substracted from the mapped reads numbers of each library. The results of this analysis are presented in Table S3.

CMV was detected, by high read numbers, at a single sampling site (TOM1), in the nightshade population but not in the corresponding tomato population (Table S3). All three genomic RNAs were assembled into unique long contigs of respectively 3,301 nt (RNA1; ca. 98.2% of the full length molecule), 2,996 nt (RNA2, ca. 98.3% of the full length molecule) and 2,155 nt (RNA3, ca. 97.2% of the full length molecule) but no evidence was found for presence of a CMV satellite. Despite the fact that no specific efforts were made to improve/validate the contigs further, all three genomic RNAs are extremely close to CMV sequences present in Genbank and, in particular to the I17F isolate, a subgroup I isolate characterized from tomato in France at the beginning of the 1980's (Jacquemond and Lot, 1981). Nucleotide identity levels of respectively 99.5%, 99.4% and 99.5% for genomic RNA1, 2 and 3 (respectively 18, 18 and 11 point mutations) are thus observed between the 1981 IF17 isolate and the contigs from HTS data on a 2017 sample, highlighting both the quality of the HTS assemblies and the relative stability of the CMV population over more than 35 years.

In the case of southern tomato virus (STV), unique long contigs representing nearly complete genomes were obtained from several plant populations, representing 92.6%-99.4% of the full length genome. Coherent with the low diversity identified so far in this virus, these contigs are nearly identical to each other (<0.4% nucleotide divergence) with the exception of one contig, which diverges by 2.6%-2.8% from the others. Identity levels with isolates present in Genbank

range from 100% to 95.9%, again highlighting the quality of the contigs assembled from the HTS data. Overall STV was detected in five of the seven tomato pools, an observation in accordance with the presence of this virus in a wide range of tomato varieties (Sabanadzovic et al., 2009). On the other hand, a surprising result is the detection, with higher reads number than for the tomato pools (Table S3) of STV in a nightshade pool (Nightshade-TOM5), extending the host range of this relatively recently discovered virus. The nightshade STV sequence belongs to the group of closely related isolates and does not present obvious specific molecular properties (data not shown).

Broad bean wilt virus 1 populations diversity

Broad bean wilt virus 1 (BBWV1) was detected in six of the sampled nightshade populations (out of a total of 11 populations, 55%) and was not detected in any of the sampled tomato populations (seven populations, Table S3). The assembly of the BBWV1 reads from the various nightshade populations highlighted a complex viral population structure with a total of five RNA1 clusters and three RNA2 clusters identified (Table S3, see below). On average, the reconstructed genomic sequences represented 94.3% +/- 3.9% of the BBWV1 RNA1 (87.8%-99.3%, depending on the contigs) and 87.8% +/- 11.9% of the BBWV1 RNA2 (73.4%-96.6%). For one sample, it was not possible to reconstruct more than 60% of the RNA2 sequence and the corresponding scaffold was therefore not included in further analyses. The average nucleotide divergence between the RNA1 clusters, calculated on representative isolates is 16.6% +/- 0.3% (13.4%-17.6%), explaining the effective separate assembly in cases of mixed infection by isolates belonging to different clusters. For the three RNA2 clusters, the corresponding divergence values are 15.8% +/- 0.6% (10.5%-18.6%). Mapping of reads at high stringency on contigs representative of the various clusters allowed to describe the BBWV1 population present in the various plant populations. Isolates

representative of between one and four RNA1 clusters and of one or two RNA2 clusters could thus be detected at individual sampling sites, with some sites providing evidence of only a single RNA1-RNA2 combination, while at the other extreme, one site showed the presence of four RNA1 clusters and a single RNA2 one. Another site showed the presence of a single RNA1 cluster but of two RNA2 ones (Table S3). Taken together, these elements suggest the frequent occurrence of reassortment between BBWV1 genomic segments in the sampled nightshade populations.

Phylogenetic analyses performed on the RNA1 and RNA2 sequences derived from the HTS data and from all full length isolates present in Genbank (Figure 2A and 2B) demonstrate that the BBWV1 isolates present in the nightshade populations sampled here largely expand the known BBWV1 diversity. Indeed, the HTS-derived sequences cluster separately from reference full-length sequences available to date and are, on average, highly divergent from them with an average intergroup distance of 17.0% +/- 0.4% for RNA1 and 19.0% +/- 0.6% for RNA2.

Presence of a novel ilarvirus in the sampled nightshade and tomato populations

Long, high quality contigs representative of an ilarvirus were identified in several libraries. The contigs corresponding to the three genomic RNAs were further extended and validated for the NIG4 sampling site, allowing to reconstruct near complete molecules. Indeed, a comparison with the genomic RNAs of Parietaria mottle virus (PMoV), the closest characterized ilarvirus (see below) indicated that all five open reading frames (ORFs) [coding respectively for P1 (RNA1), P2 and P2b (RNA2) and the movement (MP) and coat proteins (CP) (RNA3)] were complete, with the exception of ORF2 which misses an estimated 62 nt (21 N-terminal amino acids missing from the P2 protein sequence). The contigs are respectively 3,445, 2,757 and 2,257 nt long for RNA1,

RNA2 and RNA3, representing respectively 97.9%, 94.4% and 100.5% of the length of the corresponding genomic RNAs of the reference PMoV isolate (NC_005848, -49 and -54). These genomic sequences have been deposited in Genbank under Accession numbers MN216370 to MN216378. Blast analyses indicated that the virus is most closely related to PMoV and to several other subgroup 1 ilarviruses and this proximity was confirmed by phylogenetic analyses performed on all genome encoded proteins (Figure 3A and 3B, Figure S2). However, these phylogenetic trees demonstrate that the virus is not substantially more related to PMoV than to any other approved species in that small ensemble. The significant divergence of the virus from PMoV is confirmed by sequence comparisons, the deduced proteins being only 81.8% (P1) to 53.9% (CP) identical with those of PMoV while the genomic RNAs show only 73.2% (RNA1) to 58.6% (RNA3) nucleotide identity (Table S4). Taken together these results suggest that the detected ilarvirus is a new subgroup 1 member for which the name Solanum nigrum ilarvirus 1 (SnIV1) is proposed.

Mapping of the reads from each plant population on the SnIV1 genomic RNAs showed that this virus was present in eight of them, corresponding to 6/11 nightshade populations (54.5%) and, represented by relatively low read numbers, to 2/7 tomato populations (28%) (Table S3).

Analysis of PVY populations in the sampled nightshade and tomato populations

As for the other viruses, long, high quality contigs were obtained in most cases for PVY. In a few cases, probably resulting from low reads numbers or from the simultaneous presence of closely related isolates in the sampled plant populations, only short PVY contigs were obtained for some isolates. However, from all plants populations with high PVY read numbers, from one to three long contigs could be assembled presenting on average 95.4% +/- 4.8% of the full length PVY

genome (85.3%-99.9%). A phylogenetic analysis of these contigs, together with representative reference isolates retrieved from Genbank (Figure 4) shows a very contrasted situation, with on the one hand, a large number of sequences forming a very tight cluster corresponding to PVY-NTN and, on the other, a much more diverse second cluster corresponding to PVY-C. No isolates representative of the PVY-O and PVY-N strains were observed nor some of their frequent recombinants such as PVY-Wilga (Figure 4). In total, 10 contigs were obtained for PVY-NTN (five from tomato and five from nightshade) and four for PVY-C (three from tomato and one from nightshade).

The reads from all plant populations were then mapped on selected contigs representative of PVY-NTN and of the three PVY-C variants identified, using stringent parameters so as to limit cross-mapping between isolates. Under these conditions, from one to four PVY variants could be detected in the analyzed plant populations. Some populations showed extremely low read numbers (<90), which is suggestive of an absence or a very low prevalence of PVY in the corresponding plant populations. Remarkably, this situation corresponds to 2/7 (28.6%) tomato populations, to 2/6 (33.3%) nightshade populations growing side by side with tomato but to 4/5 (80%) of the nightshade population growing away from tomato.

As judged from the mapping results, the two most frequent PVY strains were PVY-NTN, which was detected in all tomato and nightshade populations in which PVY was detected, and isolates with mapping affinities with isolate TOM7-C, which clusters together with the French PVY-C1 SON41 pepper isolate (Table S3 and Figure 4). By contrast, isolates corresponding to the two other PVY-C mapping references used were only detected in one to three of the sampled plant

populations. The frequency of detection of the various clades does not seem to differ much between tomato and nightshade (Table S3).

DISCUSSION

The viromes highlighted in the present work vary greatly between the sampled plant populations and, for some of them, showed only a limited number of OTUs or of viral families despite the size of the composite plant samples analyzed. This might reflect the impact of fungicide treatments in the sampled crops which might have reduced fungal diversity and in turn the ability to detect mycoviral communities associated with the sampled plants. It should however be stressed that the OTU-based analysis provides a lower bound estimate of viral diversity since viruses for which the genome region encoding the conserved RdRp motif is not represented in the assembled contigs will not be identified by a corresponding OTU. On the other hand, competition between viruses for representation in the sequencing reads is unlikely to have adversely impacted the richness of the identified viromes since the three richest viromes were identified in plant populations for which the percentage of mapped viral reads was not obviously higher (or lower) than that observed in samples with a much lower viral diversity (Tables S1 and S2).

In contrast to a recent virome study of 170 tomato samples which indicated that diverse ssRNA viruses represented 77% of the identified viruses (Xu et al., 2017), they represented only 12.6% of the viral OTUs identified here (Table S1). The corresponding value for dsRNA viruses is 26.4% while unassigned or unannotated agents accounted for a cumulated 60.9%. Whether this difference is a consequence of differences in the methodology used or actually reflects differences in the

analyzed viromes cannot easily be ascertained. However, some frequent viruses of tomato such as PVY, CMV or STV were detected in both studies (Xu et al., 2017).

Despite the use of complex plant pools composed of 100 individual plants, we were able to assemble long, high quality contigs for some viruses (PVY, BBWV1, STV, and the new SnIV1), covering a very high proportion of the genome of these agents. In a few cases, such long contigs could not be assembled, possibly as a consequence of too low coverage and read numbers, or because mixed infection involving closely related variants created problems during contig assembly. Indeed, there is some evidence that at least one additional clade of PVY existed in some tomato samples as judged by the detection of some partial contigs diverging from the fully assembled genomes (data not shown).

For read mapping, stringent parameters were used so that there is no or extremely limited cross talk between isolates of different clades, as seen by reads numbers in the case of BBWV1 and PVY (Table S2). At the same time, it is difficult to know precisely how to interpret the samples with a very low number of reads mapped. Even if the background observed in the negative control was substracted, this cross-talk background likely due to index hoping (Illumina, 2017; van der Valk et al., 2019) may not be completely uniform from sample to sample. These low read numbers may therefore either reflect an absence of the virus but a low, slightly uneven cross-talk with other samples or a true, very low prevalence of the virus in the sampled population. It is not possible to decide between these two options here.

A very large and unexpected BBWV1 diversity was identified in the sampled nightshade populations. The analysis of BBWV1 populations suggests the existence of frequent reassortment

between RNA1 and RNA2 variants, an observation in line with the results of (Ferriol et al., 2014). BBWV1 is a Fabavirus with a relatively wide host range and which is pathogenic on a range of crops including broad bean, pea, spinach, lettuce, pepper and, occasionally, tomato (Blancard, 2012; Carpino et al., 2019; Taylor and Stubbs, 1972). It is therefore surprising that this aphid-transmitted virus was only detected from nightshade samples in this study. This observation suggest the existence of a biological or epidemiological barrier limiting the spread of BBWV1 from nightshade to tomato. In this respect, it is noteworthy that during a recent comparison of BBWV1 isolates, infection rates in tomato following artificial inoculations ranged only from 40% to 60% for four genetically different BBWV1 isolates (Carpino et al., 2019).

The novel ilarvirus here named Solanum nigrum ilarvirus 1 (SnIV1) was detected in both tomato and nightshade samples. However, both the prevalence and, with one exception, the read numbers of SnIV1 appear to be higher in the nightshade populations than in the tomato ones. On the other hand, the presence of SnIV1 in nightshade samples does not seem to be affected by whether they were growing side by side with tomato or not (respectively 3/6 and 3/5 cases, Table S2). Interestingly, reanalysis of metagenomics data showed that this virus was already present in 2011 at the TOM3 site,in *S. villosum* (hairy nightshade) a close relative of *S. nigrum*. Whether this novel ilarvirus is pathogenic to tomato or whether it has the potential to emerge at some point as a tomato pathogen in the same fashion as its close relatives Parietaria mottle virus (Roggero et al., 2000) and tomato necrotic spot virus (Batuman et al., 2011) remains to be evaluated.

The main PVY strains identified in this study were PVY-NTN and PVY-C1. PVY-C1 isolates were mainly detected from tomato, with one isolate shared between tomato and nightshade in the TOM3 site (Figure 4). On the other hand, PVY-NTN isolates were found in both tomato (5/7

samples) and nightshade samples (6/11 samples) from a total of seven of the 12 sampling sites. Interestingly, PVY populations at the TOM3 site had been studied 2011-2012 using specific RT-PCR assays (Moury et al., 2017). At the time, PVY-C1 and recombinant isolates likely to represent PVY-NTN were detected in tomato, while a more diverse population involving PVY-O, PVY-NTN, PVY-N and PVY-C1 was detected in nightshade and in the related *S. villosum* (Moury et al., 2017). The results reported here therefore suggest a simplification of the PVY nightshade population at that site, with the loss of PVY-O and PVY-N, possibly as a consequence of the competition with PVY-NTN and C1.

A noteworthy observation concerns PVY prevalence in nightshade populations at tomato sites (4/6 sites, 66.6%) and at non-tomato sites (1/5 sites, 20%). This suggests that infection in nightshade is greatly increased by the presence of tomato, reflecting a likely spillover effect from tomato crops to the wild nightshade population (Power and Mitchell, 2004).

Taken together the results reported here provide evidence for viral exchanges between tomato and nightshade populations growing side by side (such as the extremely closely related tomato and nightshade PVY isolates shared at the TOM3 site or the low detection of the new ilarvirus in tomato only at sites were it is also present in nightshade). At the same time, our results also highlight situations where an expected transfer is not observed, likely as a consequence of unforeseen biological or ecological barriers. This concerns in particular BBWV1 only found in nightshade when there are numerous indications that this virus should be able to infect tomato (Carpino et al., 2019). These results also highlight the power of metagenomics to analyze viral exchanges in complex plant populations, from the overall virome structure down to the intra-specific variability level, revealing unknown novel agents but also unforeseen biological processes.

ACKNOWLEDGMENTS

The authors would like to thank Eric Sclaunich (INVENIO, Sainte Livrade sur Lot, France) for assistance in accessing some of the tomato and nightshade sampling sites and the Genotoul Platform (INRA, Toulouse, France) for the Illumina sequencing. We would like to thank Laurence Svanella-Dumas and other colleagues in the INRA virology team for help in sample processing. Yuxin Ma was supported by a China Scholarship Council PhD grant.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.

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 Ecol. Evol. 19, 535-544.

Bateman, A., Smart, A., Luciani, A., Salazar, G.A., Mistry, J., Richardson, L.J., Qureshi, M., El-Gebali, S., Potter, S.C., Finn, R.D., Eddy, S.R., Sonnhammer, E.L L., Piovesan, D., Paladin, L., Tosatto, S.C E., Hirsh, L., 2018. The Pfam protein families database in 2019. Nucleic Acids Res. 47, D427-D432.

Batuman, O., Chen, L., Gilbertson, R., 2011. Characterization of Tomato necrotic spot virus (ToNSV), a new ilarvirus species infecting processing tomatoes in the Central Valley of California, Phytopathology 103, 1391-1396.

Blancard, D., 2012. Tomato diseases: identification, biology and control: A Colour Handbook. CRC Press, USA.

Brunt, A., 1996. Plant Viruses Online: Descriptions and Lists from the VIDE Database. Ver. 20. http://biology.anu.edu.au/Groups/MES/vide/.

Carpino, C., Elvira-González, L., Rubio, L., Peri, E., Davino, S., Galipienso, L., 2019. A comparative study of viral infectivity, accumulation and symptoms induced by broad bean wilt virus 1 isolates. J. Plant Pathol. 101, 275-281.

Elena, S.F., Fraile, A., Garcia-Arenal, F., 2014. Evolution and emergence of plant viruses. Adv. Virus Res. 88, 161-191.

Ferriol, I., Ferrer, R.M., Luis-Arteaga, M., Guerri, J., Moreno, P., Rubio, L., 2014. Genetic variability and evolution of broad bean wilt virus 1: role of recombination, selection and gene flow. Arch. Virol. 159, 779-784.

García-Andrés, S., Monci, F., Navas-Castillo, J., Moriones, E., 2006. Begomovirus genetic diversity in the native plant reservoir Solanum nigrum: evidence for the presence of a new virus species of recombinant nature. Virology 350, 433-442.

Hanssen, I.M., Lapidot, M., Thomma, B.P., 2010. Emerging viral diseases of tomato crops. Mol. Plant. Microbe Interact. 23, 539-548.

Holm, L., Pancho, J.V., Herberger, J.P., Plucknett, D.L., 1979. A geographical atlas of world weeds. John Wiley and Sons, New York, USA.

Huerta-Cepas, J., Bork, P., Serra, F., 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol. Biol. Evol. 33, 1635-1638.

Illumina, 2017. Effects of index misassignment on multiplexing and downstream analysis. https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf.

Jacquemond, M., Lot, H., 1981. L'ARN satellite du virus de la mosaïque du concombre I. - Comparaison de l'aptitude à induire la nécrose de la tomate d'ARN satellites isolés de plusieurs souches du virus. Agronomie 1, 927-932.

Jamuna, S., Rajendran, L., Haokip, B.D., Nagendran, K., Karthikeyan, G., Manoranjitham, S.K., 2017. First Report of Natural Infection of Solanum nigrum with Tomato mosaic virus in India. Plant Dis. 102, 1044-1044.

Javed, T., Ashfaq, U.A., Riaz, S., Rehman, S., Riazuddin, S., 2011. In-vitro antiviral activity of Solanum nigrum against Hepatitis C Virus. Virol J. 8, 26-26.

Kumar, S., Nei, M., Dudley, J., Tamura, K., 2008. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Briefings in Bioinformatics 9, 299-306.

Lefebvre, M., Theil, S., Ma, Y., Candresse, T., 2019. The VirAnnot pipeline: a resource for automated viral diversity estimation and operational taxonomy units (OTU) assignation for virome sequencing data. Phytobiomes Journal, online first, DOI:10.1094/PBIOMES-07-19-0037-A.

Luria, N., Smith, E., Reingold, V., Bekelman, I., Lapidot, M., Levin, I., Elad, N., Tam, Y., Sela, N., Abu-Ras, A., 2017. A new Israeli Tobamovirus isolate infects tomato plants harboring Tm-22 resistance genes. PLoS One 12, e0170429.

Marais, A., Faure, C., Bergey, B., Candresse, T., 2018. Viral Double-Stranded RNAs (dsRNAs) from Plants: Alternative Nucleic Acid Substrates for High-Throughput Sequencing. Methods Mol. Biol. 1746, 45-53.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17, 10-12.

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

McLeish, M.J., Fraile, A., García-Arenal, F., 2019. Evolution of plant–virus interactions: host range and virus emergence. Curr. Opin. Virol. 34, 50-55.

Moury, B., Simon, V., Faure, C., Svanella-Dumas, L., Marais, A., Candresse, T., 2017. Host groups of Potato virus Y: Vanishing barriers. In Potato virus Y: biodiversity, pathogenicity, epidemiology and management. C. Lacomme, L. Glais, D.U. Bellstedt, B. Dupuis, A.V. Karasev &, E. Jacquot, Eds. dir., *Potato virus Y: biodiversity, pathogenicity, epidemiology and management* (p. 243-261). Springer, Cham. pp. 243-261.

Murtagh, F., Legendre, P., 2014. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J Classif 31, 274-295.

Power, Alison G., Mitchell, Charles E., 2004. Pathogen Spillover in Disease Epidemics. The American Naturalist 164, S79-S89.

Roggero, P., Ciuffo, M., Katis, N., Alioto, D., Crescenzi, A., Parrella, G., Gallitelli, D., 2000. Necrotic disease in tomatoes in Greece and southern Italy caused by the tomato strain of Parietaria mottle virus. J. Plant Pathol. 82, 159.

Roossinck, M.J., Martin, D.P., Roumagnac, P., 2015. Plant Virus Metagenomics: Advances in Virus Discovery. Phytopathology 105, 716-727.

Sabanadzovic, S., Valverde, R.A., Brown, J.K., Martin, R.R., Tzanetakis, I.E., 2009. Southern tomato virus: The link between the families Totiviridae and Partitiviridae. Virus Res. 140, 130-137.

Simmonds, P., 2015. Methods for virus classification and the challenge of incorporating metagenomic sequence data. J Gen Virol 96, 1193-1206.

Taylor, R.H., Stubbs, L.L., 1972. Broad bean wilt virus 1.CMI-AAB Description of plant viruses, 81, http://www.dpvweb.net/dpv/showadpv.php?dpvno=81

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673-4680.

van der Valk, T., Vezzi, F., Ormestad, M., Dalén, L., Guschanski, K., 2019. Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. Mol. Ecol. Resour. 2019; 00:1-11. https://doi.org/10.1111/1755-0998.13009.

van der Vlugt, R.A., Verbeek, M., Dullemans, A.M., Wintermantel, W.M., Cuellar, W.J., Fox, A., Thompson, J.R., 2015. Torradoviruses. Annu. Rev. Phytopathol. 53, 485-512.

Verbeek, M., Dullemans, A., van den Heuvel, H., Maris, P., van der Vlugt, R., 2010. Tomato chocolate virus: a new plant virus infecting tomato and a proposed member of the genus Torradovirus. Arch. Virol. 155, 751-755.

Verbeek, M., Dullemans, A.M., van den Heuvel, J.F., Maris, P.C., van der Vlugt, R.A., 2008. Tomato marchitez virus, a new plant picorna-like virus from tomato related to tomato torrado virus. Arch. Virol. 153, 127-134.

Verdin, E., Gognalons, P., Wipf-Scheibel, C., Bornard, I., Ridray, G., Schoen, L., Lecoq, H., 2009. First report of Tomato torrado virus in tomato crops in France. Plant Dis. 93, 1352-1352.

Villamor, D.E.V., Ho, T., Al Rwahnih, M., Martin, R.R., Tzanetakis, I.E., 2019. High Throughput Sequencing For Plant Virus Detection and Discovery. Phytopathology 109, 716-725.

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, e80.

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. J. Virol. 91, e00173-00117.

LEGENDS TO THE FIGURES

Figure 1. Barplot illustrating the presence/absence data based on Blast annotation for identified viral families in each sampled plant population.

Figure 2. Neighbor-joining trees reconstructed from the alignment of near complete nucleotide sequences of RNA1 (A) and RNA2 (B) of broad bean wilt virus 1 (BBWV1) isolates and other *Fabavirus* members. Statistical significance of the branches was evaluated by bootstrap analysis (1,000 replicates) and only bootstrap values higher than 70% are indicated. The scale bars represent 5% nucleotide divergence. Sequences of BBWV1 determined in this work are indicated by a black diamond. The abbreviations followed by the accession numbers are: BBWV2: broad bean wilt virus 2; GeMV: gentian mosaic virus; LLMV: Lamium mild mosaic virus; PeLaV: peach leaf pitting-associated virus; PrVF: Prunus virus F; ChVF: cherry virus F; GFabV: grapevine fabavirus.

Figure 3. Neighbor-joining trees reconstructed from the alignment of amino acid sequences of the P1 protein (A) and coat protein (B) of representative members of the genus *Ilarvirus*. Statistical significance of branches was evaluated by bootstrap analysis (1,000 replicates) and only bootstrap values higher than 70% are indicated. The scale bars represent 10% amino acid divergence (A) and 5% amino acid divergence (B). Solanum nigrum ilarvirus 1 (SnIV1) characterized in this study is indicated by a black diamond.

Figure 4. Phylogenetic analysis of the near complete nucleotide genome sequences of potato virus Y (PVY) isolates determined in this study (indicated by black diamonds) and reference sequences. PVY isolates from tomato samples are colored in green and those from nightshade samples in orange. The tree was constructed by the neighbor-joining method from strict nucleotide identity distances and the statistical significance of branches was evaluated by bootstrap analysis (1,000 replicates). Only bootstrap values higher than 70% are indicated. The scale bar represents 5% nucleotide divergence.

Figure 1

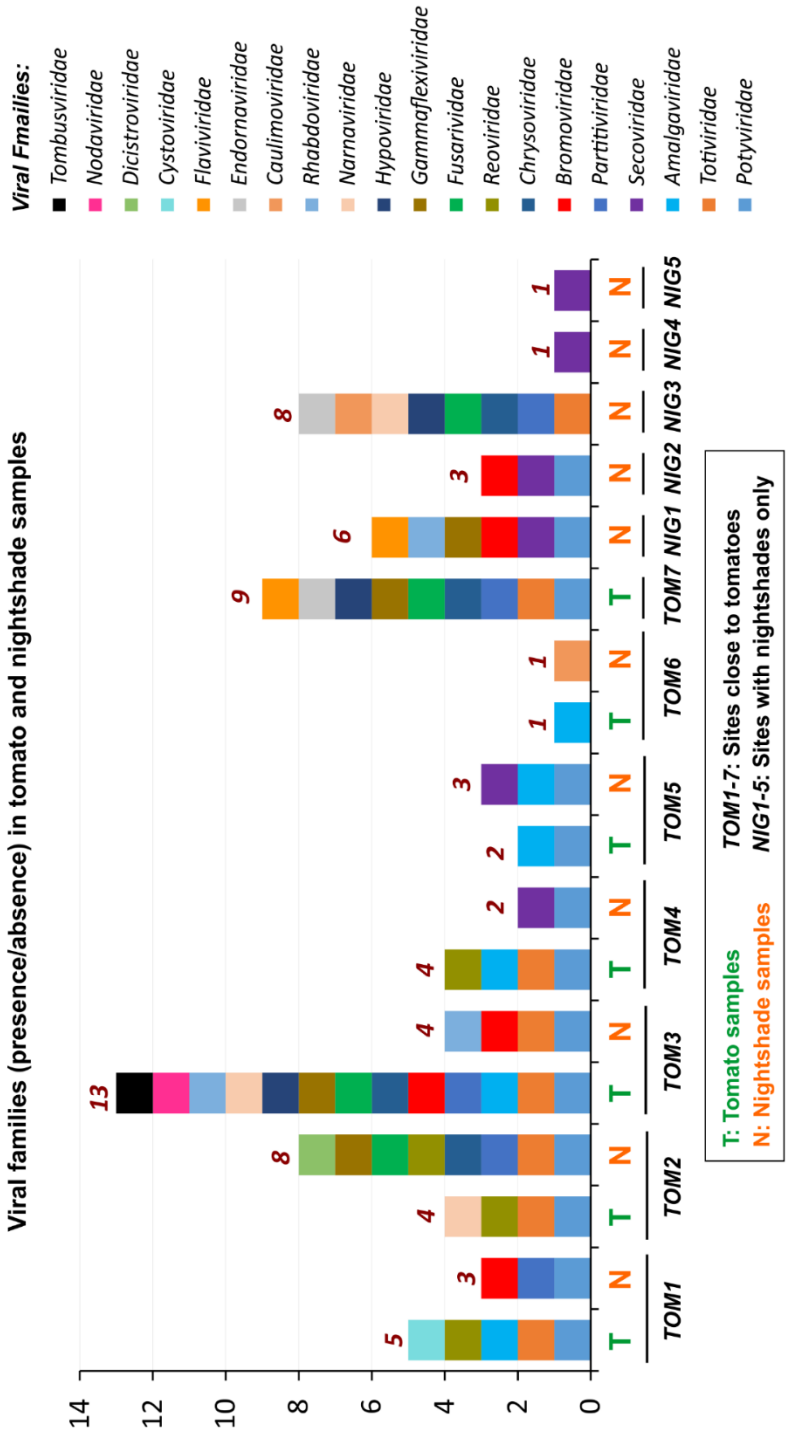

Viral Fmailies:
Tombusviridae
Nodaviridae
Dicistroviridae
Cystoviridae
Flaviviridae
Endornaviridae
Caulimoviridae
Rhabdoviridae
Narnaviridae
Hypoviridae
Gammaflexiviridae
Fusariviridae
Reoviridae
Chrysoviridae
Bromoviridae
Partitiviridae
Secoviridae
Amalgaviridae
Totiviridae
Potyviridae
Viral families (presence/absence) in tomato and nightshade samples
0
2
4
6
8
10
12
14
T: Tomato samples
N: Nightshade samples
TOM1-7: Sites close to tomatoes
NIG1-5: Sites with nightshades only
TOM1
TOM2
TOM3
TOM4
TOM5
TOM6
TOM7
NIG1
NIG2
NIG3
NIG4
NIG5

Figure 2.


A RNA1 (nt)
BBWV1 TOM5-N-A RNA1
BBWV1 NIG4-A RNA1
BBWV1 NIG1-A RNA1
BBWV1 NIG2 RNA1
BBWV1 NIG5-E RNA1
BBWV1 NIG1-B RNA1
BBWV1 RNA1 NC 005289
BBWV1 NIG1-C RNA1
BBWV1 RNA1 MH645161
BBWV1 RNA1 MF770979
BBWV1 RNA1 KT988973
BBWV1 RNA1 AY781171
BBWV1 NIG4-D RNA1
BBWV1 NIG1-D RNA1
BBWV1 TOM4-N-D RNA1
BBWV2 RNA1 NC 003003
GeMV RNA1 NC 011190
LLMV RNA1 NC 023016
PeLaV RNA1 KY867750
PrVF RNA1 NC 039077
ChVF RNA1 MH998212
GFabV RNA1 NC 039073
0.05
B RNA2 (nt)
BBWV1 NIG1-H RNA2
BBWV1 TOM4-N-H RNA2
BBWV1 NIG4-H RNA2
BBWV1 NIG5-G RNA2
BBWV1 RNA2 AY781172
BBWV1 RNA2 KT988974
BBWV1 NIG2-F RNA2
BBWV1 RNA2 NC 005290
BBWV1 RNA2 AB110537
BBWV1 RNA2 MF770980
BBWV1 RNA2 MH645162
BBWV1 RNA2 AF225955
BBWV1 RNA2 AB018703
BBWV2 RNA2 NC 003004
GaMV RNA2 NC 011189
LLMV RNA2 NC 023017
GFabV RNA2 NC 039072
PeLaV RNA2 KY867751
PrVF RNA2 NC 039078
ChVF RNA2 MH998220
0.05

Figure 3.

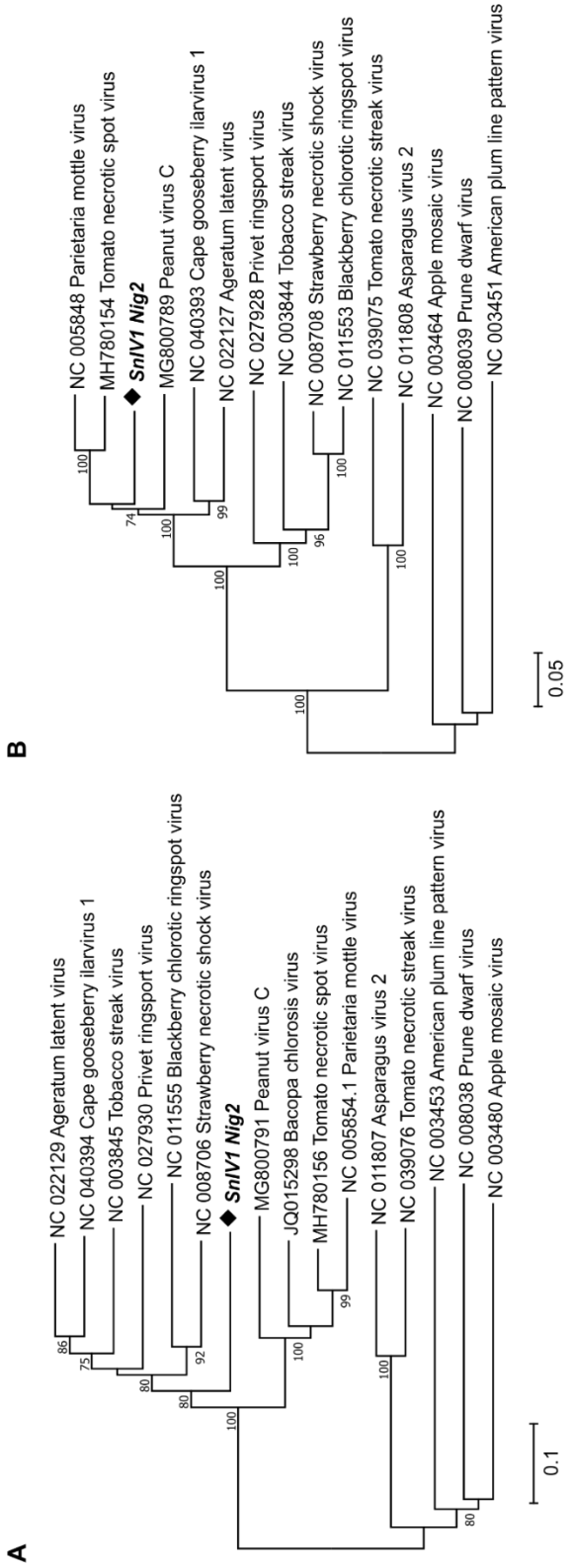
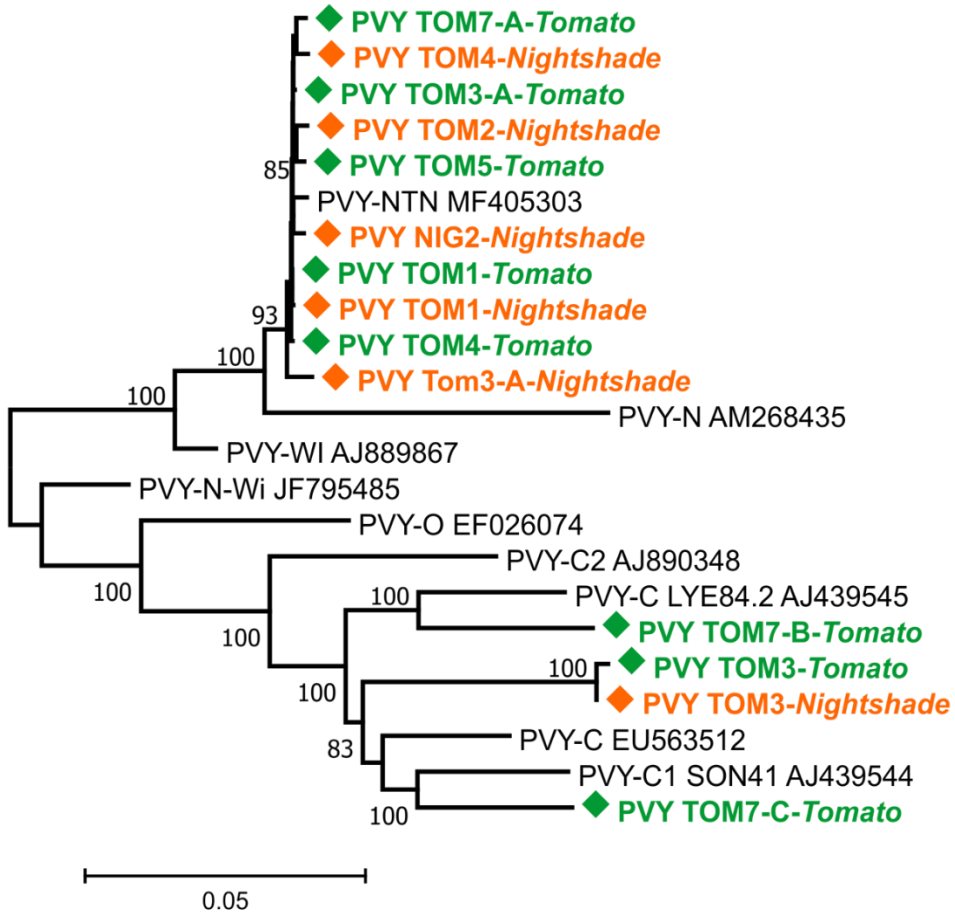

A
NC 022129 Ageratum latent virus
NC 040394 Cape gooseberry ilarvirus 1
NC 003845 Tobacco streak virus
NC 027930 Privet ringsport virus
NC 011555 Blackberry chlorotic ringspot virus
NC 008706 Strawberry necrotic shock virus
◆ SnlV1 Nig2
MG800791 Peanut virus C
JQ015298 Bacopa chlorosis virus
MH780156 Tomato necrotic spot virus
NC 005854.1 Parietaria mottle virus
NC 011807 Asparagus virus 2
NC 039076 Tomato necrotic streak virus
NC 003453 American plum line pattern virus
NC 008038 Prune dwarf virus
NC 003480 Apple mosaic virus
0.1
B
NC 005848 Parietaria mottle virus
MH780154 Tomato necrotic spot virus
◆ SnlV1 Nig2
MG800789 Peanut virus C
NC 040393 Cape gooseberry ilarvirus 1
NC 022127 Ageratum latent virus
NC 027928 Privet ringsport virus
NC 003844 Tobacco streak virus
NC 008708 Strawberry necrotic shock virus
NC 011553 Blackberry chlorotic ringspot virus
NC 039075 Tomato necrotic streak virus
NC 011808 Asparagus virus 2
NC 003464 Apple mosaic virus
NC 008039 Prune dwarf virus
NC 003451 American plum line pattern virus
0.05

Figure 4.


PVY TOM7-A-Tomato
PVY TOM4-Nightshade
PVY TOM3-A-Tomato
PVY TOM2-Nightshade
PVY TOM5-Tomato
PVY-NTN MF405303
PVY NIG2-Nightshade
PVY TOM1-Tomato
PVY TOM1-Nightshade
PVY TOM4-Tomato
PVY Tom3-A-Nightshade
PVY-N AM268435
PVY-WI AJ889867
PVY-N-Wi JF795485
PVY-O EF026074
PVY-C2 AJ890348
PVY-C LYE84.2 AJ439545
PVY TOM7-B-Tomato
PVY TOM3-Tomato
PVY TOM3-Nightshade
PVY-C EU563512
PVY-C1 SON41 AJ439544
PVY TOM7-C-Tomato
85
93
100
100
100
100
100
100
100
83
100
0.05

Supplementary Materials [Supplementary Tables are available at https://doi.org/10.15454/RWOLLQ, Table S4, which fits easily in an A4 format is also provided here]

Table S1. OTUs with their annotation and the number of corresponding reads in each library.

Table S2. Virus isolates and the GenBank accession numbers of their nucleotide sequences.

Table S3. Number of reads mapped on representative contigs of the genomic RNAs of cucumber mosaic virus (CMV), southern tomato virus (STV), potato virus Y (PVY), bean broad wilt virus 1 (BBWV1) and the novel ilarvirus Solanum nigrum ilarvirus 1 (SnIV1).

Table S4. Amino-acid and nucleotide identity levels between the genomic RNAs, open reading frames and proteins of Solanum nigrum ilarvirus 1 and Parietaria mottle virus.

		Nucleotide identity (%)	Amino acid identity (%)
RNA1	Genomic RNA1	73.2%	na
	P1 ORF/P1 Protein	74.3%	81.8%
RNA2	Genomic RNA2	69.3%	na
	P2 ORF/P2 protein	69.7%	72.6%
	P2b ORF/P2b protein	64.6%	61.6%
RNA3	Genomic RNA3	58.6%	na
	MP ORF/MP protein	58.6%	58.8%
	CP ORF/CP protein	61.6%	53.9%

LEGENDS TO SUPPLEMENTARY FIGURES

Figure S1. Virome cross-talk at OTU level between samples. The sample/library and identified number of OTUs are indicated at bottom-left; the interactions between different viromes were shown in the matrix layout at the bottom-right, the aggregates based on the groupings and the corresponding numbers of OTUs were plotted and shown in the upper part.

Figure S2. Neighbor-joining trees reconstructed from the alignment of amino acid sequences of the P2a, P2b and movement (MP) proteins of representative members of the genus *Ilarvirus*. Statistical significance of branches was evaluated by bootstrap analysis (1,000 replicates). Only bootstrap values higher than 70% are indicated. The scale bars represent 5% (P2a) or 10% (P2b and MP) amino acid divergence. Solanum nigrum ilarvirus 1 (SnIV1) characterized in this study is indicated by a black diamond.

Figure S1.

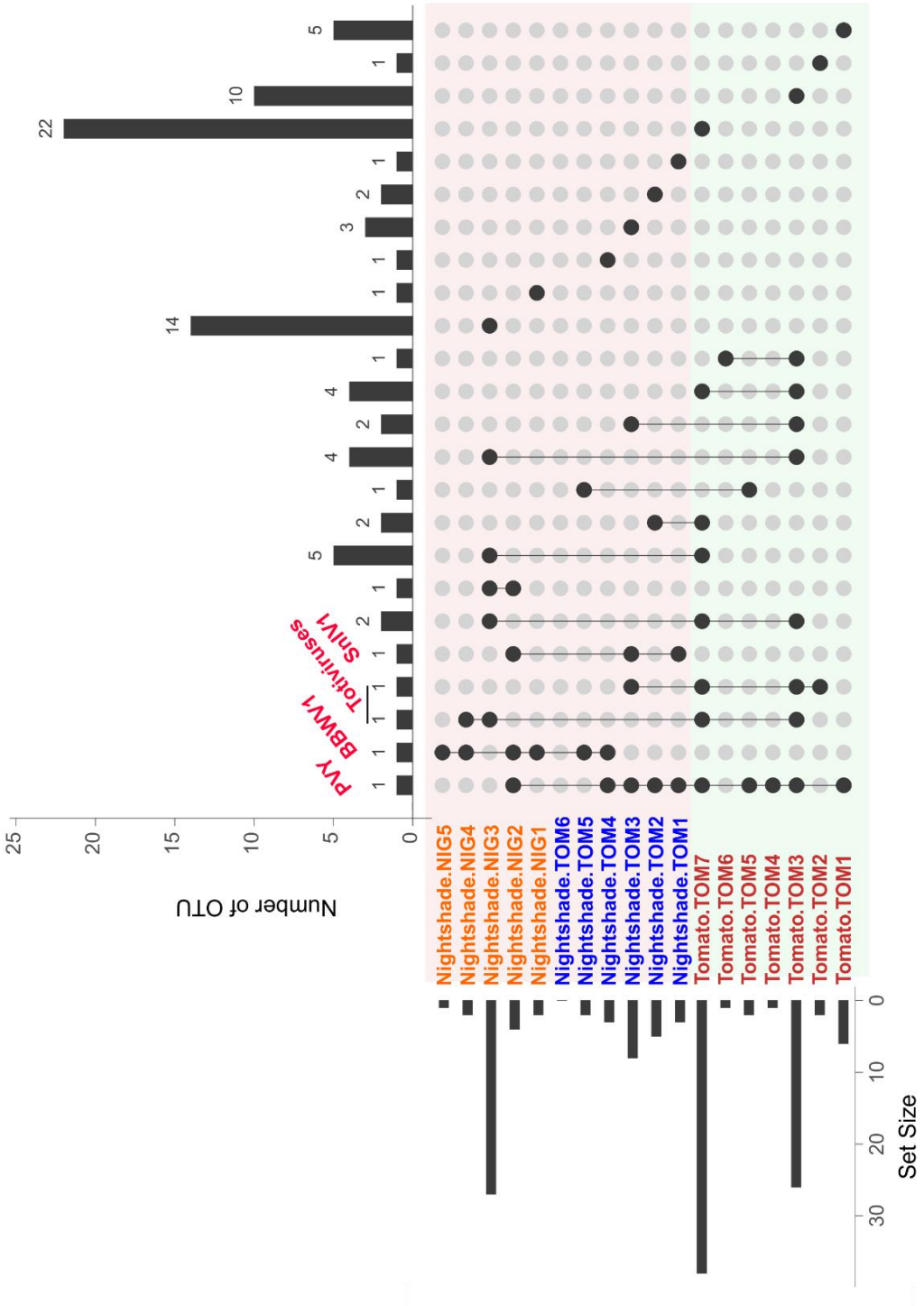
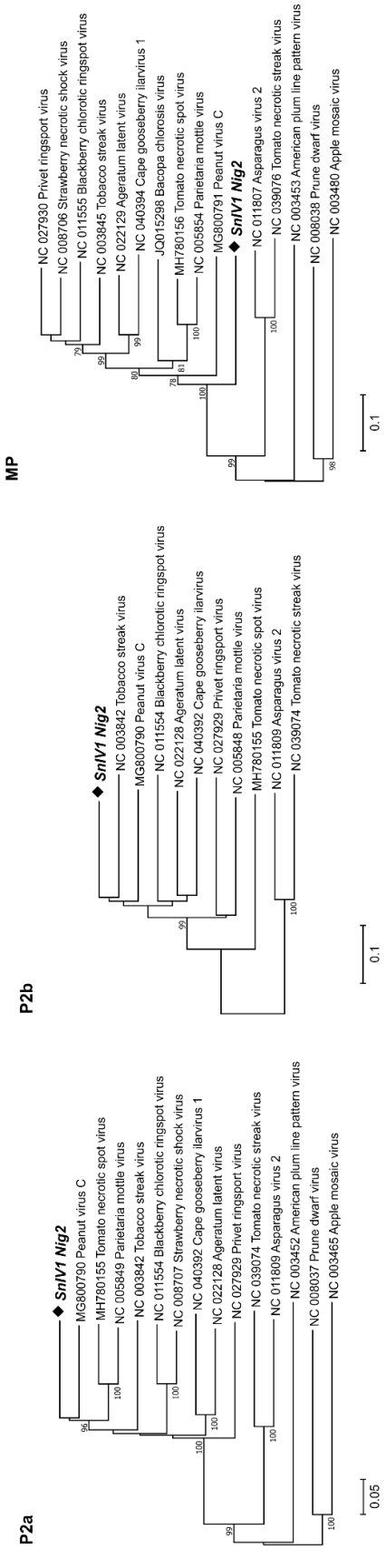

Number of OTU
0
5
10
15
20
25
PVY
BBWV1
Totiviruses
SnIV1
Nightshade.NIG5
Nightshade.NIG4
Nightshade.NIG3
Nightshade.NIG2
Nightshade.NIG1
Nightshade.TOM6
Nightshade.TOM5
Nightshade.TOM4
Nightshade.TOM3
Nightshade.TOM2
Nightshade.TOM1
Tomato.TOM7
Tomato.TOM6
Tomato.TOM5
Tomato.TOM4
Tomato.TOM3
Tomato.TOM2
Tomato.TOM1
Set Size
0
10
20
30

Figure S2.


MP
NC 027930 Privet ringsport virus
NC 008706 Strawberry necrotic shock virus
NC 011555 Blackberry chlorotic ringspot virus
NC 003845 Tobacco streak virus
NC 022129 Ageratum latent virus
NC 040394 Cape gooseberry ilarvirus 1
JQ015298 Bacopa chlorosis virus
MH780156 Tomato necrotic spot virus
NC 005854 Parietaria mottle virus
MG800791 Peanut virus C
SnIV1 Nig2
NC 011807 Asparagus virus 2
NC 039076 Tomato necrotic streak virus
NC 003453 American plum line pattern virus
NC 008038 Prune dwarf virus
NC 003480 Apple mosaic virus
0.1
P2b
SnIV1 Nig2
NC 003842 Tobacco streak virus
MG800790 Peanut virus C
NC 011554 Blackberry chlorotic ringspot virus
NC 022128 Ageratum latent virus
NC 040392 Cape gooseberry ilarvirus
NC 027929 Privet ringsport virus
NC 005848 Parietaria mottle virus
MH780155 Tomato necrotic spot virus
NC 011809 Asparagus virus 2
NC 039074 Tomato necrotic streak virus
0.1
P2a
SnIV1 Nig2
MG800790 Peanut virus C
MH780155 Tomato necrotic spot virus
NC 005849 Parietaria mottle virus
NC 003842 Tobacco streak virus
NC 011554 Blackberry chlorotic ringspot virus
NC 008707 Strawberry necrotic shock virus
NC 040392 Cape gooseberry ilarvirus 1
NC 022128 Ageratum latent virus
NC 027929 Privet ringsport virus
NC 039074 Tomato necrotic streak virus
NC 011809 Asparagus virus 2
NC 003452 American plum line pattern virus
NC 008037 Prune dwarf virus
NC 003465 Apple mosaic virus
0.05

DISCUSSION AND PERSPECTIVES

DISCUSSION AND PERSPECTIVES

Through my thesis, the main objective was simply to explore the diversity of plant viruses using high throughput sequencing based metagenomics, to understand the prevalence and the dynamics of viruses in space and time and further to begin to discover the potential biotic and abiotic drivers shaping phytovirome compositions under various conditions. I used HTS-based metagenomics as a key technique through a series of experiments addressing current questions in plant phytovirome studies. I also applied a metabarcoding approach in one chapter of this thesis for the description of leaf-associated mycobiomes in various ecosystems. The most significant results were obtained on the following two aspects: first, I explored the impacts of different methods on virome or mycobiome description and for the interpretation of the biological and ecological importance of the results obtained. I systematically investigated and compared the richness and compositions of a number of phytoviromes (and a few mycobiomes or mycoviromes), using either monospecific or plurispecific pools of cultivated or wild/weeds plants from cultivated or unmanaged environments, allowing me to improve our understanding of plant virus diversity, to provide a rich virus database for future studies and to begin to unravel parameters influencing in space and time virome composition and richness.

Key methodological aspects

1. OTUs, an applicable proxy to ICTV species for virome composition and richness estimation!

The Operational Taxonomic Unit (OTU) classification implemented with the virAnnot pipeline (Lefebvre *et al*., submitted for publication) was applied through almost all the virus diversity analyses in this thesis for the richness assessment (Chapter I and Chapter IV), and also for beta diversity analysis (Chapter II and Chapter III). As we know the relatively short reads obtained by second generation HTS and the incomplete coverage often seen in metavirome studies bring assembly and annotation difficulties that can strongly impair diversity assessment and comparisons (Simmonds, 2015). Illuminated by the possible strategy proposed by Simmonds (2015) and others (Klingenberg *et al.*, 2013), a routine for automated OTU classification was developed, repeatedly debugged, validated, and integrated in virAnnot during this thesis. The workflow is freely available and explained at https://github.com/marieBvr/virAnnot. The short submitted Resource Announcement is included as an annex to the present thesis (Lefebvre *et al*., submitted for publication).

I would like to summarize a few characteristics of this OTU classification scheme, mainly for RNA dependent RNA polymerase (RdRp)-OTUs:

1) it is not an extremely fine net since instead of capturing all viral sequences, only contigs (or reads) containing viral conserved domains (ex. RdRp, CP, Helicase…) are captured. In addition, to avoid richness overestimation, we generally concentrate efforts on count sequences encoding an RdRp signature, which eliminates viruses for which there is no coverage of genome region. The richness estimation obtained is thus clearly a lower bound value, but one that can be reproducibly calculated in multiple samples, allowing direct, meaningful comparisons for beta diversity analyses among different communities just like the analysis used for bacteria and fungi diversity. In fact, our results show that there is more variability between independent WGA reactions from the same nucleic acids extract than between repeated clustering analyses of resampled data.
2) It is however possible to also consider other conserved viral signatures. In particular it is possible to analyse and develop OTUs based on conserved signatures specific of DNA viruses to compensate for the fact that they are by nature excluded by an RdRp-based approach.
3) A 10% nt or aa divergence on a RdRp alignable region with a minimum overlap of 20 aa has been shown to represent a decent threshold to cluster related sequences into OTUs mimicking taxonomic species in several plant RNA virus families. Ability to use OTUs as a proxy to ICTV species level, has been examined and statistically validated by comparing the distance characteristics between OTUs and between *bona fide* species (Annex C). It is however conceivable to refine the clustering routine, for example by further refining the clustering threshold or even by adopting different thresholds for different viral families in order to reflect the existing variability in taxonomic criteria between viral taxa.
4) While OTU-counting provides, as described above, a lower bound estimate of viral richness, it should also be considered that this approach might also allow the detection of highly divergent agents that are not picked up by Blast-based approaches because no close relative to these agents currently exist in databases. Indeed although the results are not included here, I have been able to identify in this way such a highly divergent agent in the virome of wild radish (*Raphanus raphanistrum*).
5) It should be stressed that beta diversity analyses and compositional dissimilarity analyses can currently only be based on presence/absence data for each identified OTU. Given the biases potentially introduced by viral biology (variable amounts of dsRNA produced by

different viral families) and by the WGA amplification process, it is unlikely that number of reads for each OTU reflects viral abundance in the sampled plant population.

6) Even with its limitations, the virAnnot strategy is likely to provide a viral diversity description closer to species level than the recently released GRAViTy pipeline (Genome Relationships Applied to Virus Taxonomy) which can assign eukaryotic viruses (especially metagenomic sequences) into viral families and orders (Aiewsakun and Simmonds, 2018).

2. Metagenomics or ecogenomics?

As repeatedly mentioned in this thesis, when the objective is to describe the virome of complex plant populations comprising different plant species and many individual plants, the two strategies most often used are the so-called "metagenomics" and "ecogenomics" approaches as defined by Roossinck (2012). The major difference is that the latter allows each sequence to be traced back to the specific geographical location and/or original host (Bernardo *et al.*, 2018; Roossinck *et al.*, 2015). However, the "metagenomics" strategy can be modified, as in Chapter III for mycobiomes, by sorting the samples by species. It is also possible to reanalyze metagenomics samples by (RT)PCR of species-specific pools, thus allowing to identify the host(s) of viruses of interest. Indeed, such an approach was used to identify wild radish as the host of the highly divergent virus mentioned above. During the present PhD, I used both "Metagenomics" and "Ecogenomics" approaches, respectively in Chapters 2 and 3 and in Chapters 1, 4.

Given the very different structure of the analyzed samples, it is not possible to determine if one of these approaches provides a higher resolution of phytoviromes at the OTU level. For a comparable number of individual plants analyzed, the amount of hands-on time is not really linked to the use of one strategy or the other but to the size of the pools analyzed. As shown in the Annex A to Chapter III, the use of larger pools probably increases the competition between individual viruses and may result is less precise and, to some extent, less reproducible results. In designing future experiments, these elements will need to be taken into consideration in order to strike a balance between hands-on wet lab time, sequencing costs and virome resolution. One possible strategy, could be to purify dsRNAs or VANA from smaller pools of plants and perform separate WGA amplifications but to then assemble the amplification products into larger homogenous ensembles (species- or site-specific) in order to limit the number of sequencing libraries produced so as to limit the associated costs.

In parallel, the results obtained here also showed a direct increase in overall virome size (aggregated number of unique OTUs) with the number of analyzed plants over time (chapter I).

Yet, in Chapter IV relatively simple viromes were obtained for most monospecific pools analyzed, despite the fact that these pools contained 100 individual plants. These seemingly contradictory observations in fact raise the question of the stability of the virome in space and time. A very dynamic virome was observed in chapter I, suggesting high variability with time. However, as discussed in this chapter, we currently do not know if this observation stems from presence/absence changes in the sampled plant population or from changes in prevalence. Likewise, we do not currently know the physical scale(s) at which the virome may vary. A consequence is that it is not possible to determine the number and complexity of the samples to be analyzes in order to obtained a truly representative description of the virome associated with a plant population.

3. Double stranded RNAs (dsRNAs) or Virion-associated nucleic acids (VANA)?

HTS-based metagenomics studies of plant viruses generally use very complex nucleic acids mixtures that contain both hosts and viral nucleic acids. Two possible nucleic acids target populations allow an enrichment of viral sequences, double stranded RNAs (dsRNAs) and virion-associated nucleic acids (VANA). In Chapter II, I directly compared the performance of the HTS analysis of highly purified dsRNAs and of VANA for phytovirome description in 6 cultivated or unmanaged sampling sites. The results obtained show that the dsRNA-based approach consistently revealed a more comprehensive diversity for RNA viruses than the VANA approach, whatever the assessment criterion. In particular, the VANA approach was less efficient for the detection of viruses in the *Chrysoviridae*, *Reoviridae* and *Rhabdoviridae* families. This could be due to a low titer of these viruses in the analyzed samples or possibly, to less stable particles. Another interesting finding, is that *Endornaviridae*, which are ssRNA persistent viruses without a true capsids and particles but that produce host-derived vesicles (Dulieu *et al.*, 1988; Horiuchi *et al.*, 2001; Lefebvre *et al.*, 1990) containing their nucleic acids, were abundantly found from many libraries using the VANA approach. This finding confirm observations from other studies (Bernardo *et al.*, 2018; Maclot *et al.*, 2019) and suggests that the VANA approach is not limited to viral particles and virions.

Several hypotheses can be proposed for the better performance obtained using dsRNA purification. One could be that a greater enrichment of viral sequences is achieved during dsRNA purification. Indeed, a slightly higher proportion of viral reads was observed in Chapter II with dsRNA than with VANA (49.9% +/- 14.3% vs 40.5% +/- 16.6%). This could in turn lead to an ability to assemble longer, more efficiently annotated contigs. However, the difference in the proportion of

viral reads is not huge, which suggests that the difference in enrichment might not be a huge one and might therefore not explain completely the difference in overall performance. A second hypothesis could be a stronger competition between viruses during the WGA procedure, possibly coming higher imbalance between viruses in the purified nucleic acids preparations. In this hypothesis, highly concentrated and stable viruses would outcompete less stable and/or concentrated ones during the amplification of VANA targets, resulting in a less complete representation(Thurber *et al.*, 2009). A more detailed exploration of the crucial methodological steps in the VANA approach may be able to separate between these hypotheses, while the use of less complex pools and/or a greater sequencing depth may be able to improve the performance of the VANA approach.

Notably, the differential performance of dsRNA and VANA not only affect virome richness assessment but also the assessment of beta diversity between different communities for ecological significance interpretation. For example, the dissimilarity analysis based respectively on the dsRNA and VANA OTU data showed different hierarchical clustering and ordinations with IT-VANA close to VO-VANA and IT-ds close to INRA-ds (Chapter II, Figure 4), indicating that the extraction methodology exerts a critical importance for virome description. In the same way, it has been reported that the use of a particular assembler to obtain contigs may also be critical for human viromes characterization (Sutton *et al.*, 2019). A consequence of these elements, is that it may prove very difficult to compare at a fine scale viromes that have been obtained using different methodological approaches and that great care should be taken when attempting to perform such comparisons.

4. Different extraction kits and target amplicons for leaf-associated mycobiomes characterization?

Different with other chapters, in Chapter III of this PhD I combined virome and mycobiome analyses to uncover the diversity of viruses and leaf-associated fungi in both cultivated and unmanaged ecosystems in an attempt to better integrate the leaf microbiome studies (Laforest-Lapointe and Whitaker, 2019). From the methodological point of view, I compared the performance of two DNA extraction kits (Powersoil and DNeasy) and of two metabarcoding strategies (ITS1 and ITS2) for mycobiomes description of complex plant pools involving a range of plant species. The results of the direct comparison between these different tools (Annex B of Chapter III) demonstrated an equal performance between the two nucleic acids extraction kits used, but ITS1 barcoding proved more robust and allowed to detect a richer fungal diversity than the

ITS2 one. So far, despite of few experiments performed to date on complex plant pools, these results are consistent with most other comparisons. For instance, the study of Nilsson *et al.* (2008) has shown that the variability of ITS1 on average exceeds that of ITS2 and that ITS2 is more variable only for 34% of compared fungal species. The detailed evaluation of HTS studies of fungal communities and practical recommendations for aspects of sampling and laboratory practices to data processing and analysis have been recently reviewed by Nilsson *et al.* (2019).

5. Culturomics or direct-metagenomics for mycobiome and virome studies?

A culture-dependent method was used in Chapter III to address the long-standing question of whether the rich dsRNA viruses diversity found in plant-associated viromes (*Endornaviridae* now moved to ssRNA viruses, *Partitivirdae*, *Totiviridae*, *Chrysoviridae* and *Amalgaviridae*) corresponds to phytoviruses or to mycoviruses.

Though culturomics has emerged recently as a successful tool to isolate high number of bacteria and to identify new species for human gut microbiota (Lagier *et al.*, 2015; Lagier *et al.*, 2016; Lagier *et al.*, 2012), Plant microbiome culturomics is substantially lagging behind the human microbiome (Sarhan *et al.*, 2019). And its application on fungi culturing are still limited in this study and some other studies (Hamad *et al.*, 2017). Here, I successfully cultured between 480 and 1279 colonies from 200 individual plant leaf tissues (representing 40 plant species) in each of the four sites (between 2.4 and 6.4 colonies per plant fragment) using a dilution culture strategy (Unterseher and Schnittler, 2009). In a large experiments on fungal endophytes 1110 axenic endophyte cultures were obtained from 810 *Bauhinia brevipes* (Fabaceae) leaf fragments (1.37 colonies per fragment; (Hilarino *et al.*, 2011). Yet, despite the efforts involved only a small fraction (5.3% to 12.7%) of fungal ASVs identified in the plant samples were recovered a mycelia cultures, a result in line with the general recognition that *in vitro* culture-based approaches grossly underestimate fungal diversity (Roossinck, 2015). Compared with ITS-amplicons, culturomics indeed lost the great majority of fungi. However, it also enabled to culture a significant proportion of fungi (17.9% to 46.2% of the cultured ASVs) which were not recovered by the barcoding approach (Chapter III, supplementary Figure S2). A similar picture was also found from human mycobiome studies (Hamad *et al.*, 2017). With a significant fungal populations solely discovered from culturomics, this study suggested that the two approaches may in fact be complementary, especially if the objective is to correlate fungal community composition with the health state of the host (Hamad *et al.*, 2017). Combining these two approaches was indeed also suggested by Nowrotek *et al.* (2019) when studying the environmental resistome.

A much more extreme situation was observed when comparing the mycoviromes derived from mycelial cultures with the viromes derived from the plant samples, with only an extremely limited number of shared OTUs (from 0 to 2 OTUs, depending on the sampling site). Remarkably, the mycovirome revealed a wide diversity of known mycovirus families (Marzano *et al.*, 2016) that were not detected in the phytoviromes, such as *Gammaflexiviridae*, *Hypoviridae*, *Narnaviridae*, *Fusariviridae and Birnaviridae*. Due to the relative low fraction of cultured fungi, it is not simple to draw an explicit conclusion of the origins of the dsRNA viruses detected in phytoviromes. In particular, if the majority of these viruses is interpreted to be mycoviruses, it is not simple to understand why these were detected but not the other mycoviral families easily detected in the mycobiome. The recent striking findings that cross-kingdom viral infections can occurr in natural or experimental condition (Andika *et al.*, 2017; Nerva *et al.*, 2017), and that cross-family horizontal gene transfer occurred among these dsRNA viruses (Liu *et al.*, 2012) are making this question more complicated and mysterious.

Key findings of ecological relevance

1. The dynamic nature of plant-associated viromes

One of the findings reported here is that while general structure of plant-associated viromes appears to be relatively stable over time, for example when it comes to the presence and proportion of viruses with different genome types (ssRNA or dsRNA viruses) or the viral families discovered at a given sampling point, the viromes appear highly dynamic over time (Chapter I, Figure 4). In line with this result, the viromes for the Villenave d'Ornon (VO site) analyzed in Chapters 2 and Chapter III shared only 16.1% of OTUs though identical plant species were compared. As discussed above, this might be due to the scale of the sampling effort and to a too low number of individual plants for a given species. Indeed, the number of unique OTUs increased linearly with the number of samples for a given plant species.

Even taking this into account, there were still a few virus OTUs (n=6) steadily detected from the same environment over time. Interestingly, most of them are dsRNA viruses (*Amalgaviridae*, *Chrysoviridae*, *Partitiviridae*, and *Totiviridae*) except one *Potyvirus* (turnip mosaic virus). Theses dsRNA viruses were stably detected, possibly because as persistent plant viruses, they were vertically transmitted to the next plant generation (this might possibly be the case of the *Amalgaviridae*, as members of this family are so far not known from fungi), or because as mycoviruses they horizontally spread with fungi to gain higher prevalence and resources in the

sampled environment (Roossinck, 2015). In most cases we know little about persistent viruses or mycoviruses to assess their symbiotic lifestyle.

Unfortunately, despite the size of the culturomics efforts made in Chapter III, we still cannot bring a clear answer to the phytovirus vs mycovirus dilemma associated with the large number of persistent viruses identified in plant-associated viromes. The understanding of the role(s) these viruses may play in the natural environment will still need to be further explored in the future.

2. Why different patterns for leaf-associated viromes and mycobiomes of wild plant populations between cultivated and unmanaged environments?

In Chapter III, I showed the existence of a core mycobiome shared by plant pools that do not have a single plant species in common but come from the same sampling site. This may reflect the fact that fungi have huge populations sizes, with millions of spores produced by a single diseased plant, great dispersal abilities, and several generations per year, enabling rapid adaptation (Gladieux *et al.*, 2011). It is thus possible that this core, site specific mycobiome may reflect surface contamination of the sampled plants by spores. It may also reflect the presence of promiscuous fungal species able to colonize a wide range of plant species. In any case, fungal communities showed sufficient divergence from site to site to allow to distinguish them. This "site-specific" signature was clearly stronger for viromes (Chapter III, Figure S4), which might reflect a superior, less stochastic, dispersal ability for fungi than for viruses.

As far as we know, leaf-associated fungal communities are highly influenced by the host plant, climatic and microclimatic variables and by microbial interactions (Vacher *et al.*, 2016). I found richer mycobiomes from unmanaged/natural ecosystems than from cultivated ones (Chapter III, Figure 3). This might result from a lower plant biodiversity at the cultivated sampling sites (Compant *et al.*, 2019) and/or from an indirect impact of fungicides applications on overall fungal diversity (Newton *et al.*, 2010). Indeed, the application of fungicides significantly decreased fungi richness on wheat leaves Karlsson *et al.* (2014).

Interestingly, following the cultivated/unmanaged factor, the geographic location (Villenave d'Ornon vs Bergerac) also showed a significant but lower ($R = 0.561$, $p = 1E\text{-}04$) effect on mycobiome composition. This is consistent with the study of Karlsson *et al.* (2014) who also found significant differences between wheat leaves mycobiomes collected from different areas in Sweden and hypothesized that this might result from differences in climatic conditions and agricultural management (Karlsson *et al.*, 2014). In our case, these observations cannot be deeply

discussed because the number of sampling site is limited so that the ability to separate the respective influence of different factors is in turn limited. Further yearly monitoring in more numerous sites and with detailed collection of metadata for each sampling site would be needed to clarify the interactions between microbes and environmental conditions.

A contrasted picture was obtained from the viromes data for the same experimental sampling points in that more viral families were found from cultivated sampling sites ecosystems than from unmanaged ones. This result is in line with that of Bernardo *et al.* (2018) who also observed a higher family-level virus diversity in cultivated areas. The results are less clear-cut when considering viral richness as estimated by the number of OTUs. The finding of a lower viral richness for sites with a higher mycobiome diversity suggests that virome and mycobiome richness may not be influenced by the same drivers. Climatic conditions are not expected to directly impact viromes because viruses are generally considered to be able to develop wherever their host plants can grow. However, vector populations are reported to be an important factor affecting virome composition (Anderson *et al.*, 2004) and, in turn, vector populations could very well be influenced by environmental conditions. Differences in dispersion mechanisms between fungi and viruses or the contrasted impact of fungicide treatments in mycobiomes and viromes are certainly among potential driver candidates. Domestication and cultivation, by reducing biodiversity have been suggested to be responsible for increased viral infections in cultivated ecosystems (Roossinck and García-Arenal, 2015).

Mycoviral diversity would be expected show a trend parallel to that of fungal diversity (Roossinck, 2015), and therefore to be highest in unmanaged sites. The observation of a lower viral diversity in unmanaged sites (with a higher fungal diversity) possibly suggests that the contribution of the mycovirome to the overall plant-associated virome may be limited. It is tempting to bridge this notion to the results obtained in Chapter IV and which showed almost no shared OTUs between the mycovirome from cultivated mycelia and the plant-associated virome as both elements point in the same direction. However, other effects or experimental biases might also explain these observations ad it is probably safer to refrain from making strong conclusions at this stage.

3. Virus exchanges between tomato (S. lycopersicum) and its wild-relative, the european black nightshade (S. nigrum)

Using metagenomcis I studied virus diversity in tomato crops and in a wild tomato relative, the European black nightshade. No clear conclusions could be reached on virome richness comparisons between these two species. On the other hand, the results obtained document the

circulation of viruses between these two plant populations. The ability to assemble near complete genomes for several viruses (PVY, BBWV1, STV, and the new SnIV1) from the virome HTS data allowed to document the circulation of viruses down to an intra-specific, strain or possibly isolate level.

BBWV1, an aphid-transmitted *Fabavirus* with a relatively wide host range (Blancard, 2012; Carpino *et al.*, 2019; Taylor and Stubbs, 1972), was surprisingly only detected from nightshade samples and not in tomato ones. This observation suggests the existence of a biological or epidemiological barrier limiting the spread of BBWV1 from nightshade to tomato. The results of Carpino *et al.* (2019) show that several BBWV1 isolates were able to infect tomato but that this infection was inefficient with only 40-60% of inoculated plants becoming infected. It is tempting to speculate that this low infection efficiency might indeed be the barrier limiting the spread of BBWV1 to tomato. However, BBWV1 infection was frequent in nightshade but is not reported to be regionally frequent in a well-known BBWV1 host, pepper. Other factors may therefore also be at play in the particularly diverse BBWV1 populations identified here and direct experimental efforts analyzing the ability of these BBWV1 isolates to infect tomato and pepper are clearly needed in order to reach firm conclusions.

We also discovered a new ilarvirus (SnIV1) in both tomato and nightshade populations. Moreover, reanalysis of the metagenomics data of Chapter I showed that this virus was already present in 2011 at the VO site, in *S. villosum* (hairy nightshade) a close relative of *S. nigrum*. Ilarviruses are not known to have insect vectors but are transmitted by pollen, this transmission being sometimes facilitated by pollinating insects or by pollen-eating thrips species. It is conceivable that this pollen-mediated transmission might have a low efficiency between nightshade and tomato. Indeed, such a barrier has been postulated as being responsible for the differentiation of the populations of another pollen-transmitted virus, cherry leaf roll virus (Rebenstorf *et al.*, 2006). On the other hand, SnIV1 would be expected to be more efficiently transmitted from tomato to tomato, so that past the initial hurdle of nightshade to tomato transfer it could generate epidemies. Indeed, the pattern seen with two SnIV1 relatives, Parietaria mottle virus (Roggero *et al.*, 2000) and tomato necrotic spot virus (Batuman *et al.*, 2011) fits this scenario, with outbreaks popping up in different areas of the world without a clear underlying logic. Whether SnIV1 is pathogenic to remains to be evaluated. The mechanism(s) allowing it to persist in nightshade populations similarly remains to be investigated but as ilarviruses are frequently seed-borne (Sastry, 2013) in some hosts, seed transmission in nightshade appears as a possible hypothesis.

Perspectives

1. HTS-based Metagenomics: Methodological Challenges and perspectives

Despite its relatively brief history, the study of microbial life through next-generation sequencing (NGS) technologies and computational biology is defining a new era in microbiology. Detailed characterization of the features of such communities is instrumental to our comprehension of ecological, biological, and clinical complexity (Laudadio *et al.*, 2019). Despite the fact that high throughput sequencing has been widely used in plant virology, this has largely been only in a virus discovery perspective and much less frequently in a true metagenomics perspective. A consequence is that we still encounter technical and methodological challenges for metagenomics approaches, such study design and sampling strategy, choice of wet-lab approach, data processing and interpretation as discussed above. To improve upon the current situation, some perspectives are suggested here for the future metagenomic study of plant viruses.

- ***Be careful with impalpable traps when dealing with the huge amounts of data involved***

  First, low level cross-contamination due to index-hopping or other artifacts seems to be very frequent if not systematic in HTS and as exemplified in this thesis, various strategies need to be implemented to deal with this issue. These can include PCR validation efforts, determination of cross-talk thresholds, the systematic use of negative and/or positive controls etc… Second the sequencing depth affects the output in terms of virus richness, and when comparisons between samples is the objective, there is a clear need for data normalization. Here, I used a normalization strategy involving resampling reads to an identical depth but other strategies are also possible and, in some cases have been advocated as being more reliable. Different normalization strategies for microbiome analysis were compared and their impacts on interpretation of ecological and statistical importance have been evaluated in Weiss *et al.* (2017). The results suggest rarefying the library size (resampling as we did) is a still a useful normalization technique which can more effectively mitigate the artifact than other normalization techniques. Particularly the normalization is very important for the beta diversity analysis of different microbial communities (ex. PERMANOVA test, the library size should be included if the dataset was not normalized)(Weiss *et al.*, 2017).

- ***Increase the sampling scales involving individual plants/species, ecosystems…***

The sampling scales involved in this thesis (individual pools of 15 plants/species, large pools of 40 selected plant species with 5 individuals per species, pools of up to 100 individual plants/species) encompass a wide range of possibilities. Yet it is unlikely that any of them captures fully the virus community in the sampled environments. This concerns both the spatial and temporal scales as well as the sampling depth for individual plant species. Because of the complexity and heterogeneity of vegetation, we cannot easily set the same sampling scale for all sites. However, depending on the question, it may be important in the future to screen more individual plants, involving the same or different species. If the objective is that of a larger scale screen of plant viruses in order to access the virome of an environment, blind sampling but well-designed scaling may provide us different information as "geometagenomics" implemented by Bernardo *et al.* (2018).

- ***Strict parallel design and test will facilitate the comparison between microorganisms involving composition, infectious pattern, prevalence…***

Summarized from the study of Chapter III, we had a deeper mycobiome vision with the use of 4 separate subpools per site than for virome, for which we used only a single megapool. We could indeed compare communities between sites but could not compare the virome between subpools and, therefore, not address the question or a core, site-specific virome as we were able to show for the mycobiome. The more discrete analysis of sample undoubtedly provides more information. On the other hand the more samples separately analyzed, the more important samples cross-talk problem become. However, for the future parallel study of different type of microorganisms, a discrete and parallel design is essential to better interpret commonalities and differences.

- ***Adventurous exploration of third generation sequencing on plant virome study***

I have suggest in what precedes some strategies or solutions for the current HTS-based metagenomic plant virus studies. However, to be at the forefront of the field, attempts towards the application of third generation sequencing to plant virome studies may provide fantastic view as was recently show for the detection of known or novel viruses (Adams *et al.*, 2017; Bronzato Badial *et al.*, 2018; Filloux *et al.*, 2018b). Since this technology may provide excellent genome reconstruction together with high consensus sequence accuracy it might

represent the future for viral metagenomics because it could solve many of the problems associated with incomplete or chimeric genome assemblies.

2. Pathology and ecology importance of plant viruses

Through this thesis, I discovered many new viruses, however this discovery is just the starting point of the long path of the exploration of plant virus diversity. In the future, a wider range of plant species (crops/weeds, trees, grass…) could be screened to gradually fill the huge gap in our knowledge of plant virus diversity. Meantime, real-time surveillance would be helpful for the early detection of emerging diseases and would allow us to take early and more efficient action (Anderson *et al.*, 2004). Studying the temporal and spatial variation of phytoviromes can help to find the drivers shaping the virus communities and may also allow us to understand the barriers or drivers of virus flow between species (Bernardo *et al.*, 2018). The parallel analysis in this thesis of fungal communities and of viral communities illuminated an avenue for the future study of the plant microbiome in a fully integrated perspective, taking into account all microbes interacting directly or indirectly with a plant (Compant *et al.*, 2019), which is likely needed to begin to understand the full complexity of plant holobiontes.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abrescia NGA, Bamford DH, Grimes JM, Stuart DI (2012). Structure Unifies the Viral Universe. Annu Rev Biochem 81**:** 795-822.

Adams I, Fox A (2016). Diagnosis of plant viruses using next-generation sequencing and metagenomic analysis. Current Research Topics in Plant Virology. Springer. pp 323-335.

Adams IP, Glover RH, Monger WA, Mumford R, Jackeviciene E, Navalinskiene M, Samuitiene M, Boonham N (2009). Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Mol Plant Pathol 10**:** 537-545.

Adams IP, Braidwood LA, Stomeo F, Phiri N, Uwumukiza B, Feyissa B, Mahuku G, Wangi A, Smith J, Mumford R, Boonham N (2017). Characterising maize viruses associated with maize lethal necrosis symptoms in sub Saharan Africa. bioRxiv 10.1101/161489**:** 161489.

Aiewsakun P, Simmonds P (2018). The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification. Microbiome 6**:** 38.

Al Rwahnih M, Daubert S, Úrbez-Torres JR, Cordero F, Rowhani A (2011). Deep sequencing evidence from single grapevine plants reveals a virome dominated by mycoviruses. Arch Virol 156**:** 397-403.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. J Mol Biol 215**:** 403-410.

Anderson PK, Cunningham AA, Patel NG, Morales FJ, Epstein PR, Daszak P (2004). Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. Trends Ecol Evol 19**:** 535-544.

Andika IB, Wei S, Cao C, Salaipeth L, Kondo H, Sun L (2017). Phytopathogenic fungus hosts a plant virus: A naturally occurring cross-kingdom viral infection. Proc Natl Acad Sci USA 114**:** 12267-12272.

Andrews S (2010). FastQC: a quality control tool for high throughput sequence data.

Bachman S (2016). State of the World's Plants Report: Royal Botanic Gardens, Kew. pp 7-84.

Bag S, Al Rwahnih M, Li A, Gonzalez A, Rowhani A, Uyemoto JK, Sudarshana MR (2015). Detection of a New Luteovirus in Imported Nectarine Trees: A Case Study to Propose Adoption of Metagenomics in Post-Entry Quarantine. Phytopathology 105**:** 840-846.

Baltimore D (1974). The strategy of RNA viruses. Harvey Lect 70 Series**:** 57-74.

Batuman O, Chen L, Gilbertson R (2011). Characterization of Tomato necrotic spot virus (ToNSV), a new ilarvirus species infecting processing tomatoes in the Central Valley of California, vol. 101.

Beerenwinkel N, Gunthard HF, Roth V, Metzner KJ (2012). Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front Microbiol 3**:** 329.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ *et al* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456**:** 53.

Bernardo P, Charles-Dominique T, Barakat M, Ortet P, Fernandez E, Filloux D, Hartnady P, Rebelo TA *et al* (2018). Geometagenomics illuminates the impact of agriculture on the distribution and prevalence of plant viruses at the ecosystem scale. ISME J 12**:** 173-184.

Blackwell M (2011). The fungi: 1, 2, 3 ... 5.1 million species? Am J Bot 98**:** 426-438.

Blancard D (2012). Tomato diseases: identification, biology and control: A Colour Handbook. CRC Press.

Blawid R, Silva JMF, Nagata T (2017). Discovering and sequencing new plant viral genomes by next-generation sequencing: description of a practical pipeline. Ann Appl Biol 170**:** 301-314.

Boccardo G, Lisa V, Luisoni E, Milne RG (1987). Cryptic Plant Viruses. In: Maramorosch K, Murphy FA, Shatkin AJ (eds). Advances in Virus Research. Academic Press. pp 171-214.

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 Res 186**:** 20-31.

Breitbart M, Rohwer F (2005). Here a virus, there a virus, everywhere the same virus? Trends Microbiol 13**:** 278-284.

Bronzato Badial A, Sherman D, Stone A, Gopakumar A, Wilson V, Schneider W, King J (2018). Nanopore sequencing as a surveillance tool for plant pathogens in plant and insect tissues. Plant Dis 102**:** 1648-1652.

Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S, Cruaud C *et al* (2015). Patterns and ecological drivers of ocean viral communities. Science 348**:** 1261498.

Candresse T, Filloux D, Muhire B, Julian C, Galzi S, Fort G, Bernardo P, Daugrois J-H, Fernandez E, Martin DP, 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.

Carpino C, Elvira-González L, Rubio L, Peri E, Davino S, Galipienso L (2019). A comparative study of viral infectivity, accumulation and symptoms induced by broad bean wilt virus 1 isolates. J Plant Pathol 101**:** 275-281.

Carroll D, Daszak P, Wolfe ND, Gao GF, Morel CM, Morzaria S, Pablos-Méndez A, Tomori O, Mazet JAK (2018). The Global Virome Project. Science 359**:** 872-874.

Cieniewicz EJ, Pethybridge SJ, Loeb G, Perry K, Fuchs M (2018). Insights Into the Ecology of Grapevine red blotch virus in a Diseased Vineyard. Phytopathology 108**:** 94-102.

Coetzee B, Freeborough M-J, Maree HJ, Celton J-M, Rees DJG, Burger JT (2010). Deep sequencing analysis of viruses infecting grapevines: Virome of a vineyard. Virology 400**:** 157-163.

Compant S, Samad A, Faist H, Sessitsch A (2019). A review on the plant microbiome: Ecology, functions, and emerging trends in microbial application. Journal of Advanced Research 19**:** 29-37.

Cooper I, Jones RAC (2006). Wild Plants and Viruses: Under-Investigated Ecosystems. Advances in Virus Research. Academic Press. pp 1-47.

Cox S, Mayo M, Jones AT (2000). The occurrence of dsRNA species in apparently healthy and virus-infected Ribes cultivars, and evidence that one such species originates from a member of the virus family Totiviridae. Eur J Plant Pathol 106**:** 353-364.

Creager ANH, Scholthof K-BG, Citovsky V, Scholthof HB (1999). Tobacco Mosaic Virus: Pioneering Research for a Century. The Plant Cell 11**:** 301-308.

Deamer D, Akeson M, Branton D (2016). Three decades of nanopore sequencing. Nat Biotechnol 34**:** 518-524.

Delwart EL (2007). Viral metagenomics. Rev Med Virol 17**:** 115-131.

Duffy S, Shackelton LA, Holmes EC (2008). Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet 9**:** 267-276.

Dulieu P, Penin F, Gautheron D (1988). Purification of virus-like particles from male-sterile Vicia faba and detection by ELISA in crude leaf extracts. Plant Sci 56**:** 9-14.

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ *et al* (2018). The Pfam protein families database in 2019. Nucleic Acids Res 47**:** D427-D432.

Elena SF, Fraile A, Garcia-Arenal F (2014). Evolution and emergence of plant viruses. Adv Virus Res 88**:** 161-191.

Emerson JB, Thomas BC, Andrade K, Heidelberg KB, Banfield JF (2013). New Approaches Indicate Constant Viral Diversity despite Shifts in Assemblage Structure in an Australian Hypersaline Lake. Appl Environ Microbiol 79**:** 6755-6764.

Esling P, Lejzerowicz F, Pawlowski J (2015). Accurate multiplexing and filtering for high-throughput amplicon-sequencing. Nucleic Acids Res 43**:** 2513-2524.

Fabre F, Montarry J, Coville J, Senoussi R, Simon V, Moury B (2012). Modelling the Evolutionary Dynamics of Viruses within Their Hosts: A Case Study Using High-Throughput Sequencing. PLoS Pathog 8**:** e1002654.

Feldman TS, Morsy MR, Roossinck MJ (2012). Are communities of microbial symbionts more diverse than communities of macrobial hosts? Fungal Biol 116**:** 465-477.

Fermin G, Mazumdar-Leighton S, Tennant P (2018). Chapter 9 - Viruses of Prokaryotes, Protozoa, Fungi, and Chromista. In: Tennant P, Fermin G, Foster JE (eds). Viruses. Academic Press. pp 217-244.

Filloux D, Dallot S, Delaunay A, Galzi S, Jacquot E, Roumagnac P (2015). Metagenomics Approaches Based on Virion-Associated Nucleic Acids (VANA): An Innovative Tool for Assessing Without A Priori Viral Diversity of Plants. Methods Mol Biol 1302**:** 249-257.

Filloux D, Fernandez E, Comstock JC, Mollov D, Roumagnac P, Rott P (2018a). Viral Metagenomic-Based Screening of Sugarcane from Florida Reveals Occurrence of Six

Sugarcane-Infecting Viruses and High Prevalence of Sugarcane yellow leaf virus. Plant Dis 102**:** 2317-2323.

Filloux D, Fernandez E, Loire E, Claude L, Galzi S, Candresse T, Winter S, Jeeva ML, Makeshkumar T, Martin DP, Roumagnac P (2018b). Nanopore-based detection and characterization of yam viruses. Sci Rep 8**:** 17879-17879.

Fonseca NA, Rung J, Brazma A, Marioni JC (2012). Tools for mapping high-throughput sequencing data. Bioinformatics 28**:** 3169-3177.

Francois S, Filloux D, Frayssinet M, Roumagnac P, Martin DP, Ogliastro M, Froissart R (2018). Increase in taxonomic assignment efficiency of viral reads in metagenomic studies. Virus Res 244**:** 230-234.

Gelderblom HR (1996). Structure and classification of viruses. Medical Microbiology. 4th edition. University of Texas Medical Branch at Galveston.

Gladieux P, Byrnes III EJ, Aguileta G, Fisher MC, Heitman J, Giraud T (2011). Epidemiology and evolution of fungal pathogens in plants and animals. Genetics and Evolution of infectious disease. Elsevier. pp 59-132.

Gu Y-H, Tao X, Lai X-J, Wang H-Y, Zhang Y-Z (2014). Exploring the Polyadenylated RNA Virome of Sweet Potato through High-Throughput Sequencing. PLoS One 9**:** e98884.

Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, Li W, Chitsaz F *et al* (2018). RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res 46**:** D851-d860.

Hamad I, Ranque S, Azhar EI, Yasir M, Jiman-Fatani AA, Tissot-Dupont H, Raoult D, Bittar F (2017). Culturomics and Amplicon-based Metagenomic Approaches for the Study of Fungal Population in Human Gut Microbiota. Sci Rep 7**:** 16788.

Hamilton AJ, Baulcombe DC (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. Science 286**:** 950-952.

Hilarino MPA, Silveira FAdOe, Oki Y, Rodrigues L, Santos JC, Corrêa Junior A, Fernandes GW, Rosa CA (2011). Distribution of the endophytic fungi community in leaves of Bauhinia brevipes (Fabaceae). Acta Botanica Brasilica 25**:** 815-821.

Horiuchi H, Udagawa T, Koga R, Moriyama H, Fukuhara T (2001). RNA-dependent RNA polymerase activity associated with endogenous double-stranded RNA in rice. Plant Cell Physiol 42**:** 197-203.

Idris A, Al-Saleh M, Piatek MJ, Al-Shahwan I, Ali S, Brown JK (2014). Viral metagenomics: analysis of begomoviruses by illumina high-throughput sequencing. Viruses 6**:** 1219-1236.

Illumina (2017). Effects of index misassignment on multiplexing and downstream analysis https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf.

Jain M, Olsen HE, Paten B, Akeson M (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol 17**:** 239.

Jeske H (2018). Barcoding of plant viruses with circular single-stranded DNA based on rolling circle amplification. Viruses 10**:** 469.

Jo Y, Choi H, Cho JK, Yoon J-Y, Choi S-K, Cho WK (2015). In silico approach to reveal viral populations in grapevine cultivar Tannat using transcriptome data. Sci Rep 5**:** 15841.

Jo Y, Choi H, Kim S-M, Kim S-L, Lee BC, Cho WK (2016). Integrated analyses using RNA-Seq data reveal viral genomes, single nucleotide variations, the phylogenetic relationship, and recombination for Apple stem grooving virus. BMC Genomics 17**:** 579.

Jo Y, Choi H, Kim S-M, Kim S-L, Lee BC, Cho WK (2017). The pepper virome: natural co-infection of diverse viruses and their quasispecies. BMC Genomics 18**:** 453.

Jo Y, Lian S, Chu H, Cho JK, Yoo S-H, Choi H, Yoon J-Y, Choi S-K, Lee BC, Cho WK (2018). Peach RNA viromes in six different peach cultivars. Sci Rep 8**:** 1844.

Karlsson I, Friberg H, Steinberg C, Persson P (2014). Fungicide effects on fungal community composition in the wheat phyllosphere. PLoS One 9**:** e111786-e111786.

Katsiani A, Maliogka VI, Katis N, Svanella-Dumas L, Olmos A, Ruiz-Garcia AB, Marais A, Faure C, Theil S, Lotos L, Candresse T (2018). High-Throughput Sequencing Reveals Further Diversity of Little Cherry Virus 1 with Implications for Diagnostics. Viruses 10.

Kinoti WM, Constable FE, Nancarrow N, Plummer KM, Rodoni B (2017). Generic Amplicon Deep Sequencing to Determine Ilarvirus Species Diversity in Australian Prunus. Front Microbiol 8**:** 1219-1219.

Klingenberg H, Aßhauer KP, Lingner T, Meinicke P (2013a). Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. Bioinformatics (Oxford, England) 29**:** 973-980.

Koonin EV, Senkevich TG, Dolja VV (2006). The ancient Virus World and evolution of cells. Biol Direct 1**:** 29-29.

Krenz B, Thompson JR, Fuchs M, Perry KL (2012). Complete genome sequence of a new circular DNA virus from grapevine. J Virol 86**:** 7715.

Krenz B, Thompson JR, McLane HL, Fuchs M, Perry KL (2014). Grapevine red blotch-associated virus Is Widespread in the United States. Phytopathology 104**:** 1232-1240.

Kutnjak D, Rupar M, Gutierrez-Aguirre I, Curk T, Kreuze JF, Ravnikar M (2015). Deep Sequencing of Virus-Derived Small Interfering RNAs and RNA from Viral Particles Shows Highly Similar Mutational Landscapes of a Plant Virus Population. J Virol 89**:** 4760.

Laforest-Lapointe I, Whitaker BK (2019). Decrypting the phyllosphere microbiota: progress and challenges. Am J Bot 106**:** 171-173.

Lagier J-C, Hugon P, Khelaifia S, Fournier P-E, La Scola B, Raoult D (2015). The rebirth of culture in microbiology through the example of culturomics to study human gut microbiota. Clin Microbiol Rev 28**:** 237-264.

Lagier J-C, Khelaifia S, Alou MT, Ndongo S, Dione N, Hugon P, Caputo A, Cadoret F, Traore SI, Dubourg G (2016). Culture of previously uncultured members of the human gut microbiota by culturomics. Nature microbiology 1**:** 16203.

Lagier JC, Armougom F, Million M, Hugon P, Pagnier I, Robert C, Bittar F, Fournous G, Gimenez G, Maraninchi M (2012). Microbial culturomics: paradigm shift in the human gut microbiome study. Clin Microbiol Infect 18**:** 1185-1193.

Laudadio I, Fulci V, Stronati L, Carissimi C (2019). Next-Generation Metagenomics: Methodological Challenges and Opportunities. OMICS 23**:** 327-333.

Lauring AS, Andino R (2010). Quasispecies theory and the behavior of RNA viruses. PLoS Pathog 6**:** e1001005.

Lefebvre A, Scalla R, Pfeiffer P (1990). The double-stranded RNA associated with the '447' cytoplasmic male sterility in Vicia faba is packaged together with its replicase in cytoplasmic membranous vesicles. Plant Mol Biol 14**:** 477-490.

Letunic I, Bork P (2017). 20 years of the SMART protein domain annotation resource. Nucleic Acids Res 46**:** D493-D496.

Liu H, Fu Y, Xie J, Cheng J, Ghabrial SA, Li G, Peng Y, Yi X, Jiang D (2012). Evolutionary genomics of mycovirus-related dsRNA viruses reveals cross-family horizontal gene transfer and evolution of diverse viral lineages. BMC Evol Biol 12**:** 91.

Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J (2000). Viruses: structure, function, and uses. Molecular Cell Biology. 4th edition. WH Freeman.

Lu Q-Y, Wu Z-J, Xia Z-S, Xie L-H (2015). Complete genome sequence of a novel monopartite geminivirus identified in mulberry (Morus alba L.). Arch Virol 160**:** 2135-2138.

Lu R, Martin-Hernandez AM, Peart JR, Malcuit I, Baulcombe DC (2003). Virus-induced gene silencing in plants. Methods 30**:** 296-303.

Luria N, Smith E, Sela N, Koren A, Lachman O, Dombrovsky A (2019). Insights Into a Watermelon Virome Contribute to Monitoring Distribution of Whitefly-Borne Viruses. Phytobiomes Journal 3**:** 61-70.

Ma Y, Navarro B, Zhang Z, Lu M, Zhou X, Chi S, Di Serio F, Li S (2015). Identification and molecular characterization of a novel monopartite geminivirus associated with mulberry mosaic dwarf disease. J Gen Virol 96**:** 2421-2434.

Maclot F, Candresse T, Filloux D, Roumagnac P, Massart S: Effect of species composition on virome diversity in various ecosystemic communities of Poaceae. *Rencontres de Virologie Végétale (RVV 2019)*; Aussois, France. 2019. http://agritrop.cirad.fr/590987/.

Maliogka VI, Minafra A, Saldarelli P, Ruiz-García AB, Glasa M, Katis N, Olmos A (2018). Recent Advances on Detection and Characterization of Fruit Tree Viruses Using High-Throughput Sequencing Technologies. Viruses 10**:** 436.

Malmstrom CM, Melcher U, Bosque-Pérez NA (2011). The expanding field of plant virus ecology: Historical foundations, knowledge gaps, and research directions. Virus Res 159**:** 84-94.

Malmstrom CM, Alexander HM (2016). Effects of crop viruses on wild plants. Curr Opin Virol 19**:** 30-36.

Marais A, Faure C, Couture C, Bergey B, Gentit P, Candresse T (2013). Characterization by Deep Sequencing of Divergent Plum bark necrosis stem pitting-associated virus (PBNSPaV) Isolates and Development of a Broad-Spectrum PBNSPaV Detection Assay. Phytopathology 104**:** 660-666.

Marais A, Faure C, Bergey B, Candresse T (2018). Viral Double-Stranded RNAs (dsRNAs) from Plants: Alternative Nucleic Acid Substrates for High-Throughput Sequencing. Methods Mol Biol 1746**:** 45-53.

Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC *et al* (2009). CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res 37**:** D205-D210.

Maree HJ, Fox A, Al Rwahnih M, Boonham N, Candresse T (2018). Application of HTS for Routine Plant Virus Diagnostics: State of the Art and Challenges. Front Plant Sci 9.

Marzano S-YL, Nelson BD, Ajayi-Oyetunde O, Bradley CA, Hughes TJ, Hartman GL, Eastburn DM, Domier LL (2016). Identification of Diverse Mycoviruses through Metatranscriptomics Characterization of the Viromes of Five Major Fungal Plant Pathogens. J Virol 90**:** 6846-6863.

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

Massart S, Chiumenti M, De Jonghe K, Glover R, Haegeman A, Koloniuk I, Komínek P, Kreuze J *et al* (2018). Virus Detection by High-Throughput Sequencing of Small RNAs: Large-Scale Performance Testing of Sequence Analysis Strategies. Phytopathology 109**:** 488-497.

McLeish MJ, Fraile A, García-Arenal F (2019). Evolution of plant–virus interactions: host range and virus emergence. Curr Opin Virol 34**:** 50-55.

Mehle N, Gutiérrez-Aguirre I, Prezelj N, Delić D, Vidic U, Ravnikar M (2014). Survival and Transmission of Potato Virus Y, Pepino Mosaic Virus, and Potato Spindle Tuber Viroid in Water. Appl Environ Microbiol 80**:** 1455-1462.

Mehle N, Gutiérrez-Aguirre I, Kutnjak D, Ravnikar M (2018). Chapter Four - Water-Mediated Transmission of Plant, Animal, and Human Viruses. In: Malmstrom CM (ed). Advances in Virus Research. Academic Press. pp 85-128.

Mokili JL, Rohwer F, Dutilh BE (2012). Metagenomics and future perspectives in virus discovery. Curr Opin Virol 2**:** 63-77.

Naidu RA, Hughes JD (2003). Methods for the detection of plant virus diseases. Plant Virology in Sub Saharan Africa**:** 233-253.

Nerva L, Varese GC, Falk BW, Turina M (2017). Mycoviruses of an endophytic fungus can replicate in plant cells: evolutionary implications. Sci Rep 7**:** 1908.

Newton A, Gravouil C, Fountaine J (2010). Managing the ecology of foliar pathogens: ecological tolerance in crops. Ann Appl Biol 157**:** 343-359.

Ng TFF, Duffy S, Polston JE, Bixby E, Vallad GE, Breitbart M (2011). Exploring the Diversity of Plant DNA Viruses and Their Satellites Using Vector-Enabled Metagenomics on Whiteflies. PLoS One 6**:** e19050.

Nilsson RH, Kristiansson E, Ryberg M, Hallenberg N, Larsson K-H (2008). Intraspecific ITS variability in the kingdom Fungi as expressed in the international sequence databases and its implications for molecular species identification. Evolutionary bioinformatics 4**:** EBO. S653.

Nilsson RH, Anslan S, Bahram M, Wurzbacher C, Baldrian P, Tedersoo L (2019). Mycobiome diversity: high-throughput sequencing and identification of fungi. Nature Reviews Microbiology 17**:** 95-109.

Nowrotek M, Jałowiecki Ł, Harnisz M, Płaza GA (2019). Culturomics and metagenomics: In understanding of environmental resistome. Frontiers of Environmental Science & Engineering 13**:** 40.

Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA (1986). Microbial ecology and evolution: a ribosomal RNA approach. Annu Rev Microbiol 40**:** 337-365.

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**:** e1002796.

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. Front Microbiol 8**:** 1998-1998.

Pooggin MM (2018). Small RNA-Omics for Plant Virus Identification, Virome Reconstruction, and Antiviral Defense Characterization. Front Microbiol 9**:** 2779-2779.

Poojari S, Alabi OJ, Fofanov VY, Naidu RA (2013). A Leafhopper-Transmissible DNA Virus with Novel Evolutionary Lineage in the Family Geminiviridae Implicated in Grapevine Redleaf Disease by Next-Generation Sequencing. PLoS One 8**:** e64194.

Prendeville HR, Ye X, Morris TJ, Pilson D (2012). Virus infections in wild plant populations are both frequent and often unapparent. Am J Bot 99**:** 1033-1042.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J (2011). The Pfam protein families database. Nucleic Acids Res 40**:** D290-D301.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005). InterProScan: protein domains identifier. Nucleic Acids Res 33**:** W116-W120.

Ravnikar M, Kutnjak D, Mehle N, Pecman A, Bacnik K, Kosel J, Dular M, Filipic A, Dobnik D, Zel J (2018). Water mediated virus transmission: Sources, detection and inactivation. Phytopathology 108.

Rebenstorf K, Candresse T, Dulucq MJ, Büttner C, Obermeier CJJov (2006). Host species-dependent population structure of a pollen-borne plant virus, Cherry leaf roll virus 80**:** 2453-2462.

Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466**:** 334.

Rhoads A, Au KF (2015). PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics 13**:** 278-289.

Rodelo-Urrego M, Pagán I, González-Jara P, Betancourt M, Moreno-Letelier A, Ayllón M, Fraile A, Piñero D, García-Arenal F (2013). Landscape heterogeneity shapes host-parasite interactions and results in apparent plant–virus codivergence. Mol Ecol 22**:** 2325-2340.

Rodríguez-Nevado C, Montes N, Pagán I (2017). Ecological factors affecting infection risk and population genetic diversity of a novel potyvirus in its native wild ecosystem. Front Plant Sci 8**:** 1958.

Roggero P, Ciuffo M, Katis N, Alioto D, Crescenzi A, Parrella G, Gallitelli D (2000). Necrotic disease in tomatoes in Greece and southern Italy caused by the tomato strain of Parietaria mottle virus. J Plant Pathol 82**:** 159.

Roossinck MJ (2005). Symbiosis versus competition in plant virus evolution. Nat Rev Microbiol 3**:** 917-924.

Roossinck MJ (2010). Lifestyle of plant viruses. Philos Trans R Soc B 365.

Roossinck MJ, Saha P, Wiley GB, Quan J, White JD, Lai H, Chavarria F, Shen GA, Roe BA (2010). Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol Ecol 19**:** 81-88.

Roossinck MJ (2011). The big unknown: plant virus biodiversity. Curr Opin Virol 1**:** 63-67.

Roossinck MJ (2012). Plant Virus Metagenomics: Biodiversity and Ecology. Annu Rev Genet 46**:** 359-369.

Roossinck MJ (2015). Metagenomics of plant and fungal viruses reveals an abundance of persistent lifestyles. Front Microbiol 5.

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

Roossinck MJ, Martin DP, Roumagnac P (2015). Plant Virus Metagenomics: Advances in Virus Discovery. Phytopathology 105**:** 716-727.

Rosario K, Nilsson C, Lim YW, Ruan Y, Breitbart M (2009). Metagenomic analysis of viruses in reclaimed water. Environ Microbiol 11**:** 2806-2820.

Rosario K, Breitbart M (2011). Exploring the viral world through metagenomics. Curr Opin Virol 1**:** 289-297.

Rosario K, Padilla-Rodriguez M, Kraberger S, Stainton D, Martin DP, Breitbart M, Varsani A (2013). Discovery of a novel mastrevirus and alphasatellite-like circular DNA in dragonflies (Epiprocta) from Puerto Rico. Virus Res 171**:** 231-237.

Rothberg JM, Leamon JH (2008). The development and impact of 454 sequencing. Nat Biotechnol 26**:** 1117-1124.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K *et al* (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature 475**:** 348-352.

Rott M, Xiang Y, Boyes I, Belton M, Saeed H, Kesanakurti P, Hayes S, Lawrence T, Birch C, Bhagwat B, Rast H (2017). Application of Next Generation Sequencing for Diagnostic Testing of Tree Fruit Viruses and Viroids. Plant Dis 101**:** 1489-1499.

Rwahnih MA, Dave A, Anderson MM, Rowhani A, Uyemoto JK, Sudarshana MR (2013). Association of a DNA Virus with Grapevines Affected by Red Blotch Disease in California. Phytopathology 103**:** 1069-1076.

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**:** 992-998.

Sarhan MS, Hamza MA, Youssef HH, Patz S, Becker M, ElSawey H, Nemr R, Daanaa H-SA *et al* (2019). Culturomics of the plant prokaryotic microbiome and the dawn of plant-based culture media – A review. Journal of Advanced Research 19**:** 15-27.

Sastry KS (2013). Seed-borne plant virus diseases. Springer Science & Business Media.

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci USA 109**:** 6241-6246.

Shates TM, Sun P, Malmstrom CM, Dominguez C, Mauck KE (2018). Addressing Research Needs in the Field of Plant Virus Ecology by Defining Knowledge Gaps and Developing Wild Dicot Study Systems. *Front Microbiol*. p 3305.

Siddell SG, Walker PJ, Lefkowitz EJ, Mushegian AR, Adams MJ, Dutilh BE, Gorbalenya AE, Harrach B *et al* (2019). Additional changes to taxonomy ratified in a special vote by the International Committee on Taxonomy of Viruses (October 2018). Arch Virol 164**:** 943-946.

Simmonds P (2015). Methods for virus classification and the challenge of incorporating metagenomic sequence data. J Gen Virol 96**:** 1193-1206.

Simmonds P, Adams MJ, Benko M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E *et al* (2017). Consensus statement: Virus taxonomy in the age of metagenomics. Nat Rev Microbiol 15**:** 161-168.

Simmons HE, Dunham JP, Stack JC, Dickins BJ, Pagan I, Holmes EC, Stephenson AG (2012). Deep sequencing reveals persistence of intra- and inter-host genetic diversity in natural and greenhouse populations of zucchini yellow mosaic virus. J Gen Virol 93**:** 1831-1840.

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

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

Sudarshana MR, Perry KL, Fuchs MF (2015). Grapevine Red Blotch-Associated Virus, an Emerging Threat to the Grapevine Industry. Phytopathology 105**:** 1026-1032.

Susi H, Filloux D, Frilander MJ, Roumagnac P, Laine A-L (2019). Diverse and variable virus communities in wild plant populations revealed by metagenomic tools. PeerJ 7**:** e6140.

Sutton TDS, Clooney AG, Ryan FJ, Ross RP, Hill C (2019). Choice of assembly software has a critical impact on virome characterisation. Microbiome 7**:** 12.

Taylor RH, Stubbs LL (1972). Broad bean wilt virus 1. http://www.dpvweb.net/dpv/showdpv.php?dpvno=81.

Temin HM, Baltimore D (1972). RNA-Directed DNA Synthesis and RNA Tumor Viruses. In: Smith KM, Lauffer MA, Bang FB (eds). Advances in Virus Research. Academic Press. pp 129-186.

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

Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009). Laboratory procedures to generate viral metagenomes. Nat Protoc 4**:** 470-483.

Unterseher M, Schnittler M (2009). Dilution-to-extinction cultivation of leaf-inhabiting endophytic fungi in beech (Fagus sylvatica L.)--different cultivation techniques influence fungal biodiversity assessment. Mycol Res 113**:** 645-654.

Vacher C, Hampe A, Porté AJ, Sauer U, Compant S, Morris CE (2016). The Phyllosphere: Microbial Jungle at the Plant–Climate Interface. Annu Rev Ecol Evol Syst 47**:** 1-24.

Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA *et al* (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res 18**:** 1051-1063.

van der Valk T, Vezzi F, Ormestad M, Dalén L, Guschanski K (2019). Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. Mol Ecol Resour https://doi.org/10.1111/1755-0998.13009.

van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014). Ten years of next-generation sequencing technology. Trends Genet 30**:** 418-426.

Villamor DEV, Ho T, Al Rwahnih M, Martin RR, Tzanetakis IE (2019). High Throughput Sequencing For Plant Virus Detection and Discovery. Phytopathology 109**:** 716-725.

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

Wang Q, Jia P, Zhao Z (2015). VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. Genome Med 7**:** 2-2.

Weber F, Wagner V, Rasmussen SB, Hartmann R, Paludan SR (2006). Double-Stranded RNA Is Produced by Positive-Strand RNA Viruses and DNA Viruses but Not in Detectable Amounts by Negative-Strand RNA Viruses. J Virol 80**:** 5059-5064.

Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-Baeza Y, Birmingham A, Hyde ER, Knight R (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 5**:** 27.

Wikipédia (2019). Plante,Wikipédia, l'encyclopédie libre. http://fr.wikipedia.org/w/index.php?title=Plante&oldid=160335526.

Wren JD, Roossinck MJ, Nelson RS, Scheets K, Palmer MW, Melcher U (2006). Plant Virus Biodiversity and Ecology. PLoS Biol 4**:** e80.

Wu Q, Ding SW, Zhang Y, Zhu S (2015). Identification of viruses and viroids by next-generation sequencing and homology-dependent and homology-independent algorithms. Annu Rev Phytopathol 53**:** 425-444.

Yadav N, Khurana SMP (2016). Plant Virus Detection and Diagnosis: Progress and Challenges. In: Shukla P (ed). Frontier Discoveries and Innovations in Interdisciplinary Microbiology. Springer India: New Delhi. pp 97-132.

Yang X, Ren Y, Sun S, Wang D, Zhang F, Li D, Li S, Zhou X (2018). Identification of the Potential Virulence Factors and RNA Silencing Suppressors of Mulberry Mosaic Dwarf-Associated Geminivirus. Viruses 10**:** 472.

Yepes LM, Cieniewicz E, Krenz B, McLane H, Thompson JR, Perry KL, Fuchs M (2018). Causative Role of Grapevine Red Blotch Virus in Red Blotch Disease. Phytopathology 108**:** 902-909.

Zhang Y-Z, Chen Y-M, Wang W, Qin X-C, Holmes EC (2019). Expanding the RNA Virosphere by Unbiased Metagenomics. Annual Review of Virology 10.1146/annurev-virology-092818-015851.

ANNEX

Accepted manuscript *"The virAnnot pipeline: a resource for automated viral diversity estimation and operational taxonomy units (OTU) assignation for virome sequencing data"*

The VirAnnot pipeline: a resource for automated viral diversity estimation and operational taxonomy units (OTU) assignation for virome sequencing data

Marie Lefebvre[1], Sébastien Theil[1,2], Yuxin Ma and Thierry Candresse[*]

UMR 1332 BFP, INRA, Univ. Bordeaux, CS20032, 33882 Villenave d'Ornon cedex, France

Abstract: 125 words

Text : 1908 words (Introduction-Materials and Methods-Results-Perspectives-References)

* Corresponding author : Thierry.candresse@inra.fr

1: equally contributed to this work.

2 : current address INRA UMRF, 20, côte de Reyne, 15000 Aurillac, France

Running title: VirAnnot: automated viral diversity estimation

Keywords: metagenomics, virome, bioinformatics, OTU, viral diversity

Abstract

Viral metagenomics relies on high-throughput sequencing and on bioinformatic analyses to access the genetic content and diversity of entire viral communities. No universally accepted strategy or tool currently exists to define Operational Taxonomy Units (OTUs) and evaluate viral alpha or beta diversity from virome data. Here we present a new bioinformatic resource, the virAnnot pipeline, which performs the automated identification of OTUs. RPS-Blastn is used to detect conserved viral protein motifs. The corresponding contigs are then aligned and a clustering approach used to group in the same OTU contigs sharing more than a set identity threshold. A 10% threshold has been validated as producting OTUs that reasonably approach, in many families, the ICTV taxonomy and can therefore be used as a proxy to viral species.

Introduction

Metagenomic approaches rely on high-throughput sequencing (HTS) and on bioinformatic analyses of HTS sequence data to access the genetic content and diversity of entire communities in an unbiased way. The use of metagenomic data has a wide variety of applications for plant pathogens studies, including the identification of potential pathogens for better food security (MacDiarmid *et al.*, 2013) or further our knowledge on the effects of microbiomes and viromes on plants (Bulgarelli *et al*., 2013; Roossinck 2015).

Whereas metagenomics face identification, storage and computational challenges, viral metagenomics is confronted to specific taxonomic assignation difficulties. Unlike cellular organisms as fungi or bacteria for which universally conserved genes (ITS, 16S ribosomal RNA) can be used to define Operational Taxonomic Units (OTUs) through a clustering approach (Caporaso *et al.*, 2010), no such universally shared pattern exists for viruses. A consequence is that no universally accepted strategy or tool currently exists to define OTUs and evaluate viral alpha or beta diversity from virome data (Simmonds, 2015; Nooij *et al*., 2018).

We have developed an automated routine addressing this OTU definition problem and integrated it in VirAnnot, a bioinformatic phytovirome sequence pipeline which already performed the assembly of reads, the identification of viral contigs and their Blast-based taxonomic assignation, steps considered as standard for the analysis of HTS data (Nooij *et al.*, 2018). These steps are then followed by a new clustering strategy which allows to group together in the same OTU contigs that share more than a set identity threshold.

Material and Method

• Sequence datasets

In order to analyse the reproducibility of the VirAnnot OTU clustering output and the closeness of the identified OTUs to taxonomic species recognized in the current International Committee for the Taxonomy of Viruses (ICTV), several datasets generated from complex pools of plants (ca. 40 species and 200 individual plants, Ma *et al*., in preparation) were used. These datasets have been sequenced on an Illumina HiSeq 3000 system at the GeT-PlaGe platform (INRA Toulouse, France) and deposited in the INRA National Data Portal under the identifier https://doi.org/10.15454/TVWBCQ.

• VirAnnot pipeline workflow

The VirAnnot pipeline integrates standard bioinformatic tools in three main steps: (i) reads cleaning, (ii) contigs assembly, (iii) taxonomic classification. Briefly, the first step consists of raw reads quality trimming with a minimum score of 20 and a minimum read length of 70. The reads are then demultiplexed followed by adapters, polyA and, if necessary, multiplex identifier (MID) tag removal using cutadapt (Martin, 2011). In order to limit inter-sample cross talk associated with index-hopping (Illumina, 2017; van der Valk *et al*., 2019), a sub-routine can be implemented to retain only reads having identical MID tags on both pairs members. The second step performs the

assembly of the reads from all selected libraries into contigs using either IDBA-UD (Peng *et al.*, 2012) or SPAdes (Bankevich *et al.*, 2012). Through the third step, the contigs are annotated using BlastN or BlastX (Altschul *et al.*, 1990) against the NCBI GenBank nr or nt sequence databases with a user defined significance threshold (default e-value of 10^{-4}). In addition to similarity searches against protein databases, a search against the PFAM database (Punta *et al.*, 2012) is carried out using RPS-Blast (reverse-position-specific BLAST; Marchler-Bauer *et al.*, 2002) with again a user defined threshold (default e-value of 10^{-4}).

- **Clustering approach of OTU identification**

After the RPS-Blast and BlastX searches, the VirAnnot pipeline identifies OTUs based on a clustering approach performed on sequences encoding conserved viral protein domains (Itzhaki 2011; Koo et al., 2009). All contigs encoding a given virus-specific conserved protein motif, identified by the RPS-Blast annotation, are aligned with reference sequences using the ETE3 Toolkit (Huerta-Cepas et al., 2016). For each viral motif, a distance matrix is then computed using pairwise distances between the aligned sequences. User-defined variables at this stage include the minimum contigs overlap [default of 20 amino acids (aa)] and a distance threshold between OTUs. We routinely use a 10% divergence value but this can be adjusted at will. The matrix is used to generate two identical phylogenetic trees: one using the ETE3 Toolkit for visualisation purposes and one using the hclust clustering method (Müllner, 2013). This second tree is cut according to the distance threshold, determining the OTUs and the contigs integrated into each OTU.

- **Validation of the OTU definition threshold in relation to ICTV taxonomic species**

Various datasets (see above) were analysed using a 10% clustering cut-off value for OTUs defined on the basis of the RNA-dependent RNA polymerase (RdRp) conserved motifs. For several single-stranded RNA (ssRNA) or double-stranded RNA (dsRNA) viral families, the average pairwise

distance between OTUs (amino acid divergence in a 100 aa small region extending on both sides of the GDD conserved triplet) was then compared to that between viral species recognized by *the International Committee on Taxonomy of Viruses (ICTV) using the* MEGA software (Kumar, Stecher, and Tamura 2016).

Results

• Pipeline availability

VirAnnot is freely available at https://github.com/marieBvr/virAnnot. All documentation about the implementation, installation and use of the pipeline is available at https://virannot-docs.readthedocs.io/en/latest/.

• Repeatability

To validate the clustering routine implemented in VirAnnot, its repeatability was evaluated by running the whole pipeline analysis five times on two datasets. A comparison of the OTUs defined on the basis of the well-known RdRp signature sequences of RNA viruses (RdRp1, RdRp2, RdRp3 and RdRp4) in these five independent analyses was then performed. For RdRp1, 26≤OTUs≤27 were obtained, for RdRp2 45≤OTUs≤46, for RdRp3 40 OTUs and for RdRp4 111≤OTUs≤117. Comparison of the OTUs identified between the different clustering repetitions showed that the same OTUs were repeatedly defined (not shown). These comparisons therefore show a good stability of the VirAnnot output and the limited variations observed are likely due to the known variability of the assembly and clustering processes.

• ICTV group / Threshold validation

Given the variability of taxonomic criteria used to define species in different viral families (Simmonds, 2015), the use of a unique simple rule cannot allow to define OTUs closely mimicking

species under all circumstances. Similarly, since the extent of conservation varies between viral conserved motifs, the OTUs cut-off distance parameter may need to be adjusted depending on the motif used. VirAnnot takes this aspect into consideration and the OTU divergence threshold is a user definable parameter. We have however found that a 10% cut-off value defines in many viral families RdRp OTUs that appear to be a reasonable proxy to viral species. We compared the amino acid pairwise distances in the short conserved region around the RdRp motifs between VirAnnot OTUs and between valid ICTV species (Table 1). For comparisons in the ssRNA virus *Potyviridae* family, the average distance between OTUs closely matched that between ICTV species. For the double-stranded RNA families *Totiviridae* and *Partitiviridae*, the average distance between OTUs was slightly higher than that between ICTV species but with a large overlap in values (Table 1), supporting the meaningfulness of the 10% threshold.

Perspectives

This Resource Announcement describes a new HTS virome analysis pipeline, VirAnnot, which allows the automated evaluation of viral OTU richness (alpha diversity) in metavirome data or the comparison of diversity between viromes (beta diversity). In addition, because of the underlying phylogenetic approach, virAnnot is also compatible with the UniFrac strategy of comparison of microbial communities (Lozupone & Knight, 2005; Chen *et al*., 2012). Given the constraints imposed by the strategy used (existence of a contig encoding a conserved protein motif, minimum overlap between contigs for a given motif…), the VirAnnot OTU output represents a lower bound estimate of the total virome complexity. The stability and repeatability of the virAnnot OTU assignation and OTU richness estimation have been confirmed, therefore allowing for an easy and direct comparison of viral richness between samples. The setting at 10% of user-defined divergence threshold between OTUs provides results that approximate, in different families, the ICTV species level, allowing to use such OTUs as a proxy to taxonomic species. The use of higher

or lower threshold values may allow the definition of OTUs representing different taxonomic levels.

Acknowledgements

The authors want to thank Sandy Contreras for her contribution to the early phases of the virAnnot pipeline development and the ANR NET-BIOME 2010 SafePGR project which supported this early development phase. Yuxin Ma was supported by a China Scholarship Council PhD grant.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic Local Alignment Search Tool. J. Mol. Biol. 215:403-410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M. Nikolenko, S.I., Pham, S, Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., and Pevzner, P.A. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19:455-477. https://doi.org/10.1089/cmb.2012.0021.

Bulgarelli, D., Schlaeppi, K., Spaepen, S., van Themaat, E.V.L., and Schulze-Lefert, P. 2013. Structure and functions of the bacterial microbiota of plants. Annu. Rev. Plant Biol. 64:807-838. https://doi.org/10.1146/annurev-arplant-050312-120106

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7:335-336. https://doi.org/10.1038/nmeth.f.303.

Chen, J., Bittinger, K., Charlson, E.S., Hoffmann, C., Lewis, J., Wu, G.D., Collman, R.G., Bushman, F.D., and Li, H. 2012. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics 28:2106-2113. https://doi.org/10.1093/bioinformatics/bts342.

Huerta-Cepas, J., Serra, F., and Bork, P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol. Biol. Evol. 33:1635-168. https://doi.org/10.1093/molbev/msw046.

Illumina. 2017. Effects of index misassignment on multiplexing and downstream analysis. https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf

Itzhaki, Z. 2011. Domain-domain interactions underlying Herpesvirus-human protein-protein interaction networks. PLOS ONE 6(7):e21724. https://doi.org/10.1371/journal.pone.0021724.

Koo, Q.Y., Khan, A.M., Jung, K.O., Ramdas, S., Miotto, O., Tan, T.W., Brusic, W., Salmon, J., and August, J.T. 2009. Conservation and variability of West Nile virus proteins. PLOS ONE 4(4):e5352. https://doi.org/10.1371/journal.pone.0005352.

Kumar, S., Stecher, G., and Tamura, K. 2016. MEGA7: molecular evolutionary genetics analysis Version 7.0 for bigger datasets. Mol. Biol. Evol. 33:1870-1874. https://doi.org/10.1093/molbev/msw054.

Lozupone, C., and Knight, R. 2005. UniFrac: A New Phylogenetic Method for Comparing Microbial Communities. Appl. Env. Microbiol. 71:8228-8235. https://doi.org/10.1128/AEM.71.12.8228-8235.

MacDiarmid, R., Rodoni, B., Melcher, U., Ochoa-Corona, F., and Roossinck, M. 2013. Biosecurity implications of new technology and discovery in plant virus research. PLOS Pathogens 9(8):e1003337. https://doi.org/10.1371/journal.ppat.1003337.

Marchler-Bauer, A., Panchenko A.R., Shoemaker B.A., Thiessen P.A., Geer L.Y., and Bryant S.H. 2002. CDD: A Database of Conserved Domain Alignments with Links to Domain Three-Dimensional Structure. Nucleic Acids Research 30 (1): 281–83. https://doi.org/10.1093/nar/30.1.281

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 17:10-12. https://doi.org/10.14806/ej.17.1.200.

Müllner, D. 2013. Fastcluster: fast hierarchical, agglomerative clustering routines for R and python. J. Statistical Software 53:1-18. https://doi.org/10.18637/jss.v053.i09.

Nooij, S., Schmitz, D., Vennema, H., Kroneman, A., and Koopmans, M.P.G. 2018. Overview of virus metagenomic classification methods and their biological applications. Front. Microbiol. 9. https://doi.org/10.3389/fmicb.2018.00749.

Peng, Y., Leung, H.C.M., Yiu, S.M., and Chin, F.Y.L. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420-1428. https://doi.org/10.1093/bioinformatics/bts174.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L., Eddy, S.R., Bateman, A., and Finn, R.D. 2012. The Pfam protein families database. Nucleic Acids Res. 40:D290-D301. https://doi.org/10.1093/nar/gkr1065

Roossinck, M.J. 2015. Plants, viruses and the environment: ecology and mutualism. Virology, 479-480:271-277. https://doi.org/10.1016/j.virol.2015.03.041.

Simmonds, P. 2015. Methods for virus classification and the challenge of incorporating metagenomic sequence data. J. Gen. Virol. 96:1193-1206. https://doi.org/10.1099/jgv.0.000016.

van der Valk, T., Vezzi, F., Ormestad, M., Dalén, L., and Guschanski, K. 2019.Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. Mol. Ecol. Resour. Epub ahead of print. https://doi.org/10.1111/1755-0998.13009.

Table 1. Statistics on amino acid pairwise distances for OTUs defined using the RNA-dependent RNA polymerase conserved motif and for formal species recognized by the ICTV.

	Category	Total no. of comparisons	Mean	Standard deviation	Min	Median	Max
ssRNA Viral Family	*Potyviridae* OTU	36	0.542	0.076	0.377	0.538	0.635
	Potyviridae ICTV	16,471	0.531	0.104	0.108	0.512	0.823
dsRNA Viral Family	*Totiviridae* OTU	1,225	0.764	0.122	0.25	0.788	0.942
	Totiviridae ICTV	378	0.646	0.13	0.273	0.651	0.969
	Partitiviridae OTU	169	0.727	0.191	0.237	0.808	0.942
	Partitiviridae ICTV	990	0.661	0.158	0.057*	0.723	0.861

* A few pairwise distance values were less than 10%.